THESIS

LEXICAL BUNDLES IN AN ADVANCED INTOCSU WRITING CLASS AND ENGINEERING TEXTS: A FUNCTIONAL ANALYSIS

Submitted by

Mohammed Abdulrahman Alquraishi

Department of English

In partial fulfillment of the requirements

For the Degree of Master of Arts

Colorado State University

Fort Collins, Colorado

Summer 2014

Master's Committee:

Advisor: Douglas Flahive

Anthony Becker
Mary Vogl

ABSTRACT

LEXICAL BUNDLES IN AN ADVANCED INTOCSU WRITING CLASS AND ENGINEERING TEXTS: A FUNCTIONAL ANALYSIS

The purpose of this study is to investigate the functions of lexical bundles in two corpora: a corpus of engineering academic texts and a corpus of IEP advanced writing class texts. This study is concerned with the nature of formulaic language in Pathway IEPs and engineering texts, and whether those types of texts show similar or distinctive formulaic functions. Moreover, the study looked into lexical bundles found in an engineering 1.26 million-word corpus and an ESL 65000-word corpus using a concordancing program. The study then analyzed the functions of those lexical bundles and compared them statistically using chi-square tests. Additionally, the results of this investigation showed 236 unique frequent lexical bundles in the engineering corpus and 37 bundles in the pathway corpus. Also, the study identified several differences between the density and functions of lexical bundles in the two corpora. These differences were evident in the distribution of functions of lexical bundles and the minimal overlap of lexical bundles found in the two corpora. The results of this study call for more attention to formulaic language at ESP and EAP programs.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Douglas Flahive for his continuous support and insightful comments. Studying under him has opened my eyes to a wide range of linguistic issues, particularly corpus linguistics. I would also like to offer my special thanks to my committee members, Dr. Tony Becker and Dr. Mary Vogl who were of great help with their comments and insights.

DEDICATION

I dedicate this work to my small and extended family for their love and support. I am forever grateful to my parents, Abdulrahman and Athraa, who instilled the love for knowledge in me from an early age. I am also grateful to my wife, Sarah, whose love, support, and companionship made this accomplishment possible. Sarah's encouragement and trust in me, even when I had doubts, kept me motivated and on track. Lastly, I am thankful to God for giving my family and me his blessings and mercy.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

It is now a common belief among linguists and language teachers that vocabulary is an essential component of the language learning process. This perception of the importance of vocabulary inclusion in language teaching curricula started as a simple reductionist concept and has become increasingly more complex and encompassing as researchers continue to advance their scientific pursuits. Currently, there is a wide range of areas of linguistic research related to vocabulary such as research on semantics, phraseology, acquisition, register, form, and assessment.

However, this current status was not always the case for vocabulary. In fact, for decades, the lexical aspect of the process of learning a language was neglected and considered to be a by-product of learning the grammar of the language (Chacón-Beltrán, Abello-Contesse, & Torreblanca-López, 2010). Thus, learning vocabulary was a vehicle to learn the grammar of language, not an aspect that is worthy of focused attention (O'Dell, 1997; Zimmerman, 1997). Moreover, not only was the inclusion of vocabulary minimized in second language teaching, the vocabulary that was included was chosen because of its literary value, rather than functional or communicative value (Zimmerman, 1997; Schmitt, 2000). The predominance of such a view made it harder for research-based approaches of looking at vocabulary to be effectively integrated into language classrooms (Schmitt, 2000).

The status of vocabulary was starting to change with the turn of the 20th century when new approaches to teaching language were developed. One example of that slow change in dealing with vocabulary was the Reform Movement's decision to choose vocabulary based on its relationship to real language use rather than literary or grammatical significance (Zimmerman,

1997). This was advanced by West's (1930) sadly neglected criticism of teaching vocabulary of no use and the fact that students were not attaining a mastery level. West (1930) advocated that vocabulary should be taught as a pressing issue and focused on what vocabulary language teachers have to teach as more relevant and more pressing. Similarly, Palmer (1922) in his discussion of practical linguistics showed the need to identify valid criteria for including vocabulary in language curriculum.

From that time foreword, relatively few researchers have been expanding what we know about language learners' needs, and vocabulary in particular. As early as in the late 1950s, Firth (1957) was investigating the multifaceted nature of vocabulary knowledge and identifying several levels of meanings such as the orthographic, grammatical, phonetic, and collocational levels. At a similar time frame, Lado (1957) was also raising several important aspects pertaining to vocabulary knowledge such as frequency, register, and receptive and productive knowledge, to name a few. Moreover, Richards (1976), concerned with vocabulary teaching, attempted to build what could be considered the first comprehensive model of vocabulary knowledge. Richards (1976) used native speakers' lexical repertoire to provide an enhanced baseline of necessary vocabulary knowledge.

In a major breakthrough of L2 vocabulary research, Nation (1990), starting from the learners' needs, developed a model of vocabulary knowledge that was formed around receptive and productive knowledge of vocabulary. The receptive part of knowledge is related to processing and understanding vocabulary while the productive part is relate producing and using vocabulary to communicate. This model has come to be recognized and commonly cited among researchers investigating vocabulary knowledge. Additionally, Nation (2001) revised and rearranged the model to account for expansions in the field. Nation's (2001) model was centered

on three elements: form, meaning, and use, under which Nation dealt with several issues such as register, references, associations, and grammatical functions. However, it is evident in most proposed models of vocabulary knowledge that collocations and the context that surrounds words are considered an important element of vocabulary knowledge (Firth, 1957; Richards, 1976; Nation, 1990; Nation, 2001). With more research, the importance of formulaic language, collocations and lexical bundles is becoming more apparent to non-native speakers (NNSs) of any language in that they help demonstrate command of language and affirm their membership in discoursal communities (Wray, 2002; Biber & Barbieri, 2007; Ädel & Erman, 2012).

Moreover, the focus on formulaic language has seen a rapid development and ample interest from researchers, due to its perceived importance and functional value. This was evident in the realization of Hakuta (1974), Ferguson (1976), and Pawley and Syder (1983) that prefabricated sequences form a sizable chunk of language and that non-native speakers, just like native speakers, use them strategically for fluency. Answering calls for more attention for formulaic sequences, Nattinger and DeCarrico (1992) and Lewis (1993), among others, led one of the seminal attempts to analyze and include formulaic sequences in language teaching curriculum. Such attempts used formal and functional criteria to build a lexical phrase model of teaching language. Moreover, Nattinger, DeCarrico, and Lewis presented lexical phrases as solid units of analysis for different kinds of investigations such as pragmatic, discoursal, and pedagogical analyses. For Lewis (1993) and Nattinger and DeCarrico (1992), since language has a very dominant formulaic nature, it is only natural to embrace formulaicity when teaching language to non-native speakers.

One of the important advances in analyzing formulaic language was the utilization of statistical information about co-occurrences of vocabulary, made possible with the development

of concordancing programs. Moreover, one line of research that has capitalized on this technological advancement is the lexical bundle approach, an approach that came to light by the authors of Longman Grammar of Spoken and Written English, led by Douglas Biber (Biber, Johansson, Leech, Conrad, & Finegan, 1999). Biber et al. (1999) define lexical bundles "as the combinations of words that in fact recur most commonly in a given register" (p.992), such as *on the other hand* and *in the case of*. Lexical bundles, in this approach, were chosen solely based on their frequency of occurrence within a corpus, despite their structural status or their perceived importance (Conrad & Biber, 2004). Additionally, the reliance on frequency as the sole identifier offered an objective and reliable criteria of lexical bundle selection and avoided the subjectivity of linguists' intuitions (Biber et al., 1999; Biber & Conrad, 1999; Conrad & Biber, 2004).

Additionally, although the lexical bundle approach to study formulaicity of language started from a grammatical standpoint in Longman Grammar, it soon attracted researchers interested in second language acquisition. It was hard not to see the value that lexical bundles research presented to the fields of English for academic purposes (EAP) and English for specific purposes (ESP), especially with the attachment of this line of research to the specificity of registers. This is evident in studies exploring lexical bundles found in general and specific academic English in different contexts, as in Biber, Conrad, and Cortes's (2004) analysis of lexical bundles in university classroom teaching and textbooks, Conrad and Biber's (2004) analysis of lexical bundles in academic prose and conversation, and Biber and Barbieri's (2007) analysis of lexical bundles in specific university contexts. Similarly, Neely and Cortes (2009) analyzed the spoken language of lecturers and students to identify the functions of lexical bundles in both corpora, which informed the design of an EAP listening curriculum.

Another prominent theme in lexical bundles research in EAP and ESP is focused on analyzing lexical bundles in specific disciplines. For instance, Hyland (2008) compared lexical bundles in four disciplines, engineering, microbiology, business studies, and applied linguistics. Similarly, Cortes (2004) analyzed lexical bundles in academic history and biology texts and compared them students' writing in those two disciplines. This line of studies is important in that it shows the gaps between writing conventions within disciplines and students' writings, whether native-speakers or non-native. Along those lines, Ädel and Erman (2012) compared the use of lexical bundles among native speakers and non-native speakers of English writing linguistic papers. However, despite the extensive research investigating lexical bundles relative to second language acquisition, to the knowledge of the researcher, only one study has compared lexical bundles found in texts language learners find in intensive English programs (IEP) and in texts they encounter when they enroll in academic programs afterword. In that study, Chen (2008) compared the nature of formulaic language used in electrical engineering introductory textbooks and ESP textbooks for engineering. The results of Chen's (2008) analysis revealed a vast difference between the two types, and a misrepresentation of the target bundles found in the engineering texts.

To fill this gap, this study is set to analyze and compare two corpora that are composed of what second language learners are exposed to in a representative IEP textbook and supplementary reading and what they are expected to be reading when they enroll in their academic programs. As it will be shown in Chapter II, many studies have looked into formulaicity in language learners' productive aspect of language and compared it to language corresponding with their fields of study. These studies have identified some differences and gaps between students' compositions and spoken language and the readings they encounter at their

respective academic programs. The gaps found were varied in terms of what lexical bundles were used, how often they were used, and how they were used in the texts.

The present study draws its methodology from similar studies that compared the functional aspects of lexical bundles between corpora. This methodology groups lexical bundles based on their discoursal functions within corpora, as it can be seen in the work of Biber, et al. (2004), Hyland (2008), Byrd and Coxhead (2010), and Ädel and Erman (2012). This study will analyze two corpora to identify the similarities and differences between them in relation to lexical bundles. The first corpus is composed of texts that are required for advanced language learners at INTOCSU Pathway program. Pathway programs are discipline-specific transitional programs that are designed to bridge the linguistic and academic gaps for students applying to CSU with insufficient qualifications.

The second corpus is composed of texts collected from required readings in the first semesters of academic studies in engineering programs. The analysis will look into how lexical bundles are used in each set, and whether there are overlapping lexical bundles between the two corpora.

This research is of significance in that it is an area of research that has not received much attention. While many research projects have looked into the productive aspect of second language learners' use of lexical bundles i.e., writing, and compared it to the target language of their fields, the research on the receptive part is still lacking. The significance is further emphasized when one considers that the recognition of textual features is the first step toward being able to produce them. This makes the case for the need to investigate the nature of lexical bundles in texts that language learners encounter in IEPs and how it relates to the nature of lexical bundles in the target language of their respective fields. The proposed research project,

also, has the potential to be expanded in the future to include more areas, and to address shortcomings and advantages found the current academic language teaching practices.

The current study starts with a review of the literature that has dealt with vocabulary in language teaching and vocabulary knowledge. The review, also, discusses corpus linguistic research and research on formulaic language. The study, then, describes the methodology of this research and the analytical framework used in this study. The following section reports the results of the computerized analyses of the corpora and results of statistical analyses used to test the significance of findings. Lastly, the study discusses in detail the findings from the analyses proposed in Chapter III and compares them to findings from previous research. Additionally, the last chapter also presents a discussion of the pedagogical implications of this study, the limitations of this study, and directions for future research.

CHAPTER II

Review of the Literature

This chapter begins with a review of the position of vocabulary in the field of language teaching, and how its value as a component changed over time. Next, the concept of vocabulary knowledge and what it entails is presented from its early development to its current models. The review of those two issues expands the brief introductions presented in Chapter I. Following the review of vocabulary status; the next section reviews corpora studies, including their development and their presence in linguistic field. Then, the main focus of this research, collocations and lexical bundles is reviewed. First, I start with a review of early research on collocations and multi-word sequences. Next, I follow with a review of definitions and analysis of lexical bundles, collocations, and formulaic sequences. Following that, the chapter concludes with a review of the research on lexical bundles in English language teaching.

2.1 Vocabulary in Language Teaching

The following overview describes the role that vocabulary has occupied in language teaching. It follows the historical evolution of early language teaching practices and describes the constantly evolving nature of vocabulary in language teaching. This overview helps situate the current study about lexical bundles and how attention to vocabulary grew to include formulaic language, and particularly lexical bundles which are a subset of formulaic language.

2.1.1 Vocabulary in the early days of language teaching. In the field of second language acquisition, there was not much focus on vocabulary as an important component of language teaching, until recently. Even when vocabulary was introduced, it was usually used to shed the light on grammatical aspects, rather than for the sake of teaching vocabulary (O'Dell,

1997, Zimmerman, 1997). It was a common practice for teachers to turn to linguistic theories of grammars and use them as teaching materials in their language classes (Nattinger & DeCarrico, 1992, p.xiii). This was evident in the Grammar-Translation Method's practices, which dominated the language teaching scene from the early 1800s (Schmitt, 2000). In the grammar-translation method, the inclusion of literary vocabulary was dependent on their facilitation of explaining grammatical rules (Zimmerman, 1997, Schmitt, 2000). Moreover, language teachers, as was the common practice, focused mostly on teaching grammar. Vocabulary development was believed to be the result of exposure to language (Chacón-Beltrán et al., 2010). Not only language teachers held this view, second language acquisition specialists had also been focusing their attention on syntax and phonology as more worthy of attention and "more central to linguistic theory and more critical to language pedagogy" (Zimmerman, 1997, p. 5).

With the rise of other methodologies of teaching language, vocabulary started to find its place within SLA research and language teaching practices. Although the Reform Movement lead by Henry Sweet, which was set to reform language teaching methodologies, emphasized spoken language, there was a slight, but nonetheless important, change in viewing vocabulary (Zimmerman, 1997). Sweet while conceding that "language is made up of words" (1899. p.97, cited in Zimmerman, 1997), affirmed that "we do not speak in words, but in sentences" (1899. p.97, cited in Zimmerman, 1997). However, despite this statement, the significant change in the Reform Movement was that vocabulary was chosen because of its relation to reality rather than its syntactic or literary value (Zimmerman, 1997).

2.1.2 Attention to deficiencies in lexical knowledge. This slow change was further advanced by the work of several scholars from Great Britain and the United States, namely West, Palmer, and Hornby (Zimmerman, 1997). West (1930) stated several deficiencies with the, then,

current language teaching practices such as working on activities that have minimal benefits, learning vocabulary that students have no use for, and not mastering the vocabulary they learn. West (1930) identified the problem as not whether language teachers should focus on vocabulary, but rather what vocabulary should they teach. For West (1930), learning vocabulary is "the primary thing in learning a language" (p.514). Moreover, West's (1930) work was ahead of its time in that it brought attention to several notions relative to vocabulary teaching that were recognized later as important criteria for choosing vocabulary. One of those notions was that the more frequent the word, the more important it is (West, 1930, Schmitt, 2000). Another notion was the range of occurrence in addition to the frequency counts, and that a wide range of occurrence of vocabulary in varying texts is essential to building general language vocabulary lists (West, 1930).

At the same time, Palmer was also influential in shifting the focus toward vocabulary in language teaching research (Zimmerman, 1997). Palmer (1922), when listing the five primary elements of "practical linguistics" (p.136), placed semantics as one component that holds no more or less importance when compared to other linguistic components. Additionally, Palmer's (1922) work clearly showed the necessity of identifying valid criteria for selecting vocabulary, such as "intrinsic utility, sentence-forming utility, grammatical function, [and] regularity…" (p.137). At that time, vocabulary emerged as one crucial component of language teaching and learning, and the attention was given to including vocabulary in the teaching syllabus through systematic and scientific methodology (Zimmerman, 1997, Schmitt, 2000).

However, the attention that vocabulary gained was not without diversions. The emergence of the audio-lingual method had driven the attention away from vocabulary in favor of teaching grammar and the structure of language (Zimmerman, 1997, Larsen-Freeman &

Anderson, 2011). The reasoning behind such balance of teaching grammar and vocabulary was attributed to the over confidence the students felt with knowing so many words in a language (Zimmerman, 1997). Such feeling was believed to give language learners a false sense of mastery of language while most lacked the mastery of grammar. With that in mind, it was important for the method adopters to scale back the focus on vocabulary and shift the attention to working on grammatical aspects and forming linguistic habits (Zimmerman, 1997, Larsen-Freeman & Anderson, 2011). The introduction of new vocabulary was only acceptable if it facilitated drill practices (Larsen-Freeman & Anderson, 2011). With such control of language, the goal was to limit the errors made by the students while they master the grammar and the phonetics of the language (Larsen-Freeman & Anderson, 2011).

2.1.3 Vocabulary in the picture. After the audio-lingual approach, the following approach of teaching language, commonly known as the communicative language teaching approach did not, at first, bring much change in regards to vocabulary (Schmitt, 2000). The communicative language teaching approach put the communicative aspect of language in the center of language teaching and revised teaching methodologies to advance fluency and communicative ability. Although the approach diverged from the audio-lingual behaviorist theory to an approach that derives from cognitive theories, vocabulary occupied a minor niche (Schmitt, 2000). This time, it was not the grammar that over-shadowed vocabulary, but it was rather the focus on functional language (Schmitt, 2000). As with previous approaches to teaching language, the communicative language teaching approach gave "little guidance about how to handle vocabulary" and it was assumed that vocabulary "would take care of itself" (Schmitt, 2000). However, extensions of the communicative practices nowadays place more focus on vocabulary, due to the perceived importance of vocabulary in language teaching. This focus is

achieved through "a principled selection of vocabulary, often according to frequency lists, and an instruction methodology that encourages meaningful engagement with words" (Schmitt, 2000, p.14). Vocabulary has, thus far, gained its place as an essential component of language that has a central role of second language acquisition. The question has shifted from should we teach vocabulary to what is the best way to facilitate the students' acquisition of vocabulary (Sökmen, 1997). Moreover, focus on vocabulary as an essential component of language has raised the question of what are the elements of vocabulary knowledge, as the following paragraphs will show.

2.2 What Does Vocabulary Knowledge Entail?

While learning the meaning of words is the most familiar manifestation of vocabulary knowledge, it is only one element of such knowledge (Schmitt, 2000). However, it is important before reviewing models of vocabulary knowledge to echo Richards' (1976) and Schmitt and McCarthy's (1997) caution that a discussion of models of word knowledge is meant to describe the manifestations of that knowledge rather than how such knowledge is acquired. To quote Richards (1976), "such information cannot be translated directly into teaching procedures" (p.77), but should be used to inform syllabus building and vocabulary assessment, nonetheless. In the following paragraphs, I review the research on vocabulary acquisition and its elements.

2.2.1 Early models of vocabulary knowledge. Early on, Firth (1957) touched on this very subject in his work titled Modes of Meaning (p.190), where he discussed different levels of meanings found in dictionaries and then expanded the concept even more as it will be shown. As Firth (1957) noted, "the lexical meaning of any given word is achieved by multiple statements of meaning as different levels" (p.192), such as the orthographic, phonetic, grammatical, and etymological levels to name a few, all found usually in dictionary entries. Firth (1957) presented

a new mode of meaning, which was the "meaning by collocation" (p.196). According to Firth (1957), a part of the meaning of a word is that it collocates with certain words, and that has nothing to do with the meaning in its conceptual sense, i.e., what is found under a dictionary's entry. Firth's presentation of levels and modes of meaning was a pioneering concept that came to fruition with the work of other linguists. For instance, models proposed by Richards (1976) and Nation (1990), as it will be shown, account for grammatical, collocational, and phonetic aspects of meaning mentioned by Firth (1957), and consider them essential components of vocabulary knowledge.

Around the same time, Lado (1957) in his book Linguistics Across Cultures provided what could be considered a first step toward a vocabulary knowledge model. With language teachers in mind, he started with noting the lack of attention that vocabulary had received, and then proceeded to state specific lexical elements that he considered important for both the language teachers and learners (Lado, 1957). Lado's (1957) work touched on several issues of importance to comprehensive vocabulary knowledge such as form, different meanings, frequency, and register. Moreover, Lado (1957), also, distinguished between receptive and productive vocabulary knowledge, and the different balance learners might need from both types. Furthermore, comparing Lado's (1957) model to Firth's remarks about modes of meaning reveal that Lado's model was more comprehensive as it presented a structured model based on *form, meaning,* and *distribution*. However, one of the shortcomings of Lado's model was building a model based on contrasting two vocabulary systems from two languages, rather than building a model that comprehensively covered the elements of vocabulary knowledge.

Moreover, Richards (1976) took on the task of building a model of word knowledge that accounted for the varying aspects related to vocabulary discussed in the research. Richards'

(1976) attempt to build such a model was prompted by the lack of theories and models that analyze vocabulary similar to the theories related to grammar and other linguistic aspects of language. To build his model of vocabulary knowledge, Richards (1976) provided several assumptions about the native speakers' (NS) knowledge about words:

1. The vocabulary of a NS continues to grow with little growth to his syntax.
2. A NS knows the approximate frequency of a word and what words collocates with it.
3. NSs know the situational and functional limitations of use of words.
4. NSs know the syntactic information related to words, thus linking the syntactic and lexical systems.
5. NSs know the stem of a word and how to make derivations and new words from it.
6. NSs know the network of associations related to a word.
7. NSs know the semantic values of words.
8. NSs know the different meanings of words.

After stating these assumptions of word knowledge and discussing several ways of addressing such issues in the language classroom, Richards (1976) concluded with a statement about how such a model could enhance the vocabulary teaching experience:

> The goals of vocabulary teaching must be more than simply covering a certain number of words on a word list. Then we must look to how teaching techniques can help realize our concept of what it means to know a word. As in all areas of the syllabus, our understanding of the nature of what we are teaching, should be reflected in the way we set about teaching it. Vocabulary has for some

time been one area of the syllabus where this link between approach, method and technique has been neglected (p.88).

Moreover, one major challenge of dealing with vocabulary knowledge with NSs in mind is that while NSs build their complex lexical system over time, non-native speakers (NNSs) are faced with such complexity in a short period of time (Laufer, 1997). In fact, it is not uncommon that NNSs may have learned only a partial knowledge about their vocabulary, while educated NSs usually have a more comprehensive knowledge about their vocabulary (Laufer, 1997). Thus, vocabulary's multifaceted knowledge often causes NNSs to encounter difficulties on one front or another when dealing with vocabulary (Laufer, 1997).

2.2.2 More recent models of vocabulary knowledge. Another important attempt to build a framework of vocabulary knowledge was Nation's (1990) often-cited work in his book Teaching and Learning Vocabulary. Nation's (1990) framework was based on second language learners' needs both receptive and productive and whether the learner needs to use the word in a receptive manner, in reading, for example, or also for productive situations, as in speaking or writing. Nation (1990) provided a list of questions about what a language user should know about a word, and distinguished receptive knowledge from productive one while doing so, as seen in Table 2.1. However, while Nation (1990) started from the language learners' needs when discussing vocabulary knowledge, the shift is clear toward a NS's vocabulary knowledge as the basis of building his model. This could be shown in his comparison to NS's limited productive knowledge in certain contexts, or his recurring statements of what "we know" (Nation, 1990, p.32) about words. Additionally, one limitation of Nation's (1990) framework is his mix of description of word knowledge and teaching implications for specific types of knowledge.

However, these limitations do not overshadow the importance of Nation's (1990) work on drawing a picture of vocabulary knowledge since this model is often-cited across the research on vocabulary knowledge.

Table 2.1

Knowing a word

Form		
spoken form	R	What does the word sound like?
	P	How is the word pronounced?
written form	R	What does the word look like?
	P	How is the word written and spelled?
Position		
grammatical patterns	R	In what patterns does the word occur?
	P	In what patterns must we use the word?
collocations	R	What words or types of words can be expected before or after the word?
	P	What words or types of words must we use with this word?
Function		
frequency	R	How common is the word?
	P	How often show the word be used?
appropriateness	R	Where would we expect to meet this word?
	P	Where can this word be used?
Meaning		
concept	R	What does the word mean?
	P	What word should be used to express this meaning?
associations	R	What other words does this word make us think of?
	P	What other words could we use instead of this one?

Note: In column 2, R = receptive knowledge, P = productive knowledge. (Nation, 1990, p. 31)

Nation (2001) expanded his model of word knowledge and rearranged it to account for developments in the field. Nation's new model still depended on receptive and productive distinctions throughout his list. His new model is formed around *Form, Meaning,* and *Use,* and

now includes *words parts, concepts,* and *references,* as shown in Table 2.2 (Nation, 2001). Another important aspect in Nation's (2001) model of word knowledge is his discussion of modeling word knowledge as a process, and its teaching implications.

Table 2.2

Revised Model of Vocabulary Knowledge

Form		
spoken form	R	What does the word sound like?
	P	How is the word pronounced?
written form	R	What does the word look like?
	P	How is the word written and spelled?
word parts	R	What parts are recognizable in this word?
	P	What word parts are needed to express the meaning?
Meaning		
form and meaning	R	What meaning does this word form signal?
	P	What word form can be used to express this meaning?
concepts and references	R	What is included in the concept?
	P	What items can the concept refer to?
associations	R	What other words does this word make us think of?
	P	What other words could we use instead of this one?
Use		
grammatical functions	R	In what patterns does the word occur?
	P	In what patterns must we use the word?
collocations	R	What words or types of words occur with this one?
	P	What words or types of words must we use with this one?
constraints on use (register, frequency ...)	R	Where, when, and how often would we expect to meet this word?
	P	Where, when, and how often can we use this word?

Note: In column 2, R = receptive knowledge, P = productive knowledge. (Nation, 2001, p.26)

It is evident in this review of vocabulary knowledge models that formulaicity of language is a component of great importance to any linguistic analysis. This focus on formulaicity of language, in its various manifestations, is important to sitiuate any research on formulaic

language in its appropriate context. Moreover, it is important to note that all vocabulary knowledge models attempt to deconstruct such vocabulary knowledge for reasons of convenience and clarity of researchers and learners alike (Schmitt, 1997). The types of vocabulary knowledge are separated to facilitate discussion and research, while in reality, such a construct is an "integrated whole" of an intertwined nature (Schmitt, 1997, p.4). Moreover, it is clear from the reviewed vocabulary knowledge models that there is a great emphasis on collocations and the environment surrounding words, which links those models tightly to any research on formulaic language. Those models were drawing the attentions to many aspects of vocabulary knowledge that are related to formulaicity of language. The following discussion reviews the use of corpus methodology in linguistic research.

2.3 Linguistic Corpus Studies

The following section provides an overview of the development of corpus studies and their contributions to linguistic research. Since this study relays heavily on corpus methodology, it seems appropriate to review the evolution of corpora studies and their addition to the field of linguistics.

2.3.1 The first corpora in linguistics. Building language corpora for research purposes was not the fruit of invention of the computer; it actually preceded that by a long time (Francis, 1992). Francis (1982, cited in Francis, 1992) defined the linguistic corpus as "a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis" (p.7). Francis (1992) identified three types of linguistic corpora that predated the computer, lexicographical corpora used to make dictionaries, dialectological corpora used to make dialect atlases, and grammatical corpora used for grammatical analysis.

The following paragraphs will provide an overview of those types of corpora research preceding computerized corpora.

Johnson's Dictionary of the English Language (1755, cited in Francis, 1992) was, according to Francis (1992), the first to organize and put forth a system of data collection for his work. Johnson's work was the result of gradual development of data collection and methodology, made by many of his predecessors (Francis, 1992). Some of the significance of this dictionary was related to publishing a plan for it in 1747, seven years before his work came to fruition (Francis, 1992). Evidence of building corpora for dictionary use is also found in Bailey's Universal English Dictionary (1731, cited in Francis, 1992), published before Johnson's. However, Bailey did not explain in detail how he went about building his corpus and collecting the data like Johnson had done (Francis, 1992). More recent dictionaries such as Oxford English Dictionary, completed in 1928, and Merriam-Webster, completed in 1934, have shown that improvements of data collection and systematic approach were taking place as in using volunteers or paid lexicographers in compiling the corpora (Francis, 1992).

Following the development occurring in lexicography, dialects studies adopted similar methodologies using volunteers and paid researchers for creating small specific dictionaries of dialects (Francis, 1992). Such publications were not introduced until the 19th century, partly since dialects were not considered worthy of study (Francis, 1992). An example of such studies was Wright's (1989-1905, cited in Francis, 1992) English Dialect Dictionary and English Dialect Grammar, in which he adopted the practices of Murray in Oxford English Dictionary (Francis, 1992). Another example was Ellis's (1889, cited in Francis, 1992) book: The Existing Phonology of English Dialects, in which his work on collecting data for his corpus took a span of twenty years (Francis, 1992). Ellis's book, similar to Murray's OED, utilized the cooperation of

volunteers, and that amounted to 811 volunteers sending him information and lending assistance (Francis, 1992). In the nineteenth century, corpora work for dialectology research was clearly becoming more systematic and collecting material for corpora was becoming more precise (Francis, 1992).

Grammatical research was no different in that it also utilized corpora to advance its pursuits. Francis (1992) provides several examples of grammarians who collected uses of the language and based their grammatical analysis on those examples. However, the reliance on intuitions as the bases for selection and inclusion in corpora was a drawback, since it favors interesting and strange instances of language and overlooks the regular patterns of language (Francis, 1992). Also, the lack of recording devices resulted in more reliance on written language and left the spoken language underrepresented (Francis, 1992). Quirk's Survey of English Usage (1974, cited in Francis, 1992) attempted to address those issues by balancing written and spoken language in the survey corpora. Quirk also focused on choosing a representative sample of standard English by specifying criteria for participants and situations from which texts were used (Francis, 1992).

2.3.2 The introduction of computerized corpora. The move to computerized corpora was instigated by W. Nelson Francis and Henry Kucera in their project, the Brown Corpus of American English that was published in 1964 (Leech, 1987; Svartvik, 2007). Its pioneering systematic approach of text selection and size were followed suit by researchers building computerized corpora and analysis programs (McEnery & Wilson, 1996; Svartvik, 2007). Moreover, the availability of the Brown Corpus provided researchers with a source of linguistic material and a tool to experiment with corpus analysis (McEnery & Wilson, 1996; Svartvik, 2007). The Brown Corpus consisted of 500 texts averaging around 2000 words each, taken from

15 written categories (Leech, 1987). The British equivalent of the Brown Corpus, the LOB Corpus, was another widely distributed corpus that was published in 1978 and it followed the same exact structure of Brown Corpus (Leech, 1987). Following Brown Corpus footsteps, the Survey of English Usage Corpus for British English was built with a similar structure of texts and size (Svartvik, 2007). An important advancement of this corpus was its inclusion of spoken language alongside the written language (Svartvik, 2007).

This move was in alignment with the survey's goals of creating representative samples that "describe the grammatical repertoire of adult educated native speakers of British English" (Svartvik, 2007, p. 16) in a variety of settings. Furthermore, Brown Corpus and the Survey of English Usage were not only pioneers in building the first computerized corpora, but also in establishing a field of research, that is, corpus linguistics (McEnery & Wilson, 1996; Svartvik, 2007). At the time, corpus linguistics was surrounded by hostile environment in the 1960s to the point of being described as intellectually discredited (McEnery & Wilson, 1996; Svartvik, 2007). Researchers from both projects were faced with skepticism of their ability to use computers and piles of texts to produce sound linguistic analyses (McEnery & Wilson, 1996; Svartvik, 2007). Lastly, by embarking on such big projects, generations of linguists were trained to work on corpora. These people later became, themselves, pioneers and leaders in the same field (McEnery & Wilson, 1996).

The advancement of computerized corpora. Following the seminal beginnings of corpus linguistics, by the 1980s, corpus studies were gaining traction and because of successful projects, researchers were interested in using empirical data in linguistic research (Renouf, 2007). Around that time, corpora projects were no longer satisfied with 1 million-word corpora and aspiration for more encompassing corpora resulted in projects covering 20-500 million words, as in the

Birmingham and the Bank of English corpora (Renouf, 2007). Such a tangible change was prompted by new questions that required larger corpora for answers, and by technological advancements that made managing and analyzing large corpora a feasible option (Renouf, 2007). However, with the rise of mega-corpora, the question about representativeness also arose, since those big corpora were designed to draw a picture of language (Renouf, 2007). As Renouf (2007) had shown, it was a common belief, in the 1980s, that a true representation of the language was an unreachable one, and that the best researchers could do is to try to form criteria of selection that is justified. Questions of corpora representation, as Renouf (2007) argued, are still relevant and evident in current corpus projects. For instance, Renouf (2007) cited several projects dating from the 1990s and continuing through 2000s that expressed the importance of designing representative corpora, as in the corpus of spoken Israeli Hebrew, presented in 2004. This is not strange considering that any validity of the results of corpus analysis is dependent on accurate representation of the target language. Thus, the issue of representation of corpora is an essential concern for the current study, and any corpus investigation that seeks applicable results.

2.3.3 Types of linguistic corpus research. The authentic material that corpus studies presented, and the empirical analysis that it has offered opened the gates for old and new venues of research in linguistics. Nowadays, wide range of linguistic investigations employ corpus analyses as its main methodology of choice; the following paragraphs will review some of those investigations.

Syntactic analysis and corpus methodology. The research on the syntactic nature of language has been very fruitful for corpus linguists, and has attracted most of the corpora research (McEnery & Wilson, 1996). Corpus addition to the syntactic field is evident on many fronts such as the representation of authentic material for analysis, the reliance on actual usage

rather than intuition, and the ability to study linguistic variation (McEnery & Wilson, 1996). This can be seen on macro-scale and micro-scale projects of grammatical analysis. One of the important macro-scale projects of grammatical studies is Longman Grammar of Spoken and Written English (Longman Grammar) (Biber et al., 1999). The project is based on the analysis of four registers: conversation, fiction, newspapers, and academic language, totaling around 20 million word corpora (Biber, 2001). The grammatical analysis of different registers in Longman Grammar proved important: for example, the use of modal verbs such as *could* and *would* was found twice as common in conversation as in academic English (Biber, 2001). Similarly, analyzing progressive aspect, as in *my husband is always telling me that,* and simple aspect as in, *do you work at GE,* in conversations showed that the simple aspect was 20 times as common as the progressive aspect (Biber, 2001). This contradicted the common assumption that the progressive aspect was the unmarked type in conversations, meaning that it occurred more than the simple aspect (Biber, 2001). This clearly had a pedagogical influence on what aspect was presented as the predominant one in ESL textbooks, when, in reality, such assumption was unfounded (Biber, 2001).

Lexical studies in corpus linguistics. Lexical studies were no strangers to corpora work as it was shown earlier. However, the use of corpora in current research in the field of lexical studies has expanded a great deal. For example, the research on word lists, collocations, and lexical bundles provides an insight about the importance of corpus methodology in shaping the field. Furthermore, Coxhead (2000) analyzed a 3.5 million-word corpus of academic text to investigate the frequency and range of vocabulary in academic texts. Coxhead reached 570 word families that covered 20% of academic texts (AWL), regardless of their registers. Nation (2004), also, conducted a similar project by creating three 1000 word family lists out of the British

National Corpus (BNC) that accounted for frequency, range, and dispersion. Nation (2004) found a great deal of overlap between the AWL, the General Service List (GSL) by West, and his BNC 3000 word list. Nation's (2004) conclusion indicated that the age of the GSL is not a reason for its abandonment but rather respect, and that learners' goals should guide the list's selection.

Discourse analysis. Another area of linguistic research that has employed corpus methodology is discourse analysis. For instance, Conrad (2002) showed that corpus research goes far beyond concordancer lists, which list results of researchers' inquires, to provide an analytical view of language and its registers, e.g., conversations, newspapers, or academic prose. Moreover, Conrad (2002) also noted that corpus analysis of discourse put more value on empirical evidence, even if it was initially guided by impressions and intuitions. Another feature that is evident in corpus discourse analysis is the use of both quantitative and qualitative analyses (Conrad, 2002). Examples of this type of research in the characterization of registers and identifying their salient features could be found in Biber's (2006) stance analysis in written and spoken academic registers, Cortes's (2004) comparison of lexical bundles in academic articles in history and biology, and Biber et al.'s (2004) investigation of lexical bundles in university teaching and textbooks. After reviewing the development of corpus linguistics, the following section reviews the advent and the development of formulaic studies in linguistics research.

2.4 Collocations, Formulaic Language, and Lexical Bundles in Lexical Studies

Before reviewing the research on formulaic language, it is appropriate to point out that while there are many terminologies suggested in this field, the terms *formulaic language* and *formulaic sequence* appear to be proper terms that encompass those varying terms. These terms, suggested by Wray (2002), are general to the extent that they could cover many concepts related

to formulaicity of language while at the same time denoting a very clear type of linguistic research. The terms *formulaic language* and *formulaic sequences* would include lexical bundles as one approach of analyzing language formulaicity.

2.4.1 Early research on collocations. The study of lexical collocations is evident in the linguistic research from the early 1900s. Jespersen (1917) in his investigation of negation clearly emphasized several notions such as combination of words, frequent collocations, and the frequency of occurrence in texts. Moreover, Palmer (1922) introduced what he called "memorized matter and constructed matter" (p.116) as one of his principles of language study. Memorized matter, according to Palmer (1922), is what language users recall from their memory, in other words, pre-fabricated, while the constructed part is what is made up at the moment of production. Palmer (1933) advanced the work on collocations and went to define collocations as those "successions of words which (for various reasons) are best learnt as integral wholes" (p.8). Palmer (1933) referred to the term collocations as "technical but conveniently vague term"(1933, p.7), when used in the linguistic field.

In 1950s, Firth (1957) introduced collocations as a technical term that described an essential part of the meaning of words. Firth (1957) distinguished meaning by collocation from the conceptual meaning of a word and described meaning by collocation as the knowledge of what habitually collocates with a word. For instance, Firth (1957) stated that "one of the meanings of *night* is its collocability with *dark*" (p.196). Moreover, one of Firth's (1957) notes about using collocations in filling the blanks games is seen as a method of testing non-native speakers' knowledge of collocation in more recent research.

However, an apparent characteristic in the early research on formulaic language is the reliance on impressions and subjective opinions. For example, Firth (1957) used expressions

such as “Cursory examination” (p.203), “we note” and “which to me seem glaringly obsolete” (p.204) to describe the selection processes employed while conducting his analysis. On that nature of analysis, Hakuta (1974) discussed the difficulty of manual analysis and how “many prefabricated patterns will escape without positively being identified” (p.289). Also, Pawley and Syder (1983), in their work on native-like selection and fluency, produced lists of clauses and phrases that they consider formulaic in certain geographic locations based on mere reflections. Pawley and Syder (1983) gave the example of *I know what you mean* as one of the longer familiar and memorized sequences used in everyday speech.

Despite not using empirical methods of data analysis, such early research of word combinations paved the way for important advancements and insights about the nature of language use. For example, Firth’s (1957) early work showed how collocations occupy an important part of our knowledge of lexical items. Moreover, Hakuta’s (1974) established that using prefabricated patterns is an important strategy for NNS of language to achieve fluency, and that it is essential for receptive and productive linguistic development. Similarly, Ferguson’s (1976) work on politeness formulas, also, called for more attention to such formulas and their importance for non-native speakers. Lastly, Pawley and Syder’s (1983) work demonstrated the importance of pre-fabricated sequences in forming a big chunk of everyday language. Their work, also, shed the light on how language is processed and how memorized clauses are used to facilitate a fluent conversation.

2.4.2 Recent research on formulaic sequences. With more research being devoted to formulaic sequences and prefabricated language, every researcher is faced with the issue of defining and approaching such units. As Wray (2002) stated, in the beginning of her book dealing with formulaic language, that “research on formulaic language has lacked a clear and

unified direction, and has been diverse in its methods and assumptions" (p.4). This statement was echoed by many researchers that many rubrics and terms have been used to investigate related, and sometime identical concepts (Biber, 2006; Biber & Barbieri, 2007; Nekrasova, 2009; Liu, 2012). Wray (2002) also warned that while there are some instances of careless assignments of terminology, it is important not to fall in the trap of thinking that most terms denote the same phenomenon, and that in most cases, there are some technical and theoretical aspects behind the terminology used in this field. It is, perhaps, the lack of uniformity among researchers in their usage of terms and their tendency to coin new terms to describe phenomena that resulted in more than sixty terms being used to discuss formulaic language (Wray, 2002). Moreover, Wray (2009) stated in a reflection on formulaic language research that the field is in a stage of "consolidation and confirmation" (p.2), and at such stage several established notions might be revised or rearranged.

With the different approaches studying formulaic language, it is imperative to review those approaches of analysis and their theoretical stances. The review will not have a comprehensive structure that covers its themes due to the nature of the field, but rather it will try to cover the major themes and the prominent names and their work in the field.

Intuition as the identifying factor. The first identifying method that has been used to study the formulaicity of language is linguists' intuitions. For a long time, linguistic research relied on linguists and linguistic community's intuitions, and the field of formulaic language did not break this tradition, at least in its initial stages. For instance, Moon (1997) showed that formulaicity of an expression is a result of a community's institutionalization. Moreover, Wray (2002) argued that researchers "often are self-appointed arbiters of what is idiomatic or formulaic in their data" (p.20). Not only that, intuition, also, had influence even when other

measures are used in research, as in the interpretations of results or in the choice what example are published (Wray, 2002). For example, Bahns, Burmeister, and Vogel (1986), after elaborating on the limits of using empirical methodology for choosing formulas, they "eventually listed a number of expressions which we intuitively regarded as formulas" (p.700). However, Sinclair (1991), in his call for a new linguistic methodology, showed that how results from corpus analysis proved that intuitive linguistic analysis was not supported by actual evidence. This was echoed by Francis's (1993) call for overhauling what we know about the language on the basis of what corpora shows.

Multiple identifying factors of formulaic sequences. Moreover, another school that attempted to analyze and categorize formulaic sequences was led by Becker (1975) and later Nattinger and DeCarrico (1992). Becker (1975) was concerned that while language is occupied by preconfigured sequences that are sewed together to form spoken and written texts, the focus was given to the lexical items on the word level or even smaller. Another area of concern to Becker (1975) was that analysis at the phrasal level demoted most phrases under scope of research and dealt with them as idioms. Additionally, the methodology of this school used a mix of identifiers such as formal and functional aspects, all guided by intuition and observation. To Becker (1975), it is clear that lexical phrases "are actually observable" (p.62), and that speakers and writers "will feel them popping of [their] own brains when [they] speak and when [they] write" (p.62).

Becker's (1975) classification of lexical phrases yielded six categories: polywords, phrasal constraints, meta-messages, sentence builders, situational utterances, and verbatim texts. Those categories vary in their nature between some formally-based, e.g., polywords and phrasal constraints, and functionally-based, e.g., situational utterances (Wray, 2002). However, Becker

(1975) acknowledged that such attempt to classify lexical phrases is rather a "messy taxonomy" (p.63), and that it speaks to the complex nature of language.

Continuing Becker's (1975) line of research, Nattinger and DeCarrico (1992) embarked on a mission to refine and expand the lexical phrase model and use it as a mean for teaching language. Nattinger and DeCarrico (1992) defined lexical phrases as "chunks of language of varying lengths" (p.1) that have salient formal, functional, and statistical characteristics. Moreover, Nattinger and DeCarrico (1992) classified lexical phrases by modifying Becker's (1975) classifications and described their members through several formal criteria. Those formal criteria were grammatical level, i.e., sentence and word levels, canonical vs. non-canonical, fixedness and variability, and lastly their continuity vs. discontinuity (Nattinger & DeCarrico, 1992). Moreover, Nattinger and DeCarrico's (1992) classifications were composed of four types: polywords, e.g., *in a nutshell*, institutionalized expressions, e.g., *a watched pot never boils,* phrasal constraints, e.g., *as I was saying/mentioning,* and sentence builders, e.g., *I'm great believer in [setting money aside/ exercising]*. However, throughout their work, Nattinger and DeCarrico (1992) suggested that it is best to think of those classifications as points in a continuum in terms of fixedness and continuity.

Moreover, working on form and function axes, Nattinger and DeCarrico (1992) functionally grouped those lexical phrases to facilitate pragmatic and discoursal analyses, and even more important for their research, pedagogical analysis. Nattinger and DeCarrico (1992) listed three functional categories of lexical phrases: social interactions, necessary topics, and discourse devices, all provided with lexical phrases that fit in functional sub-categories. Moreover, register and genre are evident elements in Nattinger and DeCarrico's (1992) analysis of lexical phrases, as in their analysis of conversational lexical phrases, and their comparison

between spoken and written discourse devices. However, despite their lengthy discussion of formal and functional differences among varying registers, it is not clear as to what methodology was used to identify such differences. For example, Nattinger and DeCarrico (1992) go to great lengths to analyze differences between lexical phrases of maintenance in written and spoken registers, but do not mention whether those differences were the fruit of intuitions, empirical analysis, or previous research.

Despite the shortcomings of Nattinger and DeCarrico's (1992) work, their work is seminal in that it brought formulaic sequences to the forefront of language teaching. The first half of their book is written with pedagogical theories and implications in mind while the second half is entirely devoted to teaching lexical phrases. This covered the four skills normally targeted in ESL classes, speaking, listening, reading, and writing (Nattinger & DeCarrico, 1992). The evidence of the importance of their work is clear in that most of the research on formulaic sequences citing their book.

Another project that has put formulaic language at the center of language pedagogy is Lewis's (1997) lexical approach, first introduced in 1993. The lexical approach was an answer to the inadequate reliance on traditional grammar and vocabulary axes for linguistic and pedagogical analysis (Lewis, 1997). According to Lewis (1997), language is composed of four different types of lexical items that form linguistic chunks. Those chunks fall across a "generative spectrum" and vary from fixed to free chunks (Lewis, 1997). The lexical items that Lewis (1997) proposed are dependent on social validation from specific communities, thus accounting for the imminent variations between social groups. Lewis (1997) classified lexical items into four groups, words and polywords, collocations, institutionalized utterances, and sentence frames and heads. Moreover, formulaic sequences dominate most of those groups

except for single words in the words and polywords category, a clear evidence of its valuation within the lexical approach (Lewis, 1997).

However, Lewis's (1997) classifications were not a clear cut, a fact Lewis acknowledged and rather embraced as a positive aspect in that it provided several angles of analysis for the same lexical items. For instance, Lewis (1997) used the several identifiers to assign lexical phrases to groups, such as having a referential meaning, appearing in dictionaries, or being idiomatic to identify polywords like *on the other* hand. Among all categories, the inconsistency of the identifying methodology is clear in that functional, formal, and discoursal aspects were not applied across all categories.

The most evident aspect in the lexical approach was Lewis's (1997) emphasis on the importance of formulaic language as a vehicle for teaching language. In the lexical approach, individual words are given a minimum attention, and even less attention to grammar. What is important according to Lewis (1997) is to raise awareness of the language, and to present the language in its natural form, formulaic and chunked, which leads to successful acquisition of language. Lewis (1997) argued that presenting an utterance such as *If I were you, I'd go/pick that one* in a traditional grammatical view would yield two conditional clauses, a challenging structure to teach. However, according to Lewis (1997), lexical analysis would present it as a chunk composed of two parts, with the free part coming after *I'd*, a much easier and simpler structure. For Lewis (1997), the ability to analyze language grammatically does not necessitate teaching language through that analysis.

However, the attempts to use formulaic language as a means for teaching language are not taken without criticism. The work of Nattinger and DeCarrico (1992), Lewis (1993), and others was prompted by the growing status of formulaic language and corpus analysis, yet it

failed to describe an approach to teaching language through formulaic language. Richards and Rodgers (2001) argue that such proposals only attain to "one component of communicative competence" (p.138). Moreover, for a lexically based theory of teaching language to be considered an approach, it need to covers a wide range of aspects, including syllabus design and teaching procedures (Richards & Rodgers, 2001). According to Richards and Rodgers (2001), those proposals are starting points "in search of an approach and a methodology" (p.138).

On another front, Wray (2000) questioned the assumptions made by Nattinger and DeCarrico (1992), Lewis (1993), and others, that exposure to formulaic language is sufficient to gaining control of linguistic structure. As Wray (2000) put it, native speakers of language use formulaic language to, among other things, reduce the processing pressure on both speakers and hearers. It is, thus, unreasonable to expect non-native learners of language to activate their analytical ability when exposed to formulaic sequences to master its internal structure deductively, when in fact formulaic language bypasses such analytic process (Wray, 2000). According to Wray (2000), the pivotal issue is that such approaches to teaching language assumed similarity between L1 and L2 acquisitions of language, and of formulaic language in particular.

2.4.3 The lexical bundles approach. One of the most prominent analyses of formulaic sequences is Biber and colleagues' use of frequency as the identifying criteria of what they called "lexical bundles" (Biber & Conrad, 1999; Biber et al., 1999; Biber et al., 2004). Their lexical bundle approach was intended as an exploratory quest to find the most frequent sequences, and how these frequent sequences differ from one register to another (Conrad & Biber, 2004). Moreover, this line of research has relied on frequency of lexical bundles, their continuous nature, and their length (Conrad & Biber, 2004). The justification of those criteria is

that frequent uninterrupted expressions will be used as integral wholes, and that lexical bundles longer than two words are more susceptible to be used for discoursal functions, thus more essential for discourse (Conrad & Biber, 2004). Within this approach, lexical bundles do not have to have complete structures, an aspect that is beneficial in identifying important sequences that are less likely to be singled out by observers intuitively (Conrad & Biber, 2004).

Moreover, there is an overlap of some extent between lexical bundle research and metadiscoursal analysis research, an overlap most likely related to the functional nature of both lines of research. For example, the functional analysis is evident in most of the research conducted on lexical bundles, which is used to group lexical bundles according to functional value within discourse. Similarly, Hyland and Tse (2004) stated that metadiscourse is a functional category that is utilized to handle writer-reader interactions and textual organizations. However, there are clear distinctions between lexical bundle research and metadiscourse research. For instance, metadiscourse research separates metadiscoursal content from propositional content, which is considered the core content of a discourse (Hyland & Tse, 2004). Such distinction is not regarded in lexical bundle research, in which computer analysis present frequent sequences that could be metadiscoursal or propositional in the eye of metadiscourse research. Moreover, the selection of units of analysis is different in both lines of research. For instance, while lexical bundle research relies on pre-selected lengths and frequency rates, metadiscourse analysis focuses on function more than the limits and criteria for selection. This is evident in Hyland and Tse's (2004) analysis which presented several examples of metadiscourse units that vary in length from one word, one clause, one sentence, to an entire paragraph. Based on such variation, they argued that there are no simple criteria that could be used to identify metadiscoursal units. However, despite the varying methodologies, the functional analyses in

both types of research show overlapping of concepts and borrowing from one line of research to another.

Furthermore, the lexical bundle methodology of extracting sequences from corpora was not pioneered by the work of Biber and his colleagues. In fact, Conrad and Biber (2004) cite Altenberg (1993, 1998, cited in Conrad & Biber, 2004) and Butler (1997, cited in Conrad & Biber, 2004) who used this methodology on analyzing spoken English texts and Spanish texts respectively. However, the coinage of the term "lexical bundle" is first found in Biber et al. (1999) Longman Grammar. To identify lexical bundles, the researchers set a semi-arbitrary frequency cut-off point where bundles recurring more or at that point are included in their analysis (Conrad & Biber, 2004).

Moreover, lexical bundles as Biber and Barbieri (2007) defined them are "the most frequently recurring sequences of words" (p.264). With that definition, this line of research distances itself from research focusing on collocations and idioms. According to Biber and Conrad (1999), idioms carry meanings that are not predictable by their lexical components, and they are mostly structurally complete, criteria that are disregarded in lexical bundle research. Another distinction between idioms and lexical bundles is that while idioms are salient expressions, lexical bundles are mostly common and simple expressions that occur at a certain frequency (Biber & Conrad, 1999). Similarly, collocations are statistical relations between two words that might influence the meaning of such words while lexical bundles are strings of more words that tend to co-occur in a given text as a whole (Biber & Conrad, 1999).

The methodology in studies within the lexical bundle approach carries small variations that adapt to each individual investigation while holding the general guidelines. For instance, one fixed aspect in this methodology is its reliance on frequency as the only identifier of lexical items

(Biber et al., 1999; Biber & Conrad, 1999; Biber & Barbieri, 2007). However, the rate at which the cut-off is set varies across studies. For example, one of the low cut-off rates is found in Longman Grammar (Biber et al., 1999) at a rate of 10 per million words (PMW), which was also used in Simpson-Vlach and Ellis's (2010) work on the academic formula list. From that point, we find investigations setting 20 PMWs cut-off (Biber and Conrad, 1999; Hyland, 2008; Liu, 2012), 25 PMWs (Ädel & Erman, 2012), and 40 PMWs (Biber, Conrad, & Cortes, 2004; Biber & Barbieri, 2007). The Lowest cut-off point is probably McCarthy and Carter's (2006) 4 PMWs, which was used on a 5 million-word corpus. The low number was selected to accommodate for the longer sequences of six-word length (McCarthy& Carter, 2006).

Additionally, another fixed aspect is the reliance on multi-word sequences as an area of interest while allowing for different lengths of lexical bundles. For instance, Biber et al. (1999) investigated three-word, four-word, five-word, and six-word lexical bundles. Their investigation of longer lexical bundles prompted using a lower cut-off rate of 10 PMWs. A similar investigation is McCarthy and Carter's (2006) analysis of sequences from two-word to six-word lengths. McCarthy and Carter (2006) notes that is not practical to look for lexical bundles longer than six words, even when using a very liberal cut-off of 4 PMWs. It is evident, however, that most of lexical bundle research is concentrating on four-word lexical bundles as in Biber et al. (2004), Cortes (2004), Biber and Barbieri (2007), and Ädel and Erman (2012) to cite a few. This choice is justified in that four-word lexical bundles will show longer lexical bundles, e.g., *if you know what, you know what I,* and *know what I mean* (Biber et al., 2004), as will as shorter lexical bundles, e.g., *as a result* and *as a result of* (Cortes, 2004; Biber et al., 1999).

2.5 The Lexical Bundle Approach to Formulaicity of Language

After reviewing the methodology used in the lexical bundle approach, it is appropriate to review one of the prominent features of lexical bundle research, its attachment to analysis of registers. This attachment is evident in the operational definition of lexical bundle used in Biber *et al.* (1999): "lexical bundles are identified empirically, as the combinations of words that in fact recur most commonly in a given register" (p.992). In this sense, by definition, lexical bundles are theoretically attached to assigned registers. Utilizing this feature, Biber et al. (1999) compared two different registers, conversation and academic prose, to show "the most striking differences in language use" (p.990).

The use of register analysis as one angle of lexical bundle research has opened new territories in different fields of linguistic research. For example, from an EAP and an ESP perspective, several studies investigated the use of lexical bundles between academic registers and other registers, and among different academic registers. For example, Biber et al. (2004) compared the functional use of lexical bundles in university classroom teaching and in academic textbooks. Their investigation revealed that lexical bundles were more evident in university classrooms than in textbooks and conversations (Biber et al., 2004). Moreover, they also found that classrooms take some characteristics of conversations (i.e., using stance bundles), and some characteristics of academic prose (i.e., using referential bundles) (Biber et al., 2004). A similar analysis of language use in the academic environment looked at lexical bundles' use in different academic registers such as course management, advising, and instructional registers (Biber & Barbieri, 2007). Biber and Barbieri (2007) findings showed that lexical bundles were more prevalent in non-instructional registers (e.g., course management), and that lexical bundles were more common in written registers, contrary to previous findings in Biber et al. (2004).

Closely related to register analysis, discourse analysis is another sub-field of linguistic research that relates directly to lexical bundles investigation. Although one could argue that any investigation of lexical bundles in any register is a form of discourse analysis, some investigations of lexical bundles were using lexical bundle methodology for discourse analysis purposes. One important example of this line was the seminal and oft-cited work of Biber *et al.* (1999) comparison of conversation and academic prose, discussed earlier. Biber and Conrad (2001) expanded that work to show that, by using lexical bundles methodology, similar lexical items functioned differently in different registers.

2.5.1 Lexical bundles and formulaic sequences in EAP and ESP research. The fields of EAP and ESP have seen a growth in the amount of research looking into formulaic language in general, and lexical bundles in particular. This interest from researchers has covered many fronts related to teaching English, including receptive and productive aspects of language, and investigating different registers of language to look into their formulaic nature. The Following paragraphs will review some of this research and identify its salient aspects.

Comparing different registers for EAP and ESP. One major theme of lexical bundles research is concerned with comparing different registers within academic contexts. For instance, Conrad and Biber (2004) analyzed lexical bundles in academic prose and conversation. Their work was carried out using parts of Longman Spoken and Written English Corpus totaling more than nine million words. Moreover, their analysis found lexical bundles more apparent in conversations than in academic prose, with conversation lexical bundles serving personal and stance functions mostly, and academic lexical bundles serving referential functions (Conrad & Biber, 2004). Building on this work, Biber et al. (2004) expanded the scope of the study and compared two academic registers: classroom teaching and textbooks using two million word

corpora. Their investigation showed that classroom language was the most formulaic type when compared to conversations and textbooks language (Biber et al., 2004). Moreover, classroom language yielded four times lexical bundles as many as textbooks yielded while taking characteristics from conversational language, i.e., declarative and interrogative clauses, and from written academic language, i.e., noun and prepositional phrases (Biber et al., 2004). One important aspect of Biber et al.'s (2004) work was their functional categorization of lexical bundles into three types: referential bundles that refer to "physical or abstract entities" (p.384), stance bundles that denote attitudes and judgments, and lastly discourse bundles that are used to organize and negotiate discoursal moves. This model is adapted in this study to analyze the lexical bundles extracted from the two corpora. More on this model of categorization is provided in Chapter III.

However, Biber and Barbieri (2007) carried out a similar analysis of university spoken and written registers which yielded different results regarding lexical bundles. Biber and Barbieri's (2007) work was based on corpora that included new contexts for both written and spoken registers, such as office hours, course management, university catalogs, syllabi, and service encounters. The findings from this study were not aligned with previous research in that non-academic written registers, e.g., institutional writing and course management, showed more lexical bundles than textbooks and academic prose, and even more than university spoken registers. Moreover, among spoken registers, class management and service encounters showed more lexical bundles than classroom teaching. Both of these findings contradicted with findings from Biber et al. (2004) about lexical bundles being more evident in spoken registers, and being more evident in classroom language.

Another analysis of university spoken registers was Neely and Cortes's (2009) analysis of university lectures and students' presentations and dissertation defenses. The goal of Neely and Cortes's (2009) analysis was to identify differences in functions in order to inform EAP listening syllabus design. Their findings showed that the variations in functions among lecturers and students call for some flexible categorization when introducing lexical bundles to students rather than labeling them with one primary function (Neely & Cortes, 2009). Additionally, the authors utilized their findings in constructing corpus-based activities that introduce lexical bundles in listening classrooms (Neely & Cortes, 2009).

Comparing different discourses. Moreover, an important theme within lexical bundles studies is to compare lexical bundles found in different discourses. Researchers have shown a great interest in investigating discoursal variability due to its importance and applicability in EAP and ESP environments. For instance, Byrd and Coxhead (2010) analyzed a 3.6 million words corpus that consisted four academic disciplines: arts, commerce, law, and science. The findings of Byrd and Coxhead's research identified 73 lexical bundles that occurred in all disciplines, shown with their frequency of occurrence in each discipline. Also, another important investigating of disciplinary variation of lexical bundles is Hyland's (2008) in which he compared lexical bundles in four different disciplines.

Hyland's (2008) work was based on 3.5-million-word corpus that included research articles, doctoral dissertations and master's theses chosen to represent electrical engineering, microbiology, business studies, and applied linguistics. Moreover, the analysis in Hyland's (2008) investigation consisted of categorizing lexical bundles based on their formal and functional characteristics, similar to Biber et al. (2004) analysis of lexical bundles in university teaching and textbooks. The analysis yielded some interesting findings, some of which were

aligned with previous research. For example, the lists of lexical bundles of each discipline show that at least half of the lexical bundles occurring in each one are unique; many of them were some of the most recurring bundles within the discipline (Hyland, 2008). Similarly, the lexical bundles found in four disciplines have distinct formal characteristics, making it possible to draw some patterns about formulaic language in each field (Hyland, 2008). Hyland's (2008) analysis showed the difficulty of establishing lexical bundles syllabus for general academic English, considering the distinctive and varying nature of lexical bundles in each discipline, and that only four bundles were found across four disciplines: *on the other hand, in the case of, as well as the,* and *the end of the*. Hyland's (2008) analysis is of relevance to the current study in that it focused on disciplinary variation of lexical bundles, which is closely related to this study's analysis of the engineering texts. It, also, established a reference point in its investigation of lexical bundles in the electrical engineering corpus, which provided a chance to compare findings from this study to its results.

Comparing learners' use of lexical bundles. Additionally, another salient theme in lexical bundles research in EAP and ESP is concerned with comparing ESL and EFL students' use of lexical bundles against native speakers of English. This line of research is mainly concerned with identifying nature of lexical bundles in non-native language and addressing deficiencies and gaps of lexical bundles' use. For instance, Cortes (2004) looked into lexical bundles found in history and biology published writing (target bundles) and compered students' use of such target bundles in both fields. Cortes's (2004) analysis was based on approximately one-million-word corpus for each field, and on approximately four hundred thousands words corpus for students writing in each field. Among other analyses and comparisons in Cortes's (2004) work, the comparison between bundles in published disciplinary readings and students'

writing revealed that most target bundles were never or rarely used, and that those that were found in students' writing did not align with similar functions of target bundles. Cortes (2004) suggested that exposure to frequent lexical bundles is not sufficient for students to start using them in writing, and that more effort should be placed on getting students to notice such frequent lexical bundles in their respective fields.

Moreover, Cortes (2006) conducted a follow-up study to find the effects of including explicit lexical bundles teaching in writing-intensive history class on students' written assignments. Cortes (2006) presented mini-lessons that were merged with the history class to cover the nature of lexical bundles, their functions, and activities targeting lexical bundles. The results of pre and post-analysis of students' writing did not show any significant improvements in terms of frequency of use of lexical bundles (Cortes, 2006). However, Cortes (2006) notes that the students reported an increased awareness of lexical bundles and that the mini-classes motivated them to further use of lexical bundles in future written assignments.

Furthermore, Ädel and Erman (2012) compared written assignments in linguistics by native-speakers of English and non-native speakers to investigate lexical bundles evident in their writing. For their corpora, Ädel and Erman (2012) used an 863-thousand-word corpus and a 247-thousand-word corpus for non-native speakers and native speakers of English, respectively. Ädel and Erman's (2012) results show that native speakers used a wider range of lexical bundles than that of non-native speakers, despite the fact the non-native group is an advanced one writing in linguistic topics. Non-native speakers, in Ädel and Erman (2012) study, used not only less lexical bundles, 115 bundles compared to 185 bundles, but also they used them with less variation. Moreover, non-native speakers displayed signs of register difficulties shown by the lexical items used within lexical bundles, e.g., *hard* vs. *difficult* (Ädel & Erman, 2012).

Another study that compared the use of lexical bundles among native and non-native speakers of English is Karabacak and Qin (2013) study which compared lexical bundles in argumentative papers written by Turkish, Chinese, and American university students. Their study collected term papers written by first and second-year university students about current and controversial topics such as violent video games and smoking in public (Karabacak & Qin, 2013). Karabacak and Qin's (2013) comparison showed that American students used lexical bundles more frequently than Chinese and Turkish students, attributing the discrepancy to insufficient knowledge of formulaic language, and failed attempts to use lexical bundles as in *In the U.S.,* which was written: **in U.S,* thus using a lexical bundle incorrectly.

Comparing EAP and ESP texts with disciplinary texts. One study of importance to the current study is Chen's (2008) comparison of lexical bundles in electrical engineering introductory textbooks and ESP textbooks. This study is important to the current investigation since it was the only study that resembled the methodology and the goals of the current study. Moreover, Chen (2008) compiled two corpora consisting of the aforementioned texts, and then compared the functional nature of lexical bundles in the two corpora. The goal of this analysis was to determine whether there was a gap in the functions of lexical bundles between the two types of texts. Chen's (2008) analysis concluded that there was a striking gap between the two corpora and that the ESP textbooks misrepresented the target bundles found in the engineering texts. It also concluded that the ESP textbooks lacked many functional types that were found in the engineering texts, as the directive and desire stance bundles.

The results of Chen's (2008) analysis were especially significant since the ESP texts used for the comparison were electrical engineering ESP textbooks that were designed around supposedly authentic materials. Chen (2008) stressed the importance of including lexical bundles

and their functions in EAP and ESP programs, and suggested using lexical bundles as identifiers of authenticity of texts in any specific field. However, there are a few differences between Chen's (2008) study and the current study. For instance, Chen (2008) study analyzed electrical engineering introductory textbooks, a specific sub-field of engineering, while this study analyzed a corpus of engineering texts compiled from several sub-fields since Pathway students go different engineering programs. Another difference between the two studies is that Chen's (2008) study compared the engineering texts to ESP textbooks designed for electrical engineering students while this study compare its engineering texts to texts used in an advanced writing class at INTOCSU pathway programs. Pathway programs are discipline-specific transitional programs that are designed to bridge the linguistic and academic gaps for students applying to CSU with insufficient qualifications. The following are the research questions and hypotheses:

2.6 Research Questions

1. What are the most frequent four-word lexical bundles in the Pathway Corpus and the Engineering Corpus?
2. Are there significant differences in the functions of the lexical bundles found in the corpora?

The answer to the first question is based solely descriptive data obtained by the concordancing program. The second question on the other hand is answered by means of inferential data.

2.7 Research Hypotheses

1. The lexical bundles from both corpora will display minimal overlap.

2. Null Hypothesis: There are no significant differences in the functions of lexical bundles found in the two corpora.

2.8 Chapter Conclusion

In conclusion, as the status of vocabulary received more attention in second language acquisition research, our understanding of vocabulary knowledge evolved and transformed tremendously. Models of vocabulary knowledge have been proposed to address the complexity of vocabulary roles in language learning and the elements that constitute vocabulary knowledge. Additionally, most of the models proposed recognize the formulaicity of language as one essential part of lexical knowledge, despite their varying takes on analyzing that formulaicity.

Moreover, among many approaches to language formulaicity, the lexical bundle approach stands out as a well thought-out approach of analyzing texts. The appeal of lexical bundles is based on several aspects of its design, namely its reliability, objectivity, and ability to identify common but important formulaic language. These futures are the products of a design that handles texts without any reliance on subjective judgments or intuitions but rather utilization of computing abilities to identify the most frequent strings of words in a text or a corpus.

Moreover, the increased focus of lexical bundle research opens new venues of linguistic analyses that were neglected before, one of which was related to the analysis of specific registers and genres. This was evident in the research reviewed in this chapter, which revealed some powerful and interesting perspectives of analyzing texts in specific contexts. Furthermore, from the inception of the lexical bundle approach, the utilization of lexical bundles research in EAP and ESP fields has provided some important insights about language, and what should be taught. As it has been shown in this chapter, lexical bundle analyses have pointed out some of the needs through comparing disciplinary texts and students writing, and through comparing and

contrasting variations in the disciplines. However, the research is lacking when it comes to analyzing formulaicity of EAP and ESP texts and disciplinary texts. After an extensive review of the literature, only one investigation was found to cover this gap in the research, despite its pedagogical importance. Thus, the current study try to add to the field of second language acquisition by extending this line of research, and it is hoped that the results of this study add to the body of knowledge and trigger the interest of other researchers.

More specifically, this study investigates the nature of formulaic language in texts used in one Pathway class and in engineering disciplinary texts that are presented to students after enrolling in their academic programs. Additionally, this study directs the attention to the language used in ESL programs and whether it reflects the formulaic nature of disciplinary texts in academic fields.

The following chapter will describe the methodology of this study. It reviews and describes the process of compiling the two corpora for this study. Further, the chapter reviews the analytical framework used in this investigation, including the concordancing process, the functional analysis of lexical bundles, and statistical analyses implemented for testing the significance of the results.

CHAPTER III

Methodology

This chapter presents the methodology and the analytical framework of the current study. First, a description of the corpora developed for the analysis is presented. Next, I describe the process of identifying lexical bundles from both corpora. Moreover, the chapter reviews the analyses chosen for this study, including the functional analysis, and the statistical tests of significance. Lastly, the chapter concludes with research questions and hypotheses.

3.1 The Corpora Used in the Investigation

3.1.1 The engineering corpus. The corpus built to investigate the nature of lexical bundles in language in the engineering academic contexts is compiled from textbooks and required readings for graduate and undergraduate engineering programs in CSU. The texts included in the corpus cover a wide range of subjects such as mathematics, applied mathematics, hydrology, structural analysis, steel construction, and water management. The list of texts was provided by experts from the engineering department. Moreover, the total number of words in the corpus is 1,264,106 words. As for the word counts by subject, Hydrology and Atmospheric Science covered 345,931 words, Construction and Structural Analysis covered 458,773 words, and Mathematics and Applied Mathematics covered a total of 459,402 words. A complete list of the corpus content is available in Appendix A.

3.1.2 The pathway program corpus. The corpus built to identify lexical bundles students are exposed to in the Pathway / INTO program at CSU was compiled from readings required for advanced academic writing class. Those readings were selected articles that deal with controversial issues such as globalization, peaceful resistance, organ donation, technology,

social equality, and war. Some of the texts were taken from a textbook required for course work (Sourcework), which constituted 27,000 words. The remaining texts were selected by INTO staff to accompany the textbook readings, amounting to 38,000 words, bringing the overall total for the pathway corpus to 65,000 words. The representation of this corpus to what pathway students are exposed to is very accurate since it represents almost all of the readings those students are exposed to at an advanced writing class. Moreover, the true representation of this corpus, despite its small size, provides an insight about the formulaicity of the language presented to students in the pathway program, and makes for a good tool to compare the nature of its formulaicity to those of disciplinary variations. A complete list of the corpus content is available in Appendix B.

3.2 Identifying Lexical Bundles

To identify the most salient lexical bundles in the corpus, the researcher uses the frequency of occurrence as the basis for selection. To facilitate this task, AntConc concordancing program by Anthony (2012) was used to identify the most frequently occurring lexical bundles. AntConc is a freeware concordancing program that offers comprehensive textual analysis options such as word lists, n-grams, collocates, and clusters for researchers and students (Anthony, 2012). In the case of lexical bundles, the concordancer scans the corpus word by word and stores the repeated instances of multiple-word bundles. The program, then, identifies the lexical bundles that occurred within the corpora by a rate at or above the cut-off number set by the researcher. The cut-off for this investigation was 40 occurrences per corpus for the engineering corpus, which equals 32 PMWs, and five occurrences per corpus for the pathway corpus, which equals 77 PMWs. The cut-off rate for the engineering corpus is common in lexical bundle research. For instance, several studies used cut-off rates between 20 PMWs and 40 PMWs, as in Biber and Conrad (1999), Hyland (2008), and Liu (2012) who used 20 PMWs cut-

off, and Biber, et al. (2004) and Biber and Barbieri (2007) who used 40 PMWs cut-off. However, the researcher did not find any investigation that used a cut-off similar to the conservative 77 PMWs in the literature, which was used in this study to accommodate for the smaller size of the pathway corpus. More on the choice of 77 PMWs cut-off is in the limitations of the current study in Chapter V.

Furthermore, the lexical bundles investigated in this research are composed of four words, e.g., *the influence lines for.* Four-word bundles are very common and are long enough to carry a functional value, thus they were chosen to be a unit of analysis. Moreover, lexical bundles that are part of proper nouns, such as institutions' names, were omitted from the lists since they do not facilitate the research objectives. Also, mathematical variables and symbols that are detected by the program as lexical bundles, as in *u v u v* were also removed for the same reason.

3.3 Analytical Framework

3.3.1 Functional analysis. As for the analysis and comparison of lexical bundles found in the corpora, the current study adapted the functional categorization used in Biber et al. (2004) and. The categorization of lexical bundles is induced from their contexts within the corpora. The major functional categorizations of lexical bundles are referential expressions, stance expressions, discourse organizers, and subject-specific bundles. Referential bundles "make direct reference to physical or abstract entities, or to the textual context itself" (Biber, et al., 2004, p.384). Furthermore, "stance bundles express attitudes or assessments of certainty that frame some other proposition" (Biber, et al., 2004, p.384). Lastly, discourse bundles, according to Biber et al. (2004), negotiate and arrange the flow of discourse by providing links to previous and coming sections. Moreover, subject-specific bundles contain those bundles that are related

directly to topics at hand such as engineering, as in *of the boundary layer* and *the plane of the*, politics, as in *democracy in the region* and *a firm stand against*, and medicine, as in *the use of placebo*.

3.3.2 Statistical tests. Several statistical tests were utilized to determine the statistical significances, or the lack thereof, found between the engineering and pathway corpora. The first statistical test used in this study was z-test for two population proportions to see if there were any significant differences in the density of four-word lexical bundles in both corpora. The second test utilized in this study was a chi-square test which was used to test whether there were any significant differences in the distribution of lexical bundles' functions in both corpora. The chi-square test of significance is common in corpus linguists' research to determine the significance of differences when comparing corpora. McEnery and Wilson (2001) showed that the sensitivity of the chi-square test and its assumptions about distribution of data coupled with its ease of use made it one of the most common statistical tests in corpus linguistic research. For instance, many published corpus studies used chi-square test to determine the significance of differences between corpora, and whether differences were due to chance (Henry & Roseberry, 2001; Bond, 2007; Hareide & Hodland, 2012). Thus, chi-square test was chosen since it was suitable to the type of data produced by the functional analysis, i.e., frequency counts. The test was conducted using Preacher's (2014) online chi-square calculator.

CHAPTER IV

Results of the Study

This chapter shows the results obtained from the analysis described in Chapter III to gain an understanding about the nature of formulaic language found in the engineering and pathway program corpora. The primary questions of this study are focused on identifying frequent lexical bundles of each corpus, and on comparing the functions of those lexical bundles. The results were extracted using a concordancing program and then analyzed to determine the discoursal functions of lexical bundles in both corpora. Moreover, a Chi-Square test was performed to determine the significance of differences between lexical bundles from both corpora. A Chi-Square is an appropriate choice to test the whether variations in frequency counts is due to statistical significance or random distribution. The following sections examine the findings from the engineering corpus, pathway corpus, and the statistical analyses of differences among them.

4.1 Findings From the Engineering Corpus

The number of four-word lexical bundles in the 1.26-million-word engineering corpus was 236 unique bundles after excluding bundles of proper names and incoherent codes. Those lexical bundles occurred a total of 16914 times within the corpus, covering more than 5% of the total words. The most frequently occurring bundles were *as shown in fig* with 417 occurrences, followed by *if and only if* with an occurrence rate of 286. Moreover, the first fifty lexical bundles occurred at a rate that was 87 or higher, with many above the 100 rate. A list of the most frequent lexical bundles is shown in Table 4.1.

4.1.1 Lexical bundles' discoursal functions in the engineering corpus. There were four different types of discoursal functions of lexical bundles in the engineering corpus:

referential bundles, stance bundles, discoursal bundles, and engineering bundles. The most prominent type of those was the referential one which covered 114 bundles out of 236 total. The second most prominent type was the engineering specific function totaling 91 bundles. However, the remaining types were not as frequent as the previous two, with the stance bundles totaling 19 bundles and the discoursal bundles totaling 12 bundles. The full lists of lexical bundles according to their functions are provided in Appendices C, D, E, and F. Lastly, there were no political or medical lexical bundles in the engineering corpus.

Table 4.1

Most Frequent Lexical Bundles in the Engineering Corpus (per corpus)

Frequency	Lexical Bundle	Frequency	Lexical Bundle
1771	as shown in fig (and variations of this bundle)	105	specification for structural joints
286	if and only if	105	using astm a or
194	shear and bending moment	104	a linear combination of
192	the influence line for	104	the shear and bending
183	the initial value problem	102	the owner s designated
162	with respect to the	101	for steel buildings and
152	on the other hand	101	in the contract documents
146	the influence lines for	100	the free body of
146	the top of the	99	in the direction of
128	is equal to the	99	of standard practice for
118	in the case of	99	practice for steel buildings
117	of the influence line	99	standard practice for steel
111	the limit state of	99	steel buildings and bridges
109	to the right of	99	the sum of the
108	influence lines for the	97	and bending moment diagrams
107	the direction of the	95	method of consistent deformations
106	code of standard practice	93	in terms of the
106	joints using astm a	92	the magnitude of the
106	structural joints using astm	91	is a solution of
105	for structural joints using	90	Owner’s designated representative for
105	in accordance with the	88	kip in n mm

4.2 Findings from the Pathway corpus

The pathway corpus contained 37 four-word lexical bundles within a 65,000-word corpus. The total number of occurrences of lexical bundles in the pathway corpus was 272 instances, with the lexical bundles covering just above 1% of the total words. Moreover, the most frequent bundle was *in the middle east,* occurring 19 times, followed by *is more important to* and *it is more important*, both occurring 17 times. Additionally, the majority of lexical bundles in the pathway corpus occurred at a rate of six or five occurrences per corpus. Full lists of lexical bundles from the pathway corpus are provided in Table 4.2.

Table 4.2

Most Frequent Lexical Bundles in the Pathway Corpus (per corpus)

Frequency	Lexical Bundle	Frequency	Lexical Bundle
19	in the middle east	5	a very serious problem
17	is more important to	5	as a result of
17	it is more important	5	avoid a military conflict
16	say it is more	5	China on economic issues
12	in the United States	5	firm stand against Iran
9	in the U S	5	from the United States
9	on the other hand	5	important to avoid a
8	at the university of	5	important to take a
8	in the case of	5	in the form of
7	a placebo controlled trial	5	more important to avoid
7	are more likely to	5	more important to take
6	a firm stand against	5	placebo controlled trials are
6	at the same time	5	placebo controlled trials of
6	democracy in the middle	5	Administration's handling of the
6	democracy in the region	5	to avoid a military
6	for the United States	5	use of placebo controls
6	in the context of	5	when it comes to
6	the use of placebo	5	with China on economic
6	the world health organization		

4.2.1 Lexical bundles' discoursal functions in the pathway corpus. There were five types lexical bundles in the pathway corpus: referential bundles, stance bundles, discourse bundles, politics-specific bundles, and medical-specific bundles. The most frequent of those types was the politics-specific type, totaling 13 occurrences, followed by the referential type, with 11 occurrences. Moreover, medical bundles occurred six times, stance bundles occurred five times, while discourse bundles occurred only two times. Lastly, there were no engineering-specific bundles in the pathway corpus. The full lists of lexical bundles according to their functions are provided in Appendices G, H, I, J, and F.

4.3 Results of Comparing the Two Corpora

4.3.1 Proportions of lexical bundles coverage between corpora. To test whether there was a significant difference between lexical bundles' proportions in each corpus, a *Z*-test was performed on the lexical bundles' coverage relative to their size. The result of comparing the coverage of lexical bundles relative to size of corpora using *Z*-test produced a *Z*-ratio of 41.3 in which $p<0.01$. This result indicated that differences in proportions of lexical bundles in the two corpora were statistically significant.

4.3.2 Overlapping lexical bundles between the two corpora. To test the first hypothesis, the two lists of lexical bundles were compared to identify bundles that occurred in both corpora at or above the cut-off levels set in Chapter III. However, only two bundles were on both lists: *in the case of,* and *on the other hand,* the former being a referential bundle while the later is a discourse-organizing bundle. The results of this comparison support the first hypothesis that the two corpora would display minimal overlapping instances of lexical bundles.

4.3.3 Differences and similarities of lexical bundles' functions. Moreover, to assess whether the results of the study support or reject the second hypothesis, a comparison of the raw

results from both corpora and a Chi-Square statistical analysis were necessary to determine if a significant difference was found. The results of the functional analysis show some vast differences between the two corpora on two fronts: types of lexical bundles and frequencies within those types. On the first hand, some topic-specific bundles were found in one corpus and were not present in the other. For example, political and medical-specific bundles were found at a relatively high frequency in the pathway corpus while they were absent in the engineering corpus. Similarly, engineering-specific bundles were found in the engineering corpus at a relatively high frequency while there was not a trace of such bundles in the pathway corpus, as shown in Table 4.3.

Table 4.3

Percentages of Lexical Bundles Coverage

Type	Engineering	Pathway
Referential	48%	30%
Stance	8%	14%
Discourse	5%	5%
Engineering	39%	0
Politics	0%	35%
Medicine	0%	16%
Totals	100%	100%

On the other hand, the results of the functional analysis showed differences between the frequencies of the same functional types between both corpora. Among types that were found in both corpora, only discourse bundles showed the same exact percentages of lexical bundles' coverage where five percent of lexical bundles in both corpora were discourse bundles. However, stance bundles covered eight percent of the engineering corpus and 14% of the

pathway corpus. Also, the big difference among shared types was found in referential bundles were they covered 48% of the lexical bundles in the engineering corpus and 30% of the lexical bundles in the pathway corpus. It is important to note that the percentages reported here are relative to the number of lexical bundles in each corpus, and that, as it has been reported earlier in this chapter, lexical bundles in the engineering corpus covered a bit more than 5% of the total words while lexical bundles in the pathway corpus covered a bit more than 1%.

Moreover, in order to see whether the difference between lexical bundles in both corpora was statistically significant, a Chi-Square test was employed. The total frequencies of occurrences, presented in Table 4.4, were computed to determine the statistical significance, or the lack thereof between the engineering and pathway corpora.

Table 4.4

Frequencies of Lexical Bundles Across Functions

Type	Engineering	Pathway
Referential	114	11
Stance	19	5
Discourse	12	2
Engineering	91	0
Politics	0	13
Medicine	0	6
Totals	236	37

The results of the Chi-Square test given five degrees of freedom yielded a Chi-Square value of 139 with $p<0.01$. The results of Chi-Square test indicated that the difference in lexical bundles between the two corpora was statistically significant, and that the differences in distribution did not occur by chance. Moreover, to accommodate for low expected values when performing the

Chi-Square test, I performed Chi-Square test with Yates' corrections. A Chi-Square test with Yates' corrections given five degrees of freedom yielded a Yates' Chi-Square value of 122.7 with $p<0.01$. Again, the corrected test also indicated a statistical significance of distribution between the two corpora. The results of these statistical analyses clearly reject the null hypothesis that there is no significant difference in distribution between lexical bundles in both corpora. Therefore, the second hypothesis that the discoursal functions of lexical bundles from both corpora will have no significant differences should be rejected since both statistical analyses indicate a significant difference.

4.4 Chapter Conclusion

This chapter presented the results of the current investigation about the nature of lexical bundles introduced in the first chapter and detailed in Chapter III. First, the most frequent lexical bundles were extracted using a concordancing program. Second, those lexical bundles were examined closely to determine their discoursal functions, which yielded three shared functions: referential, discourse organizers, and stance expressions. The functional analysis, also, produced three unique subject-specific functions that were attached to specific corpora: engineering, political, and medical functions. Moreover, this chapter presented a comparison that identified overlapping lexical bundles, which revealed a minimal overlap of two lexical bundles. Lastly, this chapter also presented an analysis of similarities and differences in lexical bundles' functions based on raw data resulting from previous analyses and on Chi-Square statistical analyses. The results of two types of Chi-Square tests showed a statistically significant difference between lexical bundles in the two corpora. The following chapter will discuss the results reported here in greater detail.

CHAPTER V

Discussion of the Results

This chapter discusses the results of the study described in Chapter III and addresses its main questions. The chapter examines the results reported in the previous chapter to see how they answer the questions put forth by the current study. Moreover, the chapter describes in details the functions of lexical bundles in both corpora, and relates those findings to similar studies in the literature. The description of functions is followed by a discussion of pedagogical implications of the study, limitations of the current investigation, and lastly directions for future investigations.

5.1 The Most Frequent Lexical Bundles

The first question of this study is concerned with identifying the most frequent four-word lexical bundles found in each corpus. The concordancing process described in Chapter III yielded two lists of the most frequent lexical bundles, found in Appendices A and B. The following paragraphs discuss those findings in more details.

5.1.1 Lexical bundles in the engineering corpus. As for lexical bundles occurring in the engineering corpus, referential bundles were the most common type with 114 bundles covering 48% of the total bundles. The following type was the engineering specific bundles, occurring 91 times and covering 39%. Stance and discourse bundles covered 13% of the bundles, occurring 31 times. To determine whether those distributions of lexical bundles functions had a statistically significant nature, a Chi-Square test was performed. The results of the Chi-Square test given three degrees of freedom yielded a Chi-Square value of 133 in which $p<0.01$. This Chi-Square

result supported the conclusion that there was a significant difference in the distribution of functions of lexical bundles from the engineering corpus.

Density of lexical bundles. Moreover, the high density of lexical bundles in the engineering corpus was in alignment with previous research in this area. This study showed that lexical bundles covered more than 5% of the total corpus, as it had been reported in Chapter IV. Hyland (2008) investigated several academic disciplines and found that electrical engineering had the highest density at 3.5%, compared to 2.2% in business studies, 1.9% in applied linguistics, and 1.7% in biology. Still, the results of analysis of the present engineering corpora showed that it was 2% denser than Hyland's electrical engineering corpus. Hyland (2008) argued that the nature of composition in engineering is tied to technical and graphical representation of information, which in turn results in formulaic traditions of showing data and constructing arguments. Hyland's (2008) investigation revealed that not only engineering discipline had denser lexical bundles, but also they were more unique and most were not found in other disciplines. This makes for a more difficult job of EAP and ESP syllabus designers and instructors to accommodate for their students' varied disciplines since each discipline requires a specific list of lexical bundles.

However, a big difference was found when results from the engineering corpus in this study were compared with Conrad and Biber's (2004) results about lexical bundles in general academic prose. Academic prose in Conrad and Biber's (2004) investigation displayed a less density of four-word lexical bundles at only 2%. However, the lack of a specified discipline in Conrad and Biber's (2004) investigation could be the reason for less dense formulaicity of texts since corpora formed around similar subjects offer more concentration of similar rhetorical characteristics, as seen in Hyland's (2008) corpora and the engineering corpus from this study.

The density of lexical bundles in the engineering corpus compared to specific disciplines, as in Hyland's (2008) study, or general academic texts further emphasize the importance of formulaicity and the usefulness of lexical bundles as building blocks of the engineering discourse.

5.1.1 Lexical bundles in the pathway corpus. The results from the pathway corpus showed a different pattern from the one observed in the engineering corpus. That different pattern was evident in two ways: the distribution of lexical bundles and the density of lexical bundles. Firstly, as for the distribution of lexical bundles, the most dominant function of lexical bundles in the pathway corpus was the politics-specific type, covering 35% of the total bundles and occurring 13 times. The second most dominant function was the referential type, covering 30% of the bundles and occurring 11 times. Moreover, medical bundles covered 16% while stance and discourse bundles were the least dominant functions, covering 14% and 5% respectively. Similar to the analysis of distribution of bundles in the engineering corpus, a Chi-Square test was performed to test whether the distribution in the pathway corpus was of statistical significant. The results of the Chi-Square test given four degrees of freedom yielded a Chi-Square value of 10.9 in which $p<0.05$. This Chi-Square result supported the conclusion that there was a significant difference in the distribution of functions of lexical bundles from the pathway corpus. However, the statistical significance of distribution in the engineering corpus was higher than the significance of distribution in the pathway corpus. This meant that that lexical bundles in the pathway corpus were relatively more normally dispersed across functions when compared to lexical bundles' dispersion in the engineering corpus.

Density of lexical bundles. Secondly, the density of lexical bundles in the pathway corpus was also dissimilar to that of the engineering corpus. As it has been reported in Chapter

IV, lexical bundles in the pathway corpus covered 1.6% of the total corpus, with 38 bundles occurring 272 times. This coverage was significantly less than that of the engineering corpus, which covered 5.3% of its total words. As reported in Chapter IV, a *z*-test was performed on those proportions which revealed a statistical significant of $z=41.3$, in which $p<0.01$. Moreover, only 49% of lexical bundles in the pathway were general enough to be beneficial multiple contexts, as in *it is important to, on the other hand,* or *in the case of.* The rest of the lexical bundles, such as *democracy in the region, a placebo controlled trail,* or *avoid a military conflict,* were of relatively less value to EAP or ESP learners.

5.2 Functional Description of Lexical Bundles

The following section provides a thorough description of functions of lexical bundles from the two corpora: the engineering corpus and the pathway corpus.

5.2.1 Functional description of lexical bundles in the engineering corpus.

Referential bundles. Almost half of the lexical bundles occurring in the engineering corpus were referential bundles. In this corpus, 114 different referential bundles occurred 8302 times, making this type the most frequently occurring function. Referential bundles, according to Biber et al. (2004), "make direct reference to physical or abstract entities, or to the textual context itself" (p.384). It was, thus, not surprising that the engineering discourse relied on this type of formulaic language to construct its flow of data and thoughts. This type of lexical bundles was heavily utilized to identify concepts, specify attributions whether tangible or abstract, and provide direct references within the text itself. This was expected since the nature of the engineering discourse utilizes ample use of graphs and diagrams, and that require more frequent reliance on formulas to convey the information contained by such diagrams and graphs. This pattern was, also, observed by Hyland (2008), who noted that engineering texts, and hard

sciences in general, used lexical bundles to describe procedures and data. Hyland (2008) contributed this frequent use of this pattern to emphasis in hard sciences on empirical discoveries over interpretations of researchers and the need to guide readers through information dispersed within texts. The following paragraphs review some of the salient types within referential bundles.

Identification and focus bundles. Many of the referential bundles were used as identification or focus bundles. Moreover, this sub-category showed frequencies of occurrence ranging from 41 to 61 occurrences within the corpus. Those lexical bundles were essential in guiding the attention of readers and directing the focus of arguments, as the following examples show:

> The effect of these assumptions is that all *the members of the* truss can be treated as axial
>
> the shape of the M=EI diagram *is the same as* that of the bending moment diagram
>
> Interestingly, *in this case the* complex integral is well-defined even when n is a negative integer

As these excerpts showed, the lexical bundles here were used to identify certain aspects and narrow the focus of an argument. The last example was used after a lengthy analysis to summarize and shift the focus to one important aspect within the context.

Bundles specifying attributes. Another sub-category of referential bundles found in the engineering corpus was bundles specifying attributes. Those bundles describe the characteristics of the following texts, whether in quantifying or framing function. For instance, some of those bundles specify quantities:

For some types of frames, a member or a joint that has a number of unknowns *less than or equal* to the number of equilibrium equation

Since the bending moment M *is equal to the* sum of the moments about the neutral axis of the forces acting at all the fibers of the beam cross section

Other bundles in this sub-category function as framing expressions, whether referring to tangible or intangible attributes. The following instances are examples of framing tangible attributes:

The virtual internal work due to bending for that segment can be obtained by integrating the quantity MvM=EI over *the length of the* segment.

Steel reinforcement *in the form of* closed ties or welded wire fabric providing confinement…

The definition of "mild" relies on *the magnitude of the* Reynolds number

As a result, the stability is governed by *the size of the* magnification factor

Similarly, the following examples frame intangible attributes:

each ordinate of an influence line gives *the value of the* response function

The result *is a basis for* the column space of the given matrix.

The total of lexical bundles occurring under this sub-category was 23 bundles, covering 20% of referential bundles. These numbers speak to the importance of such bundles in framing arguments and conducting analyses in the engineering discourse.

Time, place, and text references. Moreover, another sub-category of referential bundles found in the engineering corpus was bundles referring to time, place, or text. This sub-category covered 45% of the total referential bundles with a total of 52 bundles. Additionally, one of the most common variations of similar bundles occurred in this sub-category which were variations centered around *shown in*. Those variations were technically separate bundles that occurred at a

very high rate of 2808 times. To show the variation of those bundles, the following examples are a few: *as shown in fig, are shown in fig, shown in the figure,* and *is shown in fig.* These bundles were clearly used to refer to drawings and graphs accompanying the text. This pattern was also found in Hyland's (2008) investigation of lexical bundles in electrical engineering where six of those variations were reported in the most frequent 50 bundles. However, in contrary to findings from Hyland (2008) and the current investigation, such variations were not as common in Chen's (2008) analysis of electrical engineering texts. The following are examples of those variations in context:

the work is equal to the area under the force displacement diagram *as shown in Fig*. 7.1(b)

The freebody diagrams of the two portions of the truss thus obtained *are shown in Fig*. 8.19.

Additionally, ten lexical bundles among the referential bundles were referring to place or direction. The following examples show some of them in context:

due to application of major axis bending moment alone to *the area of the* compression

suppose the initial data represent a taller solitary wave *to the left of* a shorter one.

The x axis of the coordinate system is oriented *in the direction of* the centroidal axis of the member.

Those bundles occurred 802 times in the engineering corpus.

Multi-function bundles. Moreover, Biber et al.'s (2004) functional analysis identified several referential bundles that served several functions, depending on its context, as in *in the end of the chapter* and *in end of the hallway,* where the former refers to a text and the later refers

to a place. However, when inspecting this type of bundles in the engineering corpus, there was a tendency of a few bundles to be used in a certain way, and not as multi-functioning bundles. The following excerpts of the same bundle show this pattern of attaching a bundle that could refer to place or text to textual references exclusively:

…multiplying *both sides of the* differential equation…

…dividing *both sides of the* equation…

…dividing *both sides of the* differential equation…

However, many bundles follow the multi-functionality described in Biber et al. (2004). For instance, the bundle *at the end of* was used to refer to textual context, time, and place:

we will use the direct sum definition to do the Jordan Form construction *at the end of the* fifth chapter

Fastener components that are not incorporated into the work shall be returned to protected storage *at the end of the* work shift.

Having *the end of the* bolt extending beyond…

Engineering bundles. The second most frequent bundles occurring in the engineering corpus were bundles of specific engineering functions. Those bundles consisted of 91 different bundles that occurred 6796 times. In this type, bundles were parts of technical expressions used to describe analytical and procedural arguments. Moreover, many of the bundles in the engineering function were related to advanced mathematical procedures, a topic that is relevant to the engineering context, as Chen (2008) noted. Additionally, the discoursal and pedagogical values of those bundles can not be understated since engineering bundles covered 2% of the total corpus and were essential to presenting information and scientific procedures. The following excerpts show a sample of those bundles:

We next discuss construction of *the shear and bending* moment diagrams by the method of sections.

Equation (4) tells us *the Laplace transform of* the solution Y.

the center of the storm weakens rapidly with height above the top *of the boundary layer*.

portion of the truss must be constrained against all possible rigid body movements *in the plane of the* truss

The next few samples are of advanced mathematical nature that is attached to the engineering discourse:

Find a fundamental matrix solution of *the system of differential* equations.

Consider the linear, homogeneous *first order differential equation*

Suppose that f(x) is a finite *linear combination of the* functions

This pattern of heavy use of very specific bundles was observed in Hyland's (2008) analysis of disciplinary variations. Although Hyland (2008) did not go to great lengths in reporting long lists of bundles from each discipline, some bundles reported were clearly engineering-specific. Hyland (2008), also, noted that engineering and biology texts showed higher concentrations of what Hyland named "research-oriented bundles" (p.14). Similar to findings from this study, Hyland (2008) showed that engineering texts utilized bundles that were essential "to the description of research objects or context, specifying aspects of models, equipment, materials or aspects of the research environment" (p.14).

Stance bundles. The next function of lexical bundles was stance bundles that had 19 distinct bundles occurring 1042 times within the engineering corpus. According to Biber et al. (2004), "Stance bundles express attitudes or assessments of certainty that frame some other

proposition" (p.384). The 19 bundles in this category resulted in 8% coverage of the total bundles found in the corpora. This coverage seems to be different than that of Chen's (2008), which found stance bundles in engineering textbooks covering 20% of the total lexical bundles. Moreover, stance bundles express two functions: epistemic functions, and attitude/modality functions (Biber et al., 2004).

One notable feature of stance bundles in this corpus is that all bundles fell in the impersonal category. The preference of impersonal expression in this type is a clear feature of this register, as can be seen in those examples: *can be seen from, it can be seen, can be written as,* and *is considered to be.* This might be because of the tendency to separate the writer from context, a feature of the engineering register reported in Hyland (2008). Biber et al. (2004) also found that impersonal stance bundles tend to occur in academic textbooks and prose, while personal ones tend to occur in classroom teaching and conversation.

The following paragraphs explore those functions in the engineering context and their relevance to previous research.

Attitude and modality bundles. Most of the stance bundles in this corpus fell in this sub-category with a total of 16 bundles. Those bundles varied between expressing directive and ability stances, with ability stances being the majority. The following examples show a variation of such bundles:

> Let us see how the solution formula (2.75) *can be used to* solve the initial value problem.
>
> Separation of variables seeks special solutions that *can be written as* the product of functions of the individual variables.

Observe that this flow *can be obtained by* rotating the preceding example by 45 degrees.

The other type in this sub-category was directive stance bundles. Those bundles were mostly found in exercises where readers are directed to engage in solving problems, as the following examples show:

Find the general solution of the given differential equation.

Draw the influence lines for the vertical reactions at supports A and C

Determine the reactions and draw the shear and bending moment diagrams for the beam shown in Fig. 13.3.

Note that in the last example, most of the sentence could be constructed directly out of the complete lexical bundles list, which speaks to the importance of formulaicity in this register.

Epistemic bundles. The other sub-category of lexical bundles was concerned with expressing stances evaluating and validating knowledge. Moreover, only three bundles were within this sub-category. The following excerpts provide examples of them in context:

Any such support displacement *is considered to be* positive if it has the same sense as that assumed for the redundant.

The companion action load factors on L and S in that equation reflect *the fact that the* probability of a coincidence of the peak time-varying load with the occurrence of a fire is negligible.

So, in order to look at equations that are correct across unit systems, we restrict our attention to those that use dimensional constants; such an equation *is said to be* complete.

Again, those bundles were used to convey the authors' evaluations of heuristic issues while maintaining distance and avoiding using explicit personal evaluative expressions. This pattern was similarly observed in Chen's (2008) analysis of epistemic stance bundles in electrical engineering discourse.

Discourse organizing bundles. The least frequent function of lexical bundles in the engineering corpus was discourse organization, as only 12 different bundles were identified as discourse bundles. Those bundles covered 5% of the bundles found in the engineering corpus and occurred 774 times. Discourse bundles, according to Biber et al. (2004) negotiate and arrange the flow of discourse by providing links to previous and coming sections. Furthermore, two sub-categories of discourse bundles were observed in the engineering texts: topic introductions and topic elaboration. The following excerpts show a few instances of discourse bundles that were used to introduce topics:

> *In this section, we* will learn how to solve the initial value problem on the entire line.
>
> *In this chapter, we* will analyze several important evolution equations, both linear and nonlinear, involving a single spatial variable.

Similarly, the following examples show a few of the bundles used for topic elaboration:

> From the foregoing discussion, *we can see that* the analysis of structures for variable loads consists of two steps...
>
> Stiffened elements, *on the other hand*, make use of the postbuckling strength inherent in a plate that is supported on both of its longitudinal edges, such as in HSS columns.

In a similar manner we can show that the result of any encounter between three animals is independent of the order in which they meet.

Column bases and base plates shall be finished *in accordance with the* following requirements…

Findings from this study were aligned with the results of Biber et al.'s (2004) analysis of textbooks and academic prose. Several instances that were reported in Biber et al. (2004) were found in this current analysis, as in *in this chapter we, as well as the,* and *on the other hand.* Moreover, many lexical bundles in the engineering corpus had very similar construction and functions to those reported in Biber et al. (2004) with slight variations, as in *in this section we, the same as the,* and *in a similar manner.* Similar to Biber et al. (2004), Hyland (2008) observed an abundance of "structuring signals" (p.17), similar to the ones reported in this function. Those discoursal signals were essential to organizing texts and handling relationships between text stages (Hyland, 2008). However, Chen's (2008) analysis of electrical engineering lexical bundles found only three discourse bundles: *on the other hand* and *as well as the* which were found in this investigation, and *as long as the,* which did not occur in this corpus at the cut-off level. Such bundles were essential to organizing the engineering discourse, especially when we look at the frequency of occurrence of some of them, which occurred well above 100 times.

5.2.2 Functional description of lexical bundles in the pathway corpus.

Political bundles. The function that received the majority of lexical bundles was the politics-specific bundles, which contained 13 different bundles and covered 35% of the total

bundles. Those bundles occurred five to six times each, with a total of 68 occurrences within the pathway corpus. Moreover, the bundles in this function were centered on topics involving Iran, the Middle East, and military conflicts. This reflected the choice of articles selected in the book and supplemental materials. The following excerpts provide an overview of bundles serving political-specific function:

> The public has long favored tough measures to prevent Iran from developing nuclear weapons, and 56% now say it is more important to take *a firm stand against* Iran's nuclear program.
>
> By contrast, Romney voters say it is more important to get tough *with China on economic* issues, by 67% to 26%.
>
> In January, 50% favored taking a firm stand against Iran and 41% said it was *more important to avoid* a confrontation.
>
> The percentage prioritizing *democracy in the region* has slipped over the past year and a half.

This high concentration of political bundles is probably the result of a few articles reporting on polling responses. This made for highly repeated expressions coupled with percentages of respondents to those polls.

Referential bundles. Referential bundles came second with a total of 11 bundles occurring 105 times in the pathway corpus. Those 11 lexical bundles covered 30% of the total bundles in the pathway corpus, contrary to 48% referential bundles in the engineering corpus.

Place and time bundles. Moreover, the most common sub-category of referential bundles was place bundles which had five different bundles. The following instances provide an overview of those bundles:

> While there is no public consensus on how changes *in the Middle East* are likely to affect the United States, few think the effects will be positive.
>
> Nonetheless, since 2002 enthusiasm for trade has declined significantly *in the United States*, Italy, France and Britain, and views of multinationals are less positive in Western countries where economic growth has been relatively modest in recent years.
>
> *At the University of* the West of England, roboticist Peter Jaeckel is studying how to get a person to feel empathy with a machine.

Similar to the political bundles, four out of the five bundles were, while referring to places, they were centered on the Middle East and the United States, a result of article choices that exploited those contexts heavily.

Also, only one bundle functioned as a time reference in the pathway corpus. The following example shows the bundle in context:

> If Americans and other people realize the importance of language, there can be a global effort to save those languages in danger of extinction, and *at the same time* preserve the cultures of those languages.

Bundles specifying attributes. Furthermore, four referential bundles were specifying attributes and framing the context that follows. Those bundles were similar to findings from the engineering corpus in this investigation, and findings from Biber et al.'s

(2004) analysis of textbooks and academic prose. All of the bundles in this sub-category were used to frame intangible aspects, as the following examples show:

> He penned his response to the Indian activists in London *in the form of* a book.
>
> The question of the ethics of such compromise became a hot issue, for UK doctors, about fifteen years ago *in the context of* health care rationing.
>
> *In the case of* the global jihadi war, this would mean affirming the positive principles of both sides - though the 'sides' in this case are not only state and non-state organizations but also the concerned publics that stand behind them.

Moreover, it is worth noting that the bundle *in the case of* is one of the two bundles that occurred in both corpora in the current study.

Medical bundles. Medical bundles were another topic-specific category that was evident in the pathway corpus, similar to engineering and political bundles. There were six medical-specific bundles that occurred in a total of 34 times. Moreover, similar to the political bundles, bundles in this group were mostly centered on the controversial issue of placebo trails. Also, some of the bundles were extensions to other ones as in *the use of placebo* and *use of placebo controls*. Other bundles had some minimal variations among them, as in *placebo controlled trails are* and *placebo controlled trails of*. Furthermore, just like the political bundles, the medical bundles were highly concentrated in a few articles that repeated similar expressions frequently. The following examples provide an overview of bundles in this category:

> The absolute prohibition against *the use of placebo controls* in every case in which an effective treatment exists is too broad; the magnitude of harm likely to be caused by using placebo must be part of the ethical consideration.

Conversely, consider *a placebo-controlled trial* with a 30 percent rate of response to placebo and a 53 percent rate of response to the investigational drug.

First, both sides agree that certain *placebo-controlled trials are* clearly unethical.

Stance Bundles. Within the pathway corpus, five bundles functioned as stance expressions and occurred at a total frequency of 62 times. Moreover, relative to functions in the pathway corpus, those stance bundles covered 14% of the functional distribution of bundles. Additionally, two sub-categories were identified within this function: epistemic and attitudinal bundles.

Epistemic bundles. The following examples show the epistemic bundles in context:

These dynamics *are more likely to* occur when an opponent's violence is not met with violent counter reprisals by the resistance campaign and when this is communicated to internal and external audiences.

Fewer Americans (49%) viewed China's growing military power as *a very serious problem* for the United States.

In such bundles, the function was to evaluate and express a judgment regarding an issue within the context. Moreover, the impersonal bundle *are more likely to* was found to be a frequent bundle in Biber et al.'s (2004) analysis of textbooks, and less frequent in academic prose.

Attitudinal bundles. The other sub-category of attitudinal bundles is shown in the following examples:

> And a majority of Americans (54%) continue to *say it is more important to* have stable governments in the Middle East, even if there is less democracy in the region.
>
> By contrast, Romney voters *say it is more important to* get tough with China on economic issues, by 67% to 26%.

Those examples showed that this string of six words produced three four-word lexical bundles. Also, similar to previous patterns in the pathway findings, articles analyzing polling results resulted in highly repetitive expressions that were essential to the flow of data. Lastly, although exact bundles were not found in Biber et al. (2004) the researchers reported that impersonal obligation expressions such as *it is necessary to* and *it is important to* were highly frequent in textbooks and academic prose alike.

Discourse bundles. The least frequent function of lexical bundles was discourse-organizing bundles, with only two bundles that occurred 14 times in the pathway corpus. Those bundles, despite being only two, covered 5% of the functional distribution of lexical bundles, a percentage that matched that of discourse bundles in the engineering corpus. The first of those bundles was *on the other hand*, a bundle that was used to elaborate and provide a contrastive context of what was expressed earlier. This is clearly evident in the following examples:

> Solid majorities in Egypt, Tunisia, Jordan and Lebanon said the 2011 popular uprising would lead to more democracy in the Middle East. Turks and Pakistanis, *on the other hand*, were less hopeful.

On the other hand, critics argue, the swelling numbers of consumers reflect the improvement in material conditions that has paralleled the progress of nations since the dawn of civilization.

On the other hand, treating resources with respect and harvesting in a sustainable way immediately reinforce environmentally appropriate behavior.

This bundle was the second bundle that was found in both corpora in the current investigation. This was not surprising, given the results of analyses finding this bundle among the most frequent lexical bundles in many registers, as seen in Biber et al. (2004), Hyland (2008), Chen (2008) and Byrd and Coxhead (2010).

The second bundle functioning as a discourse organizer in the pathway corpus was *when it comes to.* This served as a topic introducer, situated at the beginning of the sentence in most instances. This bundle, also, was evident in articles discussing poll results, where many subjects are introduced and contrasted. The following examples display the bundle in context:

When it comes to economics, most say women should be able to work outside the home, but most also believe that when jobs are scarce, jobs for men should be the first priority.

When it comes to democracy, the public does "not just support the general notion of democracy -- they also embrace specific features of a democratic system, such as competitive elections and free speech," the Pew report said.

5.3 Pedagogical Implications of This Study

While the context of this study is closely related to pedagogical environment, the purpose of this study was not to test pedagogical treatments of lexical bundle. Byrd and Coxhead (2010) state that there is not much empirical research of the efficacy of different approaches of teaching lexical bundles, and that language teachers are facing a challenge of how to tackle this very important aspect of language. Nonetheless, the results of this investigation of the nature of lexical bundles in the engineering and pathway corpora have some pedagogical implications.

The density of lexical bundles identified in the engineering corpus echo the calls for more attention to formulaic language in EAP and EAP contexts. Such density echoes proposals from Nattinger and DeCarrico (1992) Lewis (1997) for more attention to formulaic language in language classroom to improve fluency and communicative ability, despite reservations about their applications in the field language teaching. This investigation, similar to others in this field showed the importance that prefabricated and conventionalized chunks of language hold in forming any discourse. It was because of the importance of lexical bundles that many projects have been devoted to the analysis of lexical bundles and formulaic language in varying academic contexts, especially with the advancement of corpus linguistics. Additionally, as it has been reported in the review of literature, those studies concur that different registers are attached to several formulaic patterns that make them distinct and unique. This uniqueness is bound to make the job instructors and course designers only harder, as noted by Hyland (2008). Thus, the results of discipline-specific analysis of lexical bundles, as in this study and similar studies,

would be of help to any learners with register-specific linguistic needs and should inform the syllabus design and material selection processes.

However, as this study uncovered, there is a big gap in the nature of formulaic language in the analyzed corpora. This gap was not only limited to the results of comparing the pathway corpus to the engineering corpus, but the gap also extended to the results of comparing the pathway corpus to studies analyzing general academic texts. To address such gaps, Hyland (2008) argued that the best way to prepare language learners is not to look for universally useful materials, however helpful, but rather is to present what those learners will need to read and write about, and to guide learners to recognize discoursal patterns. Perhaps there is a need for designing adaptive classes, especially at advanced levels, that give language learners the chance to interact with discourses they will become very attached to through their academic journeys. This need is evident in the texts used in the Pathway class investigated in this study which lacked any materials that connected students to their respective fields, and rather focused on controversial and argumentative topics such as peace, technology, and placebo trails. Additionally, there are several types of texts in science-oriented publications that could serve as better replacements for the texts used in Pathway class. For example, articles that discuss engineering-related topics that are written for the non-specialized reader could be more suited for language learners wishing to enroll in engineering programs. Moreover, guiding learners to write about topics related to their fields of studies and choosing references that fit this criteria might be more helpful to their advancement than writing about topic of irrelevance to them. Benson (1993, cited in Chen, 2008) showed that using materials that

are relevant to learners' fields of study would result in positive effects on their learning process.

Moreover, one of the challenges of teaching lexical bundles reported in the field was the lack of tangible gains from explicit instruction when it comes lexical bundles usage, as Cortes (2006) found after a series of mini-lessons. Cortes's (2006) experiment found that although native-speakers' perception of lexical bundles was greatly raised, their compositions did not reflect an increased usage of lexical bundles. Cortes (2006) argued that it is "possible that the learning of these expressions could be connected to the development of students' knowledge of the discipline and identity in the academic community" (p.401). If the case was that only time and engagement with the disciplinary community will boost and enhance learners' control of lexical bundles, we could only benefit from engaging students as early as appropriately possible with their respective fields of study. For more pedagogical implications related to lexical bundle, Byrd and Coxhead (2010) provided a lengthy discussion of central issues to teaching lexical bundles, such as what lexical bundles should be taught in EAP classes, and at what length and context.

Additionally, developing field-specific materials is the bare minimum for Pathway classes to be considered true bridges to academic fields of study. With the current state of affairs uncovered by this investigation, it is hard to see how such Pathway classes are preparing students for specific academic fields. It is understandable that material development is a long and taxing process, but the current practices of language teaching might me failing to live up to the promises made to students and academic institutions alike.

5.4 Limitations of the Current Study

While working on this study, there were several apparent limitations that could be addressed in future research. Those limitations were centered on three issues: the corpora building process, the size of the corpora, and generalizability of the results. The first of the limitations is related to the process of compiling corpora for this study. As for the engineering corpus, several texts that were suggested by the field-experts were not available as computer-read texts, which resulted in a very difficult and time consuming process of finding texts and preparing them for the concordancing program. Moreover, in many cases, I had to resort to older editions that were available when the most current ones were not accessible. As for the pathway corpus, the process of compiling the texts was much easier since most the corpus components were available on the Internet.

However, the pathway corpus, despite being a true representative of the target context, was limited in its size. The issue with the pathway corpora being small was related to comparing it to a million-word corpus, as in the case of the engineering corpus. On the corpora size issue, Sinclair (1991) suggested that corpora should be as large as possible, while Sinclair (2004) also conceded that it is necessary to deal with smaller corpora when they fit their purpose of representation. Since this study dealt with corpora of two sizes, it was necessary to employ some leniency with regard to the set cut-off points. Although it is common in lexical bundles research to use normalized cut-off points, using normalization meant to accept very low rate of occurrence in the pathway corpus. For instance, a rate of 40 per the engineering corpus of 1.26-million-word equals a rate of 32 PMWs. On the other hand, a rate of 32 PMWs equals a rate of 2 per the pathway corpus of 65000 words. The decision was made to choose a cut-off of 5 per

65,000 words, which equals a rate of 77 per million words, to accommodate for the smaller size of the pathway corpus.

Additionally, the cut-off point of 5 per 65,000, despite being overly conservative, combined with the small sized corpus resulted in many topic-specific bundles, such as the political and medical bundles. Thus, on one hand, the smaller size lead to choosing a conservative cut-off point, and on the other, even after using such cut-off point, an overwhelming majority of the bundles were topic-specific. However, the inclusion of other teaching materials that language learners encounter in the pathway program might be a remedy for this issue since it will results in a larger corpus. Such inclusion would also provide a broader picture of the language presented at INTOCSU.

The last of the limitations of this study is related to the generalizability of its results. Although the corpora compiled for this research were designed to be true representatives of their context, those corpora present very specific instances taken from one ESL school. This coupled with the limited research comparing the language presented at ESL programs and language used in various academic disciplines make the results of this study limited to what they investigate, one instance at an ESL program. Further research in different contexts is needed to gain a better understanding of the nature of current practices.

5.5 Directions for Future Research

The results of this study open the question about what would similar investigations in different environments reveal about the nature of formulaic language. The scarcity of research in this niche is evident, with this study being the second of its nature, to the extent of my knowledge. There is still a lot to be known about how formulaic language is situated in ESL programs, and how its nature fit with learners' future academic programs. Additionally, more

research in this area is bound to improve the methodology carried in this line of research, and to provide a broader and clearer picture of the current status of formulaic language.

Moreover, future research similar to this study needs to use larger corpora, especially the ones involving ESL texts. There is an inherent difficulty attached to research investigating very specific corpora, in that availability of texts is harder attain. For instance, to build a representative sample of texts used in an institute, a researcher needs the exact texts used and cannot substitute with texts of perceived similar functions, since doing so would limit the applicability of the results. This is not the case when one is investigating more general contexts such as general academic language, engineering introductory textbooks, or research articles in any discipline, where one could substitute one text for another. Despite the discussed hardship, employing larger corpora would lead to more accurate analysis and more generalizable findings.

Another direction for future research involves more longitudinal approach to this line of research. Although a bit ambitious, adding second language learners writing as another corpus for analysis could shed light on both the nature of formulaic language in texts ESL programs and target fields, and how it compare to the nature of formulaic language used by language learners exposed to both. Such study could also identify deficiencies in learners' use of formulaic language, among other linguistic aspects, and then inform syllabus designers to address such needs. However, such project will probably require the cooperation of language institutes and university departments to achieve this ambitious goal.

Additionally, another future direction for research could utilize the nature of the highly specific corpora similar to the ones developed in this study is to create contexts-specific words lists. Although there is a great deal of interest in developing vocabulary lists, whether for academic or general contexts, the context-specific corpora developed in this line of research

could be helpful in building specific lists that are not covered in lexical research. Such lists of vocabulary would be very helpful for people enrolling in specific programs, such as electrical engineering, to give an example.

5.6 Chapter Conclusion

The goal of this study was to analyze the nature of formulaic language and its various functions in two corpora: a corpus of required readings in one advanced writing class in the pathway program, and a corpus of engineering texts used at different engineering programs at CSU. The compiled corpora consisted of 65,000 words for the pathway corpus and 1,264,106 words for the engineering corpus. The results of the analyses in this study revealed that there is a significant difference in the nature of formulaic language in the two corpora. The first difference was observed in the density of lexical bundles in the engineering corpus, which was significantly denser than the pathway corpus. The second observed difference was evident in the distribution of functions between the engineering and pathway corpora. Both of those differences were analyzed statistically and the analyses showed that the differences were statistically significant. Moreover, the overlap between the bundles from both corpora was minimal, with two overlapping bundles in a pool of 273 bundles. The analysis of formulaic language in the two corpora revealed that there was a big gap between the formulaic language that learners encounter in an advance writing class in the pathway program and what they encounter when they enroll in engineering programs.

This study provides a preliminary look into how language from different disciplines and from English language programs compare with regard to formulaic language. It is hoped that results from this study trigger more research in this area of analyzing formulaic language. The potential gains from undertaking similar analyses could prove helpful to the fields of EAP and

ESP, since they point out how language-learning materials are living up to expectations. In this study, I outlined some of the limitations that were found in this study and how to improve on it. I, also, suggested some directions of future research that extend from the current study to add to the field of second language acquisition.

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*, 81-92.

Anthony, L. (2012). AntConc (3.3.5m) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Becker, J. (1975). The phrasal lexicon. In R. Schank & B. L. Nash-Webber (Eds.), *Proceedings of the First Interdisciplinary Workshop on Theoretical Issues in Natural Language Processing, Cambridge, MA* (pp. 60-63).

Bahns, J., Burmeister, H., & Vogel. T. (1986). The pragmatics of formulas in L2 learner speech: Use and development. *Journal of Pragmatics 10*, 693-723.

Biber, D. (2001). Using corpus-based methods to investigate grammar and use: Some case studies on the use of verbs in English. In R. Simpson & J. Swales (Eds.), *Corpus linguistics in North America* (pp. 101-115). Ann Arbor: The University of Michigan Press.

Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes, 5,* 97-116.

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

Biber, D., & Conrad, S. (1999) Lexical bundles in conversation and academic prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of Corpora: Studies in Honour of Stig Johansson*

(181-190). Rodopi.

Biber, D., & Conrad, S. (2001). Register variation: A corpus approach. In D. Schiffrin, D. Tannen, & H. Hamilton (Eds.), *The handbook of discourse analysis* (pp.175-196). Oxford: Blackwell.

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

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

Bond, L. L. (2007). Metaphors in context: Domains and lemmas as indicators of disciplinary differences in conceptualizations of "reality". Master's Thesis. Colorado State University, Fort Collins, CO.

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.

Chacón Beltrán, R., Abello-Contesse, C., & Torreblanca-López, M. d. M. (2010). Vocabulray teaching and learning: Introduction and overview. In R. Chacón Beltrán, C. Abello-Contesse, & M. d. M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp.1-12). Bristol: Multilingual Matters.

Chen, L. (2008). An investigation of lexical bundles in electrical engineering introductory textbooks and ESP textbooks (MA thesis). Retrieved from ProQuest database.

Conrad, S. (2002). Corpus linguistics approaches for discourse analysis. *Annual Review of Applied Linguistics 22,* 75-95.

Conrad, S., & Biber, D. (2004). The frequency and use of lexical bundles in conversation and

academic prose. *Lexicographica, 20,* 56-71.

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. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and education, 17,* 391-406.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Ferguson, C. (1976). The structure and use of politeness formulas. *Language in Society, 5*, 137–151.

Firth, J. R. (1957). *Papers in linguistics, 1934-1951.* London: Oxford University Press.

Francis, G. (1993). A corpus-driven approach to grammar. Priniples, methods and examples. In M. Baker, G. Francis, & E. Togini-Bonelli (Eds.), Text and Technology: In Honour of John Sinclair (pp. 137-156). Amsterdam: Benjamins.

Francis, W. N. (1992) Language corpora BC. In J. Svartvik (ed.), *Directions in corpus linguistics* (pp. 17–32). Berlin: Mouton.

Hakuta, K. (1974). Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning, 24*, 287‐297.

Hareide, L., & Hodland, K. (2012). Compiling a Norwegian-Spanish parallel corpus. In M. P. Oakes, & M. Ji, (Eds.), *Quantitative methods in corpus-based translation studies: A practical guide to descriptive translation research* (pp. 75-114). Amsterdam/Philadelphia: John Benjamins Pub. Co.

Henry, A., & Roseberry, R. L. (2001). Using a small corpus to obtain data for teaching a genre.

In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small corpus studies and ELT: Theory and practice* (pp.93-134). Amsterdam: J. Benjamins Pub. Co.

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

Hyland, K., & Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. *Applied Linguistics, 25*(2), 156-177.

Jespersen, O. (1917). *Negation in English and other languages*. Copenhagen: A.F. Host.

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.

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.

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

Larsen-Freeman, D., & Anderson, M. (2011). *Techniques and principles in language teaching.*3rd ed. Oxford; New York: Oxford University Press.

Laufer, B. (1997). What's in a word that makes it hard or easy: Some intralexical factors that affect the learning of words. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 140-155). Cambridge: Cambridge University Press.

Leech, G. (1987). General introduction. In R. Graside, G. Leech, & G. Sampson (Eds.), *The computational analysis of English: A corpus-based approach* (pp. 1-15). New York, NY: Longman Inc.

Lewis, M. (1993). *The lexical approach.* Boston: Heinle/Cengage.

Lewis, M. (1997). Pedagogical implications of the lexical approach. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 255-270). New York: Cambridge University Press.

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes, 31,* 25-35.

McCarthy, M., & Carter, R. (2006). This that and the other: Multi-word clusters in spoken English as visible patterns of interaction. In M. McCarthy (Ed.), *Explorations in Corpus Linguistics* (pp. 7-26). New York: Cambridge University Press.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics.* Edinburgh: Edinburgh University Press.

Moon, R. (1997). Vocabulary connections: Multi-word items in English. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 40-63). Cambridge: Cambridge University Press.

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

Nation, I. S. P. (2001). *Learning vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 3-14). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford

[England]: Oxford 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.

Nekrasova, T. (2009). English L1 and L2 speakers' knowledge of lexical bundles. *Language Learning, 59*(3), 647-686.

O'Dell, F. (1997). Incorporating vocabulary into the Syllabus. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 258-279). Cambridge: Cambridge University Press.

Palmer, H. E. (1964). *The principles of language-study.* London: Oxford University Press.

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: Longman.

Preacher, K. J. (2014). Calculation for the chi-square test: An interactive calculation tool for chi-square tests of goodness of fit and independence [Computer software]. Available fromhttp://quantpsy.org.

Renouf, A. (2007). Corpus development 25 years on: From super-corpus to cyper-corpus. In R. Facchinetti (Ed.), *Corpus linguistics 25 years on* (pp. 27-50). New York, NY: Rodopi.

Richards, J. (1976). The role of vocabulary teaching. *TESOL Quarterly* 10, 77-89.

Richards, J. C., & Rodgers, T. S. (2001). *Approaches and methods in language teaching*. 2nd ed. Cambridge: Cambridge University Press.

Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt & M. McCarthy (Eds.),

Vocabulary: Description, acquisition and pedagogy (pp. 199-227). Cambridge: Cambridge University Press.

Schmitt, N. (2000). *Vocabulary in language teaching.* Cambridge: Cambridge University Press.

Schmitt, N., & McCarthy, M. (Eds.) (1997). *Vocabulary: Description, acquisition, and pedagogy.* Cambridge: Cambridge University Press.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list (AFL). *Applied Linguistics*, *31*(4), 487–512.

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

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.

Sökmen, A., J. (1997). Current trends in teaching second language vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 237-257). Cambridge: Cambridge University Press.

Svartvik, J. (2007). Corpus linguistics 25+ years on. In R. Facchinetti (Ed.), *Corpus linguistics 25 years on* (pp. 11-26). New York, NY: Rodopi.

West, M. (1930). Speaking-vocabulary in a foreign language. *The Modern language Journal, 14*(7), 509-521.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics, 21*(4), 463-489.

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

Wray, A. (2009). Future directions in formulaic language research. *Journal of Foreign Languages, 32*(6), 2-17.

Zimmerman, C. B. (1997). Historical trends in second language vocabulary instruction. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 5-19). New York: Cambridge University Press.

APPENDIX A

A List of the Content of the Engineering Corpus

Braun, M. (1983). *Differential equations and their applications: An introduction to applied mathematics* (3rd ed.). New York: Springer-Verlag.

Dingman, S. L. (1994). *Physical hydrology.* Upper Saddle River, NJ: Prentice Hall.

Hefferon, J. (2012). *Linear algebra.* Publisher: Author. Retrieved from: http://joshua.smcvt.edu/linearalgebra

Holmes, M. H. (2009). *Introduction to the foundations of applied mathematics.* Dordrecht: Springer. Available from https://ezproxy2.library.colostate.edu/login?url=http://dx.doi.org/10.1007/978-0-387-87765-5

Kassimali, A. (2010). *Structural analysis* (4th ed.). Stamford, CT: Cengage Learning.

Oliver, P. J. (2013). *Introduction to partial differential equations*. New York, NY: Springer.

Rodrigo, D., Calva, E. J. L., & Cannan, A. (2012). *Total water management*. Cincinnati, OH: U.S. Environmental Protection Agency.

Steel construction manual (13th ed.). (2005). Chicago, Ill.: American Institute of Steel Construction.

Wallace, J. M., & Hobbs, P. V. (2006). *Atmospheric science: An introductory survey* (2nd ed.). Amsterdam: Elsevier Academic Press.

APPENDIX B

A List of the Content of the Pathway Corpus

Baynham, J. (2009, February 3). Come hell or high water, the Burmese Junta endures. *Slate.* Retrieved from http://www.slate.com/id/2210439/ .

Birdwell, A. F. (2007, January 18). Addicted to phones? Cell phone use becoming a major problem for some, expert says. Retrieved from http://news.ufl.edu/2007/01/18/cell-addiction/ .

Bond, M. (2009, May 18). Smart robots. *Engineering and Technology Magazine, 4*(9). Retrieved from http://eandt.theiet.org/magazine/2009/09/smart-robots.cfm .

Bradley, W. (2007, December 15). "Tinkering with nature" Asia's disturbing trend is tipping the balance between the sexes. *Rocky Mountain News.*

Bruemmer, D. (2006). Humanoid robotics, ethical considerations.

Carr, N. (2010, June 5). Does the Internet make you dumber? *The Wall Street Journal.* Retrieved from http://online.wsj.com/news/articles/SB10001424052748704025304575284981644790098 .

Charles, J. A. (2004, July 14). The environmental benefits of globalization. *Global Envision.* Retrieved from http://www.globalenvision.org/library/1/645 .

Christie, B. (2000, October 14). Doctors revise declaration of Helsinki. *BMJ.* Retrieved from http://www.bmj.com .

Cook, J. (2004, August 31). AA Gandhi in Jerusalem: A nonvilont protest offers little hope for

Palestinians. *The New York Times.* Retrieved from http://www.nytimes.com/2004/08/31/opinion/31iht-edcook_ed3_.html .

Duvall, J. (2004, September 22). Outside view: Liberation by the people. *United Press International.* Retrieved from http://www.upi.com/Business_News/Security-Industry/2004/09/22/Outside-view-Liberation-by-the-people/UPI-41171095876933/ .

Emanuel, E. J., & Miller F. G. (2001, September 20). The ethics of placebo-controlled trails—A middle gound. *N Engl J Med, 345*(12), 915-919.

Few believe U.S. backs democracy (2012, July 10). *Pew Global Attitudes Project.*

Haass, R. N. (2009, May 5). When is war justifiable? *Washington Post.*

Hastings, M., Thiel, S., & Thomas, D. (2003, January 20). The deadly noodle. *Newsweek, 141(*3).

Hope, T. (2000). The best is the enemy of the good- Can research ethics learn from rationon? *The Journal of Medical Ethics, 26,* 417-418.

Ireland, C. (2008, February 14). Ethicists, philosophers discuss selling of human organs. *Harvard News Office.*

Juergensmeyer, M. (2007). Gandhi vs. terrorism. *Deadalus.*

Just War. (2008). In W. A. Darity, Jr. (Ed.), *International Encyclopedia of the Social Sciences* (2nd ed., Vol. 4, pp. 235-237). Detroit: Macmillan Reference USA. Retrieved from http://go.galegroup.com/ps/i.do?id=GALE%7CCX3045301234&v=2.1&u=coloradosu&it=r&p=GVRL&sw=w

Kakabadse, Y. (2002, September 3). Thinking aloud. Is ethics the missing link?*UN Cronicle, 39.*

Retrieved from http://international.vlex.com/vid/is-ethics-the-missing-link-53072557 .

Kelly, M. (2012). Martin Luther King Jr. Retrieved from http://www.about.com .

Knickerbocker, B. (2004, January 22). If poor get richer, does world see progress? *The Christian Science Monitor.* Retrieved from http://www.csmonitor.com/2004/0122/p16s01-wogi.html .

Kreiter, M. S. (2012, July 15). Arab spring: Hopes for democracy spring eternal. *United Press International.* Retrieved from http://www.upi.com/Top_News/US/2012/07/15/Arab-Spring-Hopes-for-democracy-spring-eternal/UPI-21871342342860/print#ixzz2BTEGIvTd .

Lawler, P. A. (2005). The problem of technology. *Perspectives on Political Science.*

Locke, E. A. (2002, May 1). Anti-globalization: The left's violent assault on global prosperity. *Capitalism Magazine.* Retrieved from http://www.CapMag.com/article.asp?ID=1559 .

Ma, Y. (2008). Paths of globalization: From the Berbers to Bach. *New Perspectives Quarterly, 25*(2), 19-21.

Mahatma Gandhi biography (2009). *bio True Story.* Retrieved from http://www.biography.com .

Merriman, H. (2008). Agents of change and nonviolent action. Publisher: Author. Retrieved from http://hardymerriman.com/wp-content/uploads/2011/10/Agents_of_Change.pdf .

Moore, J. (2009, September 7). Extreme do-gooders- What makes them tick? *The Christian Science Monitor.* Retrieved from http://www.csmonitor.com/World/Making-a-difference/2009/0907/p02s05-lign.html .

On eve of foreign debate, growing pessimism about Arab spring aftermath. (2012, October 18). *Pew Research Center.*

Park, A. (2012). Should people be able to sell their organs? *Time.*

Pinker, S. (2010, June 10). Mind over mass media. *The New York Times.* Retrieved from http://www.nytimes.com/2010/06/11/opinion/11Pinker.html?_r=0 .

Polster, M. (2001). Eve's daughters. *Gestalt Journal Press.*

Rettner, R. (2009, August 10). Great debate: Should organ donors be paid? *Scienceline.*

Richardson, K. (2007, February 16). My friend the robot. *Times Higher Education.* Retrieved from http://www.timeshighereducation.co.uk/features/my-friend-the-robot/207843.article .

Rosenthal, E. (2002, February 25). Buicks, Starbucks and fried chicken: Still China? *The New York Times, Beijing Journal.*

Smith, J. (2006). Living on the edge: Extreme sports and their role in society. Retrieved from http://www.summitpost.org/ .

Stephan, M. J. (2005). War without violence: The potential and pitfalls of nonviolent struggle in East Timorese independence movement.

Stephan, M. J., & Chenoweth, E. (2011). Why civil resistance works. The strategic logic of nonviolent conflict.

Taft, S. (2008, April 29). Globalization and language. Retrieved from http://voices.yahoo.com/globalization-language-1412151.html .

Tellefson, T. (1993). People who make changes: Is a hero really nothing but a sandwich. *Utne*

Reader.

The ethics of buying and selling kidneys (2008). *Harvad University, Mass General.*

Tu, J. I. (2008, November 30). 'Embryo adoption' reopens controversy. *Post Gazette.* Retrieved from http://www.post-gazette.com/stories/news/us/embryo-adoption-reopens-controversy-623425/ .

White, H. (2006, January 10). Pew survey finds opinion divided on physician assisted suicide. *LifeSiteNews.com.* Retrieved from http://www.lifesitenews.com .

Wolpert, L. (1999, Match 25). Is science dangerous? *Nature, 398,* 281-282.

World publics welcome global trade- But not immigration. (2007). *Pew Global.* Retrieved from http://www.pewglobal.org/files/pdf/258.pdf .

Younkins, E. (2000, January 1). Technology, progress, and freedom. *Foundation for Economic Education.* Retrieved from http://www.fee.org/the_freeman/detail/technology-progress-and-freedom .

Yunus, M. (2000) The role of the corporation in supporting local development. *Reflections, 9*(2).

Zinn, H. (2001, December). A just cause, not a just war. *The Progressive.* Retrieved from http://www.progressive.org .

APPENDIX C

A Full List of Referential Bundles in the Engineering Corpus

Frequency	Lexical Bundle	Frequency	Lexical Bundle
417	as shown in fig	69	truss shown in fig
278	shown in fig a	68	chapter three maps between
240	shown in fig b	68	is referred to as
205	is shown in fig	66	in the absence of
162	with respect to the	65	a solution of the
146	the top of the	65	top of the atmosphere
145	shown in fig c	65	at the ends of
140	beam shown in fig	65	shown in fig e
139	are shown in fig	64	due to the external
128	is equal to the	64	the ends of the
118	in the case of	63	the temperature of the
109	to the right of	63	in the form of
107	the direction of the	62	the surface of the
106	shown in fig p	62	frame shown in fig
104	shown in fig d	61	in the earth s
99	the sum of the	61	the members of the
99	in the direction of	61	the total number of
93	in terms of the	61	on the basis of
92	the magnitude of the	61	than or equal to
87	the general solution of	60	of the beam shown
87	the right hand side	59	in fig a by
86	as a function of	59	the m ei diagram
84	referred to as the	58	the solution to the
83	to the left of	58	the center of the
83	the beam shown in	58	the numerical values of
81	at the top of	57	is the same as
80	by the method of	57	both sides of the
76	the slope of the	56	the method of consistent
74	the value of the	56	the dimension of the
72	the solution of the	56	to the axis of
72	as shown in the	55	in this case the
71	of the initial value	53	determined in accordance with
69	the member end moments	53	the location of the
69	the length of the	53	the values of the
69	shown in the figure	53	right hand side of

Frequency	Lexical Bundle
51	on the order of
51	in fig a the
51	shown in figure c
50	for the analysis of
50	of the earth system
50	of the primary beam
50	the area of the
50	the truss shown in
49	are referred to as
49	each of the following
49	is based on the
49	is known as the
49	equal to or less
49	to or less than
49	in fig b the
47	the set of all
47	the ratio of the
47	the size of the
47	the frame shown in

Frequency	Lexical Bundle
46	as defined in section
45	each member of the
45	in the northern hemisphere
45	the earth s atmosphere
45	shown in fig f
45	shown in figs p
44	less than or equal
43	in fig b and
42	a basis for the
42	the position of the
42	just to the right
41	of each of the
41	the right of the
41	a function of the
41	is a basis for
41	of the truss shown
40	as shown in figure
40	at the time of

APPENDIX D

A Full List of Engineering Bundles in the Engineering Corpus

Frequency	Lexical Bundle	Frequency	Lexical Bundle
286	if and only if	76	of the differential equation
194	shear and bending moment	76	the bending moment diagram
192	the influence line for	74	order linear differential equations
183	the initial value problem	74	the equations of equilibrium
146	the influence lines for	72	free body of the
117	of the influence line	70	the virtual work method
111	the limit state of	69	as t approaches infinity
108	influence lines for the	69	second order linear differential
106	code of standard practice	68	three maps between spaces
106	joints using astm a	67	first order differential equations
106	structural joints using astm	64	the engineer of record
105	for structural joints using	64	to the free body
105	specification for structural joints	63	the axis of symmetry
105	using astm a or	62	the freebody diagram of
104	a linear combination of	61	deflected shape of the
104	the shear and bending	60	the system of equations
102	the owner s designated	59	at the earth's surface
101	for steel buildings and	59	considering the equilibrium of
101	in the contract documents	58	systems of differential equations
100	the free body of	57	theory of differential equations
99	of standard practice for	56	system of differential equations
99	practice for steel buildings	54	of a vector space
99	standard practice for steel	54	the bending moment at
99	steel buildings and bridges	53	find the general solution
97	and bending moment diagrams	52	is a linear combination
95	method of consistent deformations	52	of the unit load
91	is a solution of	51	is a vector space
90	Owner’s designated representative for	51	linear combination of the
88	kip in n mm	51	qualitative theory of differential
79	for the limit state	50	solution of the initial
78	the boundary value problem	49	bending moment at point

Frequency	Lexical Bundle	Frequency	Lexical Bundle

49	by considering the equilibrium
49	designated representative for construction
49	shop and erection drawings
48	the method of joints
48	the strength of the
47	and bending moment at
47	bending moment diagrams for
47	in the plane of
47	of the bending moment
47	the limit states of
46	and only if it
46	designated representative for design
46	off type tension control
46	the thickness of the
46	twist off type tension
45	to the heat equation
44	is one to one
44	the equilibrium solution x
43	and only if the
43	influence line for the
43	of the boundary layer
43	of the response function
43	every solution x t of
42	moment diagrams for the
42	the plane of the
42	the system of differential
42	type tension control bolt
41	the structural steel frame
40	the laplace transform of

APPENDIX E

A Full List of Stance Bundles in the Engineering Corpus

Frequency	Lexical Bundle
73	can be used to
70	draw the influence lines
70	can be determined by
65	it can be seen
61	can be expressed as
60	can be written in
58	can be written as
56	the fact that the
56	be written in the
53	find the general solution
51	can be obtained by
50	can be seen from
49	written in the form
48	is said to be
46	can be found in
	determine the reactions
45	and
45	is considered to be
44	solve the initial value
42	can now be determined

APPENDIX F

A Full List of Discourse Organizing Bundles in the Engineering Corpus

Frequency	Lexical Bundle
152	on the other hand
105	in accordance with the
71	we can see that
	in accordance with
63	section
56	in a similar manner
53	the same as the
51	as well as the
48	in this section we
46	in this chapter we
45	in addition to the
44	with respect to b
40	the other hand if

APPENDIX G

A Full List of Political Bundles in the Pathway Corpus

Frequency	Lexical Bundle
6	a firm stand against
6	democracy in the middle
6	democracy in the region
5	avoid a military conflict
5	china on economic issues
5	firm stand against iran
5	important to avoid a
5	important to take a
5	more important to avoid
5	more important to take
5	Administration's handling of the
5	to avoid a military
5	with china on economic

APPENDIX H

A Full List of Referential Bundles in the Pathway Corpus

Frequency	Lexical Bundle
19	in the middle east
12	in the united states
9	in the u s
8	at the university of
8	in the case of
6	at the same time
6	in the context of
6	for the united states
5	in the form of
5	as a result of
5	from the united states

APPENDIX I

A Full List of Medical Bundles in the Pathway Corpus

Frequency	Lexical Bundle
7	a placebo controlled trial
6	the use of placebo
6	the world health organization
5	placebo controlled trials are
5	placebo controlled trials of
5	use of placebo controls

APPENDIX J

A Full List of Stance Bundles in the Pathway Corpus

Frequency	Lexical Bundle
17	is more important to
17	it is more important
16	say it is more
7	are more likely to
5	a very serious problem

APPENDIX K

A Full List of Discourse Organizing Bundles in the Pathway Corpus

Frequency	Lexical Bundle
9	on the other hand
5	when it comes to