Lexical Bundles in University First Year Economics Textbooks

by

Ayman Sholkani

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of
the requirements for the degree of

Master of Arts

in

Applied Linguistics and Discourse Studies

Carleton University
Ottawa, Ontario

©2018
Ayman Sholkani

Abstract

As part of formulaic language, lexical bundles (LBs) have been a rich topic of study for the last few decades. Following a frequency-driven approach, the present corpus-based study aims to identify the most frequently used LBs in first year Economics textbooks used in a number of Canadian universities. It also aims to classify the LBs according to the functional taxonomy of lexical bundles suggested in Biber et al. (2004a). The First Year Economics Textbook Corpus (FYETC) is a specialized corpus built especially for the purpose of this study. The text analysis software, *WordSmith Tools 6.0*, was used to extract the LBs from FYETC. To verify the results, three expert judges reviewed the FYETC LBs list. The results show that referential bundles are dominant in FYETC, and that quantity reference is by far the biggest subcategory. The final product of the study is a list of 165 LBs classified functionally along with some suggestions to improve the performance of the functional taxonomy. The findings of this paper can be of interest to EAP/ESP teachers, curriculum designers, and/or business and economics students.

Keywords: formulaic language, lexical bundles, corpus studies, WordSmith Tools, FYETC, functional taxonomy, economics

Acknowledgements

There have been many people who helped me accomplish this work. To each and every one of them, I would like to express my deepest thanks and heartfelt gratitude. First, I would like to thank my thesis supervisor, Associate Professor David Wood, the Director of the School of Linguistics and Language Studies at Carleton University. I am indebted to him for his patience and continuous support throughout my years of study at Carleton University. It is a great privilege for me to have worked under your supervision. I would also like to thank Professor Ellen Cray for her eye-opening guidance and for encouraging me to do my Master of Arts degree in the first place. Professor Cray showed me that she believed in my potential to pursue my studies. Also, I would like to thank Professor Eva Kartchava whose office door was always open whenever I needed her guidance. She always found the time to help me. I cannot thank her enough. Thanks also to Professor Guillaume Gentil who introduced me to corpus analysis and taught me how to make use of such a great invention in my studies. A big ‘thank you’ must go also to Joan Grant, whose kindness and helpfulness are an example to follow.

I would also like to extend my special thanks to my colleagues Lin Chen, Patricia Severenuk, and Elshimaa Eltayeb for their valuable feedback on the FYETC LBs List. Without their feedback and help, I would not have been able to achieve this work.

Finally, I would like to express my deepest appreciation to my family (my mother, my father, and my wife). There are not enough words to describe my gratitude for their endless support and countless favours. I am forever indebted to them. This accomplishment would not have been possible without their reliable and continuous support.

Table of Contents

Abstract.....	ii
Acknowledgements.....	iii
Table of Contents.....	iv
List of Tables	v
List of Figures	vi
1. Introduction.....	1
2. Literature Review.....	8
2.1 Theoretical Background on Formulaic Language.....	8
2.1.1 Nature and characteristics of formulaic language.	10
2.1.2 Mental processing of formulaic language.	12
2.1.3 Identification of formulaic language.	15
2.2 Lexical Bundles	24
2.2.1 Research on lexical bundles.	25
2.2.2 Identification of lexical bundles.....	28
2.2.3 Structural characteristics of lexical bundles.....	30
2.2.4 Pragmatic functions of lexical bundles.	32
2.3 Corpus Linguistics	38
2.3.1 Introduction to corpus linguistics.....	38
2.3.2 Types of corpora.	39
3. Methodology	45
3.1 Choice of the Setting.....	45
3.2 Choice of the Textbooks	48
3.3 WordSmith Tools 6.0 and Building the Corpus.....	49
3.4 The Identification Criteria of Lexical Bundles	55
4. Results and Discussion	58
4.1 Results.....	59
4.2 Discussion.....	66
5. Conclusion	83
References.....	88

List of Tables

Table 1 The Functional Taxonomy of Lexical Bundles (Biber et al., 2004a)	36
Table 2 Stance Bundles in FYETC LBs List	59
Table 3 Discourse Organizing Bundles in FYETC LBs List.....	60
Table 4 Referential Bundles in FYETC LBs List.....	62

List of Figures

Figure 1 Distribution of inte'l students in Canada by province and territory, 2004-2010.....	46
Figure 2 Top 5 fields of study for inte'l university students, Canada, 2004-2005 to 2013-2014.	47
Figure 3 A WordList Index file is created as the first step of building the FYETC	50
Figure 4 Computing the clusters	50
Figure 5 Cluster choices.....	52
Figure 6 Saving results as Excel spreadsheet for easier access and editing	53
Figure 7 Raw results before expert/human judgment contain irrelevant data	54
Figure 8 The distribution of FYETC LBs according to function.....	70
Figure 9 The dominance of specification of attributes as part of referential expressions	71
Figure 10 The dominance of quantity bundles as part of the specification of attributes	71
Figure 11 The functional distribution of all LBs in FYETC	73
Figure 12 The frequencies of 'in case of the' and 'in case of' in the COCA	81

1. Introduction

Over the past decades, the nature and role of formulaic language has been the focus of many research studies. Formulaic language is a general term for multi-word strings that act as single units such as idiomatic expressions (e.g., *speaking of the devil*, *hold your horses*, *beat around the bush*), fixed phrases (e.g., *by and large*, *on the other hand*, *all in all*, *in a nutshell*), collocations (e.g., *take a break*, *make a difference*), lexical bundles (e.g., *in the case of*, *as a result of*, *are more likely to*), discontinuous expressions or fillable slots (e.g., *not only...but also*, *if X ...then Y...*) among others. Numerous attempts have been made to define this ubiquitous phenomenon. Wray and Perkins (2000) define formulaic language as “A sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar” (p. 1). Likewise, Erman and Warren (2000) refer to formulaic language as frequent “combinations of at least two words favoured by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization” (p. 31). Also, Wood (2006) identifies formulaic language as “fixed combinations of words that have a range of functions and uses in speech production and communication, and seems to be cognitively stored and retrieved by speakers as if they were single words” (p. 1).

Many technical terms have been used to refer to formulaic language, such as multi-word strings, multi-word sequences, lexical phrases, wordstrings, phraseologisms, prefabricated language, prefabricated patterns, prefabs, set phrases, routines, fixed expressions, formulae/formulas, and more. Regardless of the name used to refer to these sequences, their common

characteristics can be summarized as being multi-word constructions, having holistic meanings, and believed to have “a processing advantage over creatively generated language” (Conklin & Schmitt, 2008, p. 72). This means that a formulaic sequence should contain more than one word and has a holistic meaning that is not necessarily understood from the meanings of its single components (e.g. *It rains cats and dogs, gone pear-shaped, talk of the devil, beat around the bush*). Nattinger and DeCarrico (1992) state that lexical phrases/ formulaic sequences “are similar to lexicon in being treated as units” (p. 36). Also, a formulaic sequence is believed by many researchers (e.g. Wray and Perkins, 2000, and Wray, 2002, among others) to be mentally stored and retrieved as a whole at the time of language production, which leads to better fluency and easier communication. However, as Wood (2015) states, the conception of formulaic language sequences being stored and retrieved as wholes “has not been proven beyond a doubt” (p. 58). He argues that “Despite the amount of theorizing and reasoning which can be brought to bear on the notion that formulaic sequences are stored and retrieved as wholes, researchers have struggled to find ways to uncover empirical support [...] It is obvious that we store individual words in the mental lexicon, but it is less clear whether we store multiword items in a similar way” (p. 60). This point will be discussed in more detail in the following chapter of this paper.

A growing body of research in psycholinguistics, second language acquisition, and corpus linguistics (e.g., Kjellmer, 1984; Baayen and Lieber, 1991; Altenberg, 1990, 1993, 1998; Barkema, 1993; Biber et al., 1999; Erman & Warren, 2000; Foster, 2001) has provided sufficient evidence of the highly formulaic nature of language. Glucksberg (1989) found that native speakers produce four multiword sequences in every minute of spoken discourse. It is, also, believed that the majority of natural language, written or spoken, consists largely of collocational ‘sets’ or ‘frameworks’ (Renouf and Sinclair, 1991; Renouf, 1992). As Sinclair (1991) notes “all

the evidence points to an underlying rigidity of phraseology, despite a rich superficial variation" (p. 121). Formulaic language is an omnipresent phenomenon. Its importance stems not only from the belief that it facilitates language acquisition and production, but also from the belief that a large portion of everyday language is 'ready-made' or 'prefabricated' rather than completely constructed word by word. Wray and Perkins (2000) argue that formulaic language works as a processing shortcut by being stored and retrieved holistically from memory at the time of use rather than being generated anew each time.

The formulaic nature of language was proposed early on by Sinclair's (1991) dichotomy model of *the open-choice principle* and *the idiom principle*, by which he argued against the Chomskyan view of generative grammar as the only way of creating meaning. According to Sinclair (1991), the *open-choice principle* is a "slot-and-filler" model which allows a range of grammatical choices and "a string of lexical choice points" (p. 109) to be used in a language unit. He indicates that "At each point where a unit is completed [a word, or a phrase, or a clause], a large choice opens up and the only restraint is grammaticalness" (Sinclair, 1991, p. 109). The *idiom principle*, on the other hand, posits that "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments" (Sinclair, 1991, p. 110). Sinclair (1991) argued that most texts are typically composed of multiword expressions that 'constitute single choices' in the mental lexicon and language meaning is conventionalized by the *idiom principle*.

The belief that formulaic language facilitates communication and language learning is somehow pervasive in the literature and research. The usefulness of acquiring and learning formulaic language extends from first language (L1) and second language (L2) acquisition, to fluency and pragmatic competence. Wood (2015) indicates "there is a certain amount of

evidence of formulaic sequences being used as learning and communication strategies by children in first and second language acquisition” (p. 67). That is, children not only acquire their L1 or L2 through putting ready-made chunks to use with no or little analysis; but also subconsciously analyze those acquired chunks, infer language rules and structures, and use this knowledge to further develop their overall language competence.

The role of formulaic language in speech fluency and pragmatic competence has been emphasized in many studies such as Pawley and Syder (1983); Levelt (1989); Kuiper (1996, 2004); Wray (2000, 2002); Wood (2006, 2009a, 2009b, 2010a, 2015); Taguchi (2007); and Lin (2010). These studies show a significant positive correlation between learners’ knowledge and use of formulaic sequences on the one hand, and their level of language proficiency on the other hand. Boers, Eyckmans, Kappel, Stengers, and Demecheleer (2006) explain that the use of formulaic sequences can help second language speakers sound more native-like, make fewer errors due to using prefabricated expressions, and speak in an easier and more fluent manner. Wood (2015, p. 86) notes that *speed of speech*, *pauses and hesitations*, and *length of runs* are three measurable aspects of speech that can show the utility of formulaic sequences in promoting speech fluency. That is, a fluent speaker would speak relatively faster than other speakers, has less pauses and hesitations, and produce longer runs of speech before pausing. These aspects of fluent speech, as Wood argues, can be measured and quantified. Wray (2000) affirms that the comprehensive knowledge and having a good command of formulaic language is a pivotal factor of successful language learning (p. 463). She states that, “the first function identified for formulaic sequences is that they save precious processing resources” (p. 473). This notion is in line with the findings of many studies. For example, Jiang and Nekrasova (2007) conclude, “both native and non-native speakers responded to the formulaic sequences significantly faster and

with fewer errors than they did to non-formulaic sequences” (p. 433). This means that speech and communication is conspicuously facilitated when interlocutors use formulaic sequences. This can be linked to the notion that using ready-made or prefabricated sequences that are processed holistically saves a lot of cognitive effort and allows for more focus on meaning and on maintaining the flow of speech, which would ultimately lead to faster, more meaningful, and more fluent speech.

Under the umbrella of formulaic language, there is a big group of multi-word constructions that are generally non-idiomatic, salient in meaning, and structurally incomplete. This group is referred to as lexical bundles (LBs). Biber and Barbieri (2007, p. 269) explain that lexical bundles form the vast majority of formulaic language sequences. LBs are semantically transparent, but are also incomplete structural units. Biber and Barbieri (2007) indicate that lexical bundles are somehow flexible, and that this specific characteristic allows them to be useful in shaping language and facilitating communication. That is, lexical bundles, unlike idioms, can have many variations that allow them to be extended; e.g. a shorter lexical bundle like ‘*as a result*’ can be extended to longer ones like ‘*as a result of*’ or ‘*as a result of the*’.

The advancement in technology allowed for better tools of investigation of the most frequent and recurrent word clusters. Researchers created general and specialized corpora in order to find those recurrent multiword constructions through using text analysis software. One of these software programmes is *WordSmith Tools 6.0*, which is used in this study.

Many research papers have been written on lexical bundles and discussed how they exist in different registers and disciplines. The research conducted in some of them resulted in lists of lexical bundles that seem to be discipline-bound to some extent. In other words, the lexical

bundles that are used in one register and/ or discipline are not necessarily the same as the ones used in a different register/ discipline.

To our knowledge, there is no research done on lexical bundles in first year Economics textbooks. The significance of this study stems from the fact that almost 40% of the international students in Canada study in Ontario, according to Citizenship and Immigration Canada; and that more than 25% of them study business as their field of study, according to Statistics Canada (2017). Also, business, as a field of study, requires Economics as a core subject. It is also important to note that first year textbooks are usually the main focus of EAP courses, since first year students are more likely to need this type of academic support. Accordingly, having a list with the most frequent lexical bundles in these textbooks can prove helpful not only for learners; but also for teachers, course and material developers, and researchers. Therefore, this study is an attempt to cover this gap and to find out what lexical bundles were the most frequent ones in first year Economics textbooks; and what functions they serve. To this end, a specialized corpus was built out of eight Economics textbooks. The First Year Economics Textbooks Corpus (FYETC) consists of 1,346,658 words. The research questions for this study are:

1. What are the most frequent lexical bundles in first year Economics textbooks that are used in Canadian universities?
2. What functions do these lexical bundles serve?

This paper contains a literature review section that briefly displays some of the most significant work written on formulaic language, lexical bundles, and corpus studies. Next, the paper shows the methodology and procedure that the researcher of this study followed in his attempt to answer the research questions listed above. Then, the results are displayed and

followed by a discussion of the significant observations that can be seen in the results. Finally, a conclusion section follows to wrap up and summarize the study.

2. Literature Review

This chapter presents a synthesis of research on formulaic language and its significance in language acquisition and production. It offers theoretical background information on lexical bundles, which is the focus of the present study, along with the role of corpus studies to assist and even direct this area of research. The chapter is divided into three major sections. The first section presents the theoretical background of formulaic language research. It explores the growth of literature on formulaic language in spoken and written discourse. It also highlights the particular nature and characteristics of formulaic language along with the terminology used to describe this phenomenon. It then outlines key arguments in the field regarding the mental processing and holistic nature of formulaic language. The section closes with a brief discussion of various research methods and techniques that have been used in the literature to identify formulaic language. The second section presents an overview of lexical bundles (the most recurring multiword sequences). It investigates some of the research conducted on lexical bundles and elucidates the different criteria used to identify the lexical bundles in discourse. It also provides an illustration of the structural characteristics and pragmatic functions of lexical bundles. Finally, the third section of this review highlights the significant role of corpus studies in formulaic language research and explores types of corpora (i.e., general, specialized, learner, comparable, and theoretical corpora).

2.1 Theoretical Background on Formulaic Language

Formulaic language has become a focal point of interest for researchers in the fields of linguistics and applied linguistics. Wood (2015) notes that “virtually every aspect of communication and language is linked to it: pragmatics; discourse; fluency in speech and

writing; first and second language acquisition; and cognitive processing of language” (p. 3). The relevance of formulaic language lies not only in its central role in facilitating communication and language learning, but also in its high representation in natural language. For example, in a small corpus study of spoken language in a TV quiz show, Jackendoff (1995) argued that the prominence of formulaic sequences might be equal if not more essential than the glossary of single words. Similarly, in examining the London Lund Corpus, Altenberg (1998) found that more than 80% of the words sequences are formulaic language. Erman and Warren (2000) found that formulaic language composed 58.6% of spoken English discourse and 52.3% of written discourse in the corpus they analyzed. Foster (2001) likewise found that 32.3% of the analyzed spontaneous speech of native speakers was made up of formulaic language. These findings indicate how widespread formulaic language is in English discourse.

The formulaic nature of language and the central role of formulaicity and phraseological competence as part of fluent, native-like and idiomatic language use has been a central point of investigation among researchers (Hoey, 2005; Pawley & Syder, 1983; Wray & Perkins, 2000). Research has suggested that to master a language, learners should develop adequate knowledge of these word sequences, where formulaic language competence is the key to second language fluency and native-like selection (Pawley & Syder, 1983). Wray (2000) notes that a major component of mastering a new language is that learners become familiar with and sensitive to native speakers’ preference of certain word sequences over others. The mastery of these word sequences is more likely to distinguish novice from proficient L2 learners. Wray (2000) also observes that second language (L2) learners are more likely to use particular utterances repeatedly in their language production instead of using other possible grammatical expressions

that might express the same concept. In this respect, the recognized role in L2 learning has led to formulaic sequences becoming one of the major issues in applied linguistics.

2.1.1 Nature and characteristics of formulaic language. As noted above, research on language acquisition has found that conventionalized expressions form a substantial portion of discourse. Sinclair (1991) suggested *the idiom principle* and *the open choice principle*, that is, most texts are predominantly composed of multiword expressions that constitute single choices in the mental lexicon. According to Sinclair's principle, speakers and writers tend to depend on units of meaning expressed by pre-constructed phrases or fixed expressions instead of selecting the words they use one at a time. These pre-constructed phrases have been identified as prefabricated chunks that are stored in and retrieved holistically from the memory (Chen, 2009; Conklin & Schmitt, 2008; Pawley & Syder, 1983; Tremblay & Baayen, 2010; Weinert, 1995; Wood, 2001, 2006, 2010a, 2010b; Wray, 2002).

While Sinclair's principle is widely accepted in formulaic language research, there is otherwise little consensus in the field. As Biber, Conrad, and Cortes (2004) state, "despite the general consensus on the importance of multi-word units, there is surprisingly little agreement on their defining characteristics, the methodologies to identify them, or even what to call them; and, as a result, there is little agreement across studies on the specific set of multi-word units worthy of description" (372). The terminology is indeed varied; Biber et al.'s "multi-word units" have also been identified in the literature as "formulaic sequences" (Wray, 2000, 2002), "formulaic expressions" (Simpson, 2004), "fixed expressions" (Moon, 1998a), "lexical phrases" (Nattinger and DeCarrico, 1992), "multiword lexical units" (Cowie, 1992), "recurrent word combinations" (Altenberg, 1998), "collocation" (Partington, 1998), "multiword constructions" (Liu, 2012; Wood & Appel, 2014), and "lexical bundles" (Biber et al., 1999; Cortes, 2002; Hyland, 2008).

Wray and Perkins (2000) comment on the remarkable range of labels –as many as forty terms– that have been used to describe formulaic language, including: “chunks, clichés, co-ordinate constructions, collocations, conventionalized forms, fixed expressions, formulas/formulae, frozen phrases, idioms, lexical/lexicalized phrases, holophrases, noncompositional, gambits, ready-made expressions, praxons” (p. 3). This list later grew to include more than fifty terms (Wray, 2002, p. 9). Further, with the development of corpus linguistics studies, new technology-derived terms have been added such as *n-grams* and *concgrams*.

Wood (2015) proposes that a possible reason of this divergence is the variety of fields- and therefore theoretical and methodological perspectives- informing this area of research: “researchers were not all examining the exact same phenomenon, and were frequently working in quite separate areas of linguistics [...], even in areas outside of linguistics in fields as diverse as social anthropology and neurology” (p. 35).

It has long been agreed that idioms, proverbs, and sayings represent the obvious types of formulaic language since they “noticeably operate as single units at some level, even though their form consists of multiple orthographic words” (Schmitt, 2004). Nevertheless, since the rise of computerized corpus studies (e.g., Biber et al., 1999), new interest in lexical patterning beyond these obvious multiword units has flourished among researchers. Examples of formulaic language categories that have been a focus of research over the years might include: idioms (e.g., *kick the bucket*, *beat around the bush*), collocations (e.g., *take a risk*, *do the homework*), lexical phrases (e.g., *long time no see*, *so far so good*), lexical bundles (e.g., *in case of*, *on the other hand*), proverbs (e.g., *the ball is in your court*), phrasal verbs (e.g., *looking forward to*, *run into*), metaphor (e.g., *speak of the devil*). The list goes on, however, Wood (2015) indicates that “other types of formulaic sequences seem lost in the shuffle, uncategorized but intuitively formulaic...

[such as] *and then or sooner or later*” (p. 50). He also indicates, “some types of word strings are difficult to classify...[and] some categories overlap with others” (p. 50).

2.1.2 Mental processing of formulaic language. The mental processing of formulaic language has been a point of debate among applied linguists and researchers. Wood (2015) points out, “despite the amount of theorizing and reasoning which can be brought to bear on the notion that formulaic sequences are stored and retrieved as wholes, researchers have struggled to find ways to uncover empirical support” (p. 60). The assumption of the holistic nature of formulaic language as proposed by Sinclair (1991) has been endorsed by different definitions of formulaic language. For example, Weinert (1995) defines formulaic sequences as “multi-word or multi-form strings which are produced or recalled as a whole chunk, much like an individual lexical item, rather than being generated from individual lexical items/forms with linguistic rules” (p. 182). That is, these sequences “are retrieved based on the order of their parts, by phonological cues or by recall of first and last words first, which then trigger recall of the entire sequence (Wood, 2015, p. 58). Likewise, Wray (2002) claims that a formulaic sequence is “a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar” (p. 9). Wood (2010a) also refers to formulaic language as “multiword or polymorphemic units of language, stored in memory as if they are single lexical units, and recalled and produced as wholes” (p. 42). Although these definitions are helpful in giving an overall sense of the mental processing and storage nature of formulaic language, there is still a need for more empirical evidence that formulaic sequences are ‘stored as wholes’ (Wood, 2015).

It is worth mentioning that the evidence for the mental representation and holistic processing of formulaic language is largely derived from research on adult native speakers and proficient L2 speakers and, more particularly, from research on idiom processing. For many years, idioms, a type of nontransparent language expression that cannot be understood from the meaning of the words in the expression, have been viewed as the prototype of formulaic language (Nekrasova, 2009). Although idioms represent a small category of formulaic language, researchers use the fact that idioms have both figurative and literal meanings in an attempt to unveil the nature of the mental processing of formulaic language. To do so, researchers tend to identify and compare the activation time of the figurative meaning of idioms with literal or novel word strings by measuring how fast people react to both of the meanings (e.g., Gibbs & Gonzalez, 1985; Swinney & Cutler, 1979; Van Lancker, Canter, & Terbeek, 1981). According to Swinney and Cutler's (1979) Lexical Representation Hypothesis, when proficient speakers confront ambiguous word strings (i.e., idioms), the figurative meaning is always activated first and faster than the literal one. This is because idioms seem to be stored like morphologically complex words. While the figurative meaning requires a singular interpretation of the word sequence (i.e., as a whole), the literal one involves more linguistic analysis. In a similar study, Gibbs and Gonzales (1985) compared the processing time of idiomatic expressions among thirty-four undergraduate native speakers of English and declared, "retrieving idioms from the lexicon operates in a similar manner to normal lexical access processes" (p. 257). Interestingly, in comparing the processing time of different types of idioms (i.e. decomposable and non-decomposable idioms), Gibbs, Nayak and Cutting's (1989) Idiom Decomposition Hypothesis found that decomposable idioms, which can be linguistically analyzed, enjoy the same

processing advantage as non-decomposable idioms. This finding is probably due to the consistency between the meaning of the idiom and its linguistic analysis.

Meanwhile, research on nonnative speakers seems to be complex. It is unsurprising that nonnative speakers react differently or more slowly than native speakers; nonnative speakers are unlikely to have been exposed to or have used idioms very often (Conklin & Schmitt, 2012). Consequently, nonnative speakers are more likely to grammatically analyze novel phrases or word strings. Cieslicka (2006) found that nonnative speakers responded faster to literal interpretations of idioms than to figurative ones. As Van Lancker-Sidtis (2003) found, while native speakers were able to use pronunciation and word stress or accent combination to determine whether particular word strings are used in their figurative or literal way, nonnative speakers were not able to do so. Siyanova-Chanturia, Conklin, and Schmitt (2011) used eye-movement tracking to measure the speed of processing time of idioms in reading passages and found that native speakers processed idioms faster than novel word strings, whereas nonnative speakers processed idioms more slowly than other word strings.

While these studies might validate the holistic nature of formulaic language, it cannot be forgotten that idioms represent only a small category of formulaic language and more empirical research still needed on the various categories of formulaic language. As Wood (2015) notes,

While this research involving idioms presents some tantalizing evidence of processing phenomena, unfortunately it is hard to generalize from it to any great extent. For one thing, the fact that idioms often have two interpretations, the literal and figurative, can slow processing as choices have to be made while processing them. As well, idioms are not all equally transparent or decompositional, which makes it hard to present sweeping conclusions even

about idiom processing, let alone processing of all other types of formulaic language. And finally, it is a fact that idioms represent a subset of formulaic language and are actually not very frequent in language (p. 61).

2.1.3 Identification of formulaic language. As mentioned earlier, there is still little agreement across studies on different issues concerning the identification of formulaic language—for example, defining the minimum number of words that comprise a word sequence, or identifying the features of word combinations and how to label each type. Wray (2000) indicates, “the difficulty, in short, is that formulaic and non-formulaic language may sometimes look identical” (p. 467). Types of formulaic language range from non-compositional, semantically opaque and syntactically irregular sequences (e.g., idioms, phrasal verbs, nominal compounds, and so on) to more compositional semantically transparent and syntactically regular word combinations (e.g., lexical bundles). Wray (2000) further highlights that “there is no clear dividing line between what is regular and irregular” (p. 485). That is, while the overall meaning of compositional word strings can be determined from the meaning of their parts (e.g., *as a result of* or *in case of*), the general meaning of non-compositional sequences cannot be understood from the meaning of their components (e.g., *by and large*). *By and large* is a clear example of the syntactic irregularity of some of the formulaic sequences, as it violates the parallel structure rule, which entails that whenever we list some items or actions, we should use the same grammatical forms (i.e., *noun* and *noun*). Further complicating the matter is that in some cases formulaic sequences include discontinuous expressions or fillable slots (e.g., *not only...but also*) that can be difficult to identify (Wood, 2015).

Researchers have attempted to classify and determine formulaic language based on their variability or fixedness, institutionalization, frequency of occurrence (i.e., if a sequence is

frequent in a corpus), syntactic structure, and discourse or pragmatic functions in spoken or written discourse (Becker, 1975; Coulmas, 1979; Howarth, 1996; Yorio, 1980). It is worth mentioning that most of the evidence for identifying formulaic language is drawn from research on spoken language, yet, there are major similarities between both spoken and written language when it comes to the identification criteria of formulaic language, such as frequency of use and specific pragmatic functions (Biber et al., 2004). Nattinger and DeCarrico (1992) offer a classification that moves from frozen or semi-idiomatic sequences to more freely compositional utterances. They classify formulaic language into:

1. Polywords; e.g. *'once and for all'*
2. Institutionalized expressions; e.g. *'Nice meeting you!'*
3. Phrasal constraints; e.g. *'for ____'* (as in *for example/for instance*), *'a year ago'*
4. Sentence builders; e.g. *'not only X, but also Y'*, *'let me start with X'*

Wray (2002) and Wood (2015) provide a comprehensive illustration of the different criteria and measurements that have been used or proposed in previous research to identify formulaic language, including: psycholinguistic analysis, phonological analysis, frequency and statistical analysis, intuition and native speaker judgment, and the use of criteria checklists. These means of identification and measurement are briefly discussed in the following sections.

2.1.3.1 psycholinguistic analysis. Some scholars are in favour of using fluency and speech rate to determine the prosodic features of formulaic sequences. Various studies have used different means to measure the processing time of formulaic language, such as eye movement tracking (e.g., Underwood, Schmitt & Galpin, 2004), reaction times (e.g., Conklin & Schmitt, 2012), and electrophysiological measures (ERP) (e.g., Tremblay & Baayen, 2010). However, psycholinguistic measures seem to provide a limited view of the use of a sequence, since they

mainly depend on a small subset category of formulaic language (e.g., idioms) that might not be common in use for some individual speakers. Wood (2015) notes that these studies “do not usually help us to know how common a given sequence may be in actual use in the community, and the formulaic sequences identified in these ways may include rare, unusual, or one-off sequences that the speaker has tended to use for a variety of idiosyncratic reasons” (pp. 22-23).

2.1.3.2 phonological analysis. In spoken language, phonological coherence was used to help determine the formulaicity of word sequences. The concept of the phonological coherence can be attributed to Peters (1977, 1983) who pinpoints the idea of what can be acquired as a *chunk*. She noticed that children can produce word sequences that surpassed their linguistic competence with no internal hesitations, which gave flow to the rest of their speech. Echoing Peters, Hickey (1993) indicates that phonological coherence is a key to formulaic language in children’s speech. However, the situation in adult speech is less clear (Wood, 2015), though there is a belief that a similar dynamic process as in children’s language applies to adults (Moon, 1997; Wray, 2004).

Coulmas (1979) and Peters (1983) argue that a formulaic sequence must be at least two morphemes long and phonologically coherent (i.e., uttered without internal hesitation or pausing). They also hold that word combinations that are more complex and greater in length than the rest of the speaker’s output are formulaic language. Taking a different approach, Bybee (2002) notes that high frequency sequences undergo phonological reduction more quickly than other word phrases. This phonological reduction can be located as schwa reduction or sometimes syllables reduction (e.g., *t/d deletion*). Other research has pointed to stress placement, intonation, pausing nature and falling/rising tone characteristics as indicators to the formulaicity of word sequences. For example, Ashby (2006) found that stress placement, accent pattern, and falling

tone are essential phonological characteristics that help distinguish idioms that have the same pattern as their literal version. Similarly, Lin and Adolphs (2009) examined the Nottingham Corpus of Learner English and found that the sequence “*I don’t know why*” was used as a single intonation unit in more than fifty percent of the cases.

Despite the significance of phonological characteristics as a key feature of formulaic language, Wood (2015) cautions, “as with psycholinguistic methods, phonological coherence provides a limited or partial sense of formulaicity. For one thing, phonological coherence is limited to analysis of spoken language only. As well, it only relates to formulas used by a particular speaker, and analysis is limited by the quality of the audio data recorded (p. 23).

2.1.3.3 frequency and statistical analysis. With the emergence of the field of corpus linguistics came the statistical approach to the identification of formulaic language. The statistical approach helps identify word combinations, specified headwords, or even discontinuous sequences of word strings by locating word sequences (i.e., two, three, or more) that are generally recurrent throughout the corpus. To do so, certain specifications and criteria (i.e., minimum frequency cutoffs and minimum length of word strings) must be set before analyzing a corpus for the multiword combinations that meet the assigned criteria. Different frequency cutoffs, typically ranging between 10 and 40 occurrences per million words, are used based on the size or type of the corpus (i.e., large general corpora vs. specialized corpora). For example, Biber et al. (2004a) uses a frequency cutoff of 40 occurrence per million words, whereas Cortes (2004, 2008), Hyland (2008), and Chen (2008) set a cutoff of 20 times per million, while in a much smaller corpus, Levy’s (2003) uses a cutoff of 10 times per million words. To identify word clusters in the corpus, text analysis software such as *Wordsmith Tools* and *SPSS* can be used. Using this type of software allows researchers to identify the word

clusters and count the occurrences. This approach will be further explored in the methodology section of this paper.

Although frequency-based analysis is a remarkable valuable tool for the identification of formulaic language, it has some drawbacks. As Wray (2002) notes, “raw frequency is not an adequate measure of formulaicity” (p. 30). Wray further explains:

It seems, on the surface, entirely reasonable to use computer searches to identify common strings of words, and to establish a certain frequency threshold as the criterion for calling a string ‘formulaic’ [...] problems regarding the procedures of frequency counts can be identified. Firstly, corpora are probably unable to capture the true distribution of certain kinds of formulaic sequences...The second serious problem is that the tools used in corpus analysis are no more able to help decide where the boundaries between formulaic sequences fall than native speaker judges are (pp. 25-28).

To illustrate, Foster (2001) indicates, “Even a corpus as large as The Bank of English at the University of Birmingham, now nearly three hundred million words, fails to show even a single example of many phrases that would be considered a normal part of any native speaker’s repertoire” (p. 81).

Biber et al. (2004a) likewise argue for the limitations of frequency data, “Frequency is only one measure of the extent to which a multiword sequence is prefabricated [...] We do not regard frequency data as explanatory. In fact, we would argue the opposite: frequency data identifies patterns that must be explained” (p. 376). Wood (2015) points to particular limitations of adopting frequency-based approach with small corpora: “It would be difficult, for example, to use only frequency as a criterion for identifying formulaicity in a set of transcribed conversations

on a range of topics. Some items which we might consider formulaic might arise only once or twice in such a data set” (p. 21). Accordingly, there is a need to complement frequency criteria with other approaches that can “help the researcher to narrow down the search and be consistent... [that is] to exploit the researcher’s own intuitions” (Wray & Namba, 2003, p. 26).

2.1.3.4 intuition and native speaker judgment. Intuition and native speaker judgment apparently play an effective role in identifying formulaic language. It helps overcome the aforementioned limitations of using frequency alone as an identification criterion. As Wray (2002) and Wood (2015) indicate that the nature and the size of a given corpus play decisive roles in the results that software gives. For example, in cases where L2 learner corpora are used, the formulaic sequences that the software recognizes can be nonstandard. Accordingly, there is a need for what Wray (2002) calls “the application of common sense” (p. 28).

However, different researchers have raised concerns regarding the reliability of this method to determine what qualifies a sequence as formulaic. For example, Sinclair (1991) argues that intuition has been revealed as an unreliable means of investigating facts of language use. He states, “human intuition about language is highly specific and not at all a good guide to what actually happens when the same people actually use language” (p. 4). Wray and Namba (2003) also indicate, “intuition is not a direct window onto word distribution in text...[and] it remains true to say that we have very little understanding of exactly what, in psycholinguistic terms, intuition is reflecting” (p. 27). Wray (2002) notes, “Intuition is little understood and not highly predictable” (p. 20). Wray (2002, p. 23) identifies internal problems with the use of raw intuition for identifying formulaic language:

- It has to be restricted to smaller data sets.

- There is no way to avoid inherent inconsistency within the range of judgments made by an individual, because of factors such as tiredness and unintended alterations in the judgment thresholds across time. Computers do not suffer from such problem.
- There is a danger of significant variation between judges.
- There is no guarantee that formulaic sequences have firm borders in the sense that we have come to expect in the context of phrase structure analysis, so, even if all judges were actually operating identical criteria, for any given string there may not be one single answer to find.
- The application of intuition makes subjective externalized insights valid, at the expense of any knowledge we may have that is not available at the surface level of our awareness

2.1.3.5 criteria checklists. With the above challenges of intuition in mind, the idea of criterion-based checklists emerged to guide judgments. A number of checklists have been developed and used in various studies for different purposes for general or specific populations. Wood (2015, pp. 25-32) gives a comprehensive discussion of some of these checklists including those developed by Coulmas (1979), Peters (1983), Wood (2006, 2009b, 2010), and Wray and Namba (2003). The use of checklists might help overcome the limitations of other means of identification, e.g., based on frequency and/or phonological characteristics as described in previous sections.

Generally, checklists are used as follows: judges first study the different criteria in a checklist, and then go through the corpus and apply these criteria to identify formulaic sequences. Each checklist outlines different conditions that can be used as a guide to determine the formulaicity of word sequences. Coulmas' (1979) early checklist, for example, focuses mainly on phonological characteristics and the length of the word sequence. According to

Coulmas' (1979) checklist, to be considered formulaic, word sequences need to be at least two morphemes long (i.e., two words) and cohere phonologically. Also, a word sequence might be more grammatically advanced compared to other language output that a learner produces. That is, the utterance involves more syntactic and complex language than the norm, and is produced without pausing or internal hesitation. Besides, a word sequence is typically situationally dependent and shared within a community, where the sequence is frequently used in the same form.

Peters (1983) proposed a checklist of six criteria for identifying formulaic sequences in children's first language speech that addressed: phonological characteristics of word sequences, length and complexity of a sequence compared to other language output, situational dependence of word strings, frequency and fixedness in form, and the non-productive use of rules in a word sequence. In a later attempt to establish reliable justifications for native speaker intuitive judgment, Wray and Namba (2003) propose a checklist of eleven criteria that can be used to "capture the major externally detectable features of formulaic sequences" (p. 28). Their checklist was originally designed to determine formulaicity in bilingual children utterances, yet the points that this checklist addresses can be applicable for a range of child as well as adult native or non-native speakers. This checklist is based on a five-point Likert scale: *Strongly agree*, *Agree*, *Don't know or Not applicable*, *Disagree*, and *Strongly disagree*. Wray and Namba (2003) clarify, "While *Agree* or *Strongly agree* may be taken as positive evidence of formulaicity, *Disagree* and *Strongly disagree* do not constitute evidence against formulaicity- just the absence of trait that sometimes indicates it" (p. 28). Similar to Peters' (1983) checklist, Wray and Namba's checklist makes reference to phonological characteristics and complexity, as well as the distinctive uses and community wide uses of word sequences.

In examining the role of formulaic language in facilitating second language fluency, Wood (2010) used a checklist of five criteria drawn from previous research on formulaic language as a guide to determine whether a sequence is formulaic. Wood (2010) indicates, “No particular criterion or combination of criteria were deemed as essential for a word combination to be marked as formulaic, they were guide only” (p. 111). Wood makes reference to: (1) phonological coherence and reduction (Coulmas, 1979; Wray, 2002); (2) the taxonomy used by Nattinger and DeCarrico (1992) that includes different syntactic word strings; (3) greater length/complexity than other output (e.g., expressions like *I don’t understand...or I would like*); (4) semantic irregularity, as in idioms and metaphors; and (5) syntactic irregularity (Wood, 2010, pp. 111-112). Notably, unlike Peters (1983) and Wray and Namba’s (2003) checklists, Wood’s (2010) does not make reference to frequency as an essential criterion in and of itself in identifying formulaic sequences in non-native speaker’s speech.

Although criteria checklists seem to rely mainly on native speaker judgment, the use of these checklists is first and foremost intended to be a reliable means of identifying formulaic language. However, despite variation in the criteria across researchers, all of these checklists depend to some extent on native speaker judgment in determining what is a formulaic sequence. Wood (2015) notes, “judgment checklists can help overcome the limitations of frequency-based psycholinguistic, or phonologically focused identification methods, and provide a sort of aggregate measure of formulaicity” (p. 31).

By and large, to obtain validity and reliability of the identification of formulaic sequences, researchers need to apply more than one approach and use a combination of measures, as none of the methods outlined in the preceding section is sufficient by itself. As Wray (2002) indicates, “Identification cannot be based on a single criterion, but rather needs to

draw on a suite of features” (p. 43). Similarly, Biber et al. (2004) confirm, “no single approach can provide the whole story. Rather, the overall importance of multi-word units in discourse can be fully understood only by undertaking empirical research studies from different perspectives” (p. 372).

2.2 Lexical Bundles

The development of corpus-based analysis software has attracted attention to the study of frequently occurring multiword combinations, variously referred to as “lexical bundles” (Biber et al., 1999), “recurrent word combinations” (Altenberg, 1998), “clusters” (Hyland, 2008a; Schmitt, Granger, & Adolphs, 2004), or “multiword constructions” (Liu, 2012; Wood & Appel, 2014). Biber et al. (1999) defined lexical bundles as “recurrent expressions, regardless of their idiomaticity, and regardless of their structural status” (p. 990). Lexical bundles are frequently used in published academic writing (e.g., journal articles) and the types of bundles differ according to the disciplines they are used in (Cortes et al., 2002). For example, using corpus-based analysis to identify lexical bundles in university textbooks, Biber (2006) found that academic disciplines use lexical bundles differently, where the natural and social sciences use lexical bundles more frequently than the humanities.

Most importantly for language teaching and learning in academic settings, Biber and Conrad (1999) indicate that the use and acquisition of these bundles do not seem to come naturally and may require focused instruction. Likewise, Cortes (2004) found:

Simple exposure to the frequent use of lexical bundles in published academic writing does not result in the acquisition of these expressions by university students...[and] a good way to help students come more familiar with lexical bundles could be to have them ‘*notice*’ the frequent use of these expressions...

becoming aware of different contexts and discourse functions they perform in academic disciplines (pp. 417-420).

2.2.1 Research on lexical bundles. The study of recurrent expressions has been an attractive area of research within the study of formulaic language over the last few decades. Altenberg (1993, 1998) initiated the use of frequency as a basic criterion in identifying particular word combinations and classifying them according to their functions. Altenberg (1998) examined recurrent word combinations in the London-Lund Corpus of Spoken English. He used a minimum frequency cutoff of ten occurrences per million words and narrowed his investigation to sequences with a length of three or more. He found that the majority of these word sequences were semantically transparent (e.g., *on the other hand, I don't think that, Yes I did*) and represented incomplete structural units, with only 10% of the sequences representing full clauses, while 76% took the form of clause constituents and 14% were incomplete phrases.

Following Altenberg's (1998) notion of recurrent word combinations, Biber et al. (1999), in their comprehensive study of English grammar, examined the Longman Corpus of Written and Spoken English (a large corpus of five million words per register). They identified the most frequent word sequences in conversation and academic prose. They proposed the concept of 'lexical bundles' and introduced the distribution criterion to avoid the over-representation of expressions that are idiosyncratically used. They examined the grammatical features of lexical bundles across four registers and concisely classified them according to their discourse functions in their original context (this classification was further developed by Biber and Conrad (1999)). Biber et al. (1999) found that the majority of lexical bundles in conversation are accidental and unplanned taking the form of *(pronoun)+verb+(complement)* (e.g., *it is going to be, I'd like to, I*

want you to), whereas, 60% of the lexical bundles in academic prose are noun phrases or prepositional phrases (e.g., *on the other hand*, *on the basis of*, *as a result of*).

Biber, Conrad, and Cortes (2003) introduced an initial functional taxonomy of four-word bundles in speech and writing (e.g., *I don't know what*, and *on the other hand*). This taxonomy was built based on the most frequent lexical bundles identified by Biber et al. (1999) in conversation and academic prose. It classifies lexical bundles into four main groups: (1) referential bundles, (2) text organizers, (3) stance bundles, and (4) interactional bundle. In an attempt to develop knowledge of the functions of the lexical bundles, Cortes (2004) used this initial taxonomy in a more discipline-specific context. She investigated the extent to which lexical bundles are unique to particular registers and how the use of these bundles by disciplinary professional writers differs from that of university students. In her study, Cortes (2004) identified the most frequent bundles in that corpus (over two million words) and classified them structurally as well as functionally. She also investigated students' use of these bundles, noting that students rarely use these bundles in their writing. Cortes concluded, "the functions that lexical bundles perform in academic writing make these expressions an important topic that needs to be exploited in teaching of this register" (p. 420). Likewise, Scott and Tribble (2006) compared students' writing to professional writing and found that novice writers used less sophisticated and less varied lexical bundles.

To further determine the main discourse functions of the lexical bundles, Biber et al. (2004) investigated the use of multi-word sequences in essential university registers (classroom teaching and textbooks) and compared the lexical bundles in these registers to those found in their previous research on conversation and academic prose. They analyzed the target bundles and presented a structural taxonomy, in which they differentiated three core groups of bundles

that differed depending on the register: *verb phrase* (VP) fragments, *dependent clause* (DC) fragments, and *noun phrase and prepositional phrase* (NP/PP) fragments (Biber et al., 2004a, p. 381). A detailed description of this taxonomy will be displayed in the following section.

In addition to the structural taxonomy, Biber et al. (2004, pp. 389-396) proposed a preliminary functional taxonomy in which the pragmatic functions of lexical bundles are grouped into three main categories: (1) stance expressions, (2) discourse organizers, and (3) referential expressions. They also added a fourth category, special conversational expressions. Biber et al. (2004) conclude, “lexical bundles-the most frequent sequences of words in a register- are a unique linguistic construct. Lexical bundles are usually not complete grammatical structures nor are they idiomatic, but they function as basic building blocks of discourse” (p. 317).

In investigating gaps in language use, Chen (2010) compared the pragmatic functions and use of lexical bundles in two academic disciplines: electrical engineering introductory textbooks (EEITC) and English for specific purposes (ESP). Two corpora were made up for each discipline (ESPC: 99,774 words and EEITC: 247, 346). A total of 105 four-word sequences were identified in the electrical engineering corpus, while a fewer number of bundles were identified in the ESP corpus. The lexical bundles were classified into their pragmatic functions following Biber et al.’s (2004) original functional taxonomy (stance bundles, discourse organizers, and referential bundles). In comparing the lexical bundles and their pragmatic functions in both EEITC and ESPTC, Chen found that there is a gap in language use between EE introductory textbooks and ESP textbooks. That is, “only one-third of the lexical bundles identified in EE introductory textbooks appear in the ESP textbooks, covering a much narrower scope of functional subcategories” (p. 123). Accordingly, Chen suggests that lexical bundles should be included in

pedagogical materials, “which would not only enrich the description of target language use but also increase novices’ awareness of the patterns of use in the discipline” (p. 123).

Another key study investigating gaps in language use is Wood and Appel’s (2014). They compared the use of “multiword constructions” (MWCs) in popular English for Academic Purposes (EAP) textbooks to those in university First Year Business and Engineering Textbooks Corpus (FYBETC), a corpus they created especially for their study. They examined whether the FYBETC MWCs are present in the reading texts of the EAP textbooks. The study sought to determine whether the EAP textbooks contained pedagogical treatment of MWCs. The results of this study clearly show the absence of FYBETC MWCs as teachable units in EAP textbooks. That is, the MWCs occurrences in EAP textbooks were not pedagogically dealt with at all. Accordingly, Wood and Appel call for more exposure of relevant MWCs in EAP textbooks.

2.2.2 Identification of lexical bundles. Unlike other forms of formulaic sequences that are classified based on their degree of fixedness and compositionality (e.g., idioms and fixed formulas), lexical bundles can be identified using a frequency-based approach (e.g., Altenberg, 1998; Biber et al., 1999; Biber et al., 2004a; Cortes, 2004; Hyland, 2008). Cortes (2004) refers to lexical bundles as “extended collocations, sequences of three or more words that statistically co-occur in a register” (p. 400). Likewise, Wood (2015) notes:

Lexical bundles have typically been defined as combinations of three or more words which are identified in a corpus of natural language by means of corpus analysis software programs, identified using a specific frequency cutoff, and present in a particular range of texts within the corpus (p. 122).

In this definition, three main components seem to be used in identifying lexical bundles, namely: frequency cutoff (i.e., the minimum number of occurrence of a word cluster per million word),

distribution or range (i.e., minimum number of texts in which the bundle appears), and functions (i.e., the meaning and purpose of a word cluster).

Frequency is a key component in identifying lexical bundles. To use frequency in identifying lexical bundles, researchers build up a corpus of texts and scan it, using corpus analysis tools such as WordSmith Tools (Scott, 2007). Then, they create a list of the most frequent multiword sequences using an appropriate frequency cutoff. It is important to note that the frequency cutoff thresholds are often determined somewhat arbitrary and decisions are often based on previous studies (Biber, 2006). Generally, the minimum frequency cutoff is set based on the size and specialization of the corpus, varying from ten occurrences per million words (Biber, 2006; Biber et al., 1999; Simpson-Vlach and Ellis, 2010) to forty (e.g., Biber et al., 2004). The use of frequency as a measure helps determine the formulaicity of the bundle (i.e., the extent to which a bundle is stored and used as a prefabricated chunk). However, as Biber et al. (2004) indicate “frequency is only one measure of the extent to which a multiword sequence is prefabricated... We do not regard frequency data as explanatory. In fact, we would argue the opposite: frequency data identifies patterns that must be explained” (p. 376).

The rate of occurrence of a bundle is mainly associated with a particular range or distribution to avoid individual idiosyncrasy in the process of identifying lexical bundles. The notion of range refers to the spread of a lexical bundle within the corpus (i.e., the number of texts that must include the word sequence before it is considered a lexical bundle). There seems to be a consensus among researchers that three to five texts (e.g., Biber & Barbieri, 2007) or 10% of all texts (e.g., Hyland, 2008) is the minimum range in which a sequence must occur to be considered a lexical bundle. Wood (2015) indicates that the distribution/range criterion “helps to

eliminate the possibility that certain sequences might be used more by a particular author, or when a particular topic is under consideration” (p. 122).

In addition to frequency and distribution criteria, identifying the functional characteristics of each word sequence has been a key in identifying lexical bundles, as these functions help create the meaning within a text.

2.2.3 Structural characteristics of lexical bundles. Unlike other prefabricated chunks or fixed expressions, lexical bundles more likely represent incomplete grammatical structural units and serve as frames that connect other structural elements in texts (spoken or written). In their distinctive study, Biber et al. (2004) found, “only 15 percent of the lexical bundles in conversation can be regarded as complete phrases or clauses, while less than 5 percent of the lexical bundles in academic prose represent complete structural units” (p. 377). Biber et al. further explain how the use of lexical bundles in bridging two structural units in conversation differs from that in academic prose. In their words, “Most of the bundles in conversation bridge two clauses [e.g. *I want to know, well that's what I*], while bundles in academic prose usually bridge two phrases [e.g. *in the case of, the base of the*]” (p. 377). Biber et al. (2004, p. 381) thus propose a structural taxonomy of the lexical bundles in which they identify three major types of bundles (i.e., *verb phrase* [VP] fragments, *dependent clause* [DC] fragments, and *noun phrase and prepositional phrase* [NP/PP] fragments), which differ based on their register. Biber et al.’s (2004, p. 381) structural classification of lexical bundles is highlighted as follow:

1) Lexical bundles that incorporate verb phrase fragments

- a. (Connector+) 1st/2nd person pronoun + VP fragment

E.g., *you don't have to, I'm not going to, well I don't know*

- b. (Connector+) 3rd person pronoun + VP fragment

E.g., *it's going to be, that's one of the, and this is a*

c. Discourse marker + VP fragment

E.g., *I mean you know, you know it was, I mean I don't*

d. Verb phrase (with non-passive verb)

E.g., *is going to be, is one of the, have a lot of, take a look at*

e. Verb phrase with passive verb

E.g., *is based on the, can be used to, shown in figure N*

f. Yes-no question fragments

E.g., *are you going to, do you want to, does that make sense*

g. WH-question fragments

E.g., *what do you think, how many of you, what does that mean*

2) Lexical bundles that incorporate dependent clause fragments

a. 1st/2nd person pronoun + dependent clause fragments

E.g., *I want you to, I don't know if, I don't know why, you might want to*

b. WH-clause fragments

E.g., *what I want to, what's going to happen; when we get to*

c. If-clause fragments

E.g., *if you want to, if you have a, if we look at*

d. (verb/adjective+) to-clause fragment

E.g., *to be able to, to come up with, want to do is*

e. That-clause fragments

E.g., *that there is a, that I want to, that this is a*

3) Lexical bundles that incorporate noun phrase and prepositional phrase fragments

a. (connector+) noun phrase with of-phrase fragment

E.g., *one of the things, the end of the, a little bit of*

b. Noun phrase with other post-modifier fragment

E.g., *a little bit about, those of you who, the way in which*

c. Other noun phrase expressions

E.g., *a little bit more, or something like that, and stuff like that*

d. Prepositional phrase expressions

E.g., *of the things that, at the end of the, at the same time*

e. Comparative expressions

E.g., *as far as the, greater than or equal, as well as the*

2.2.4 Pragmatic functions of lexical bundles. Employing functional analysis to classify lexical bundles dates back to Altenberg (1993, 1998), who pioneered the use of frequency as a main criterion in identifying particular word sequences and classifying them according to their function. Altenberg (1998) noted that recurrent word combinations in spoken language are closely associated with particular situations, where their meaning can only be interpreted in context. He suggested, “Each expression has a distinctive conventionalized discourse function, serving to express the speakers’ view of how an utterance is to be interpreted” (p. 110). Lexical bundles have been also functionally classified by a number of other researchers, such as Cortes (2002) and Biber et al. (2003, 2004). Biber et al. (2004a) proposed a sort of standard functional taxonomy that has been widely used in studies on lexical bundles (e.g., Chen, 2008; Cortes, 2002, 2004; Hyland, 2008a, 2008b; Wood & Appel, 2014). In this taxonomy, three main categories of lexical bundle are identified; with each category having several subcategories that serve more specific functions and meanings. These categories are: (1) stance expressions, (2) discourse organizers, and (3) referential expressions. A description of these categories is discussed in detail in the following section, based on Biber et al. (2004a, pp. 389-396).

2.2.4.1 stance bundles. Stance bundles are groups of sequences that are particularly used to reveal a writer's attitude, perspective, or judgment and to interpret the content of what immediately follows. This group of bundles is divided into two major subcategories: (1) epistemic stance bundles, which represent the status of knowledge (e.g., *I don't know if, I don't think so*); and (2) attitudinal/modality stance bundles, which reveal the writer's/speaker's attitude or feeling towards an upcoming proposition (e.g., *I want you to, I am not going to*). Both of these subcategories can be further categorized as personal or impersonal expressions. Biber et al. (2004) explain, "personal stance bundles are overtly attributed to the speaker/writer...[while] impersonal stance bundles express similar meanings without being attributed directly to the speaker/writer -e.g., it is possible to, can be used to" (p. 389). According to Biber et al., most epistemic bundles that occur in classroom teaching and conversation are personal. These bundles might be used to express uncertainty (e.g., *I don't think, I don't know if*), or to convey a sense of imprecision (e.g., *what do you think, and stuff like that*). Unlike personal epistemic bundles, impersonal epistemic bundles tend to express a degree of certainty rather than uncertainty (e.g., *are more likely to, in the fact that the*).

Attitudinal/modality bundles are also mostly personal and express "speaker attitudes towards the actions or events described in the following proposition" (Biber et al., 2004, p. 390). These bundles can be further divided into four subcategories. (1) Desire bundles express wishes or desires or inquire about another participant's desire (e.g., *if you want to, I would like to, what I want to do*). (2) Obligation/directive bundles can be either personal (e.g., *you look at the, you need to know*) or impersonal (e.g., *it is necessary to*), where most obligation/directive bundles are personal stance expressions and usually contain a second-person pronoun (you) as the subject (e.g., *you might want to*). However, in some cases the first-person pronoun is used along with

verbs of desire to directly reveal the speaker's desire that the recipient carry out some action (e.g., *I want you to*). Next, (3) intention/prediction stance bundles are mostly personal and signify the speaker's/writer's intention to do something (e.g., *what we're going to*). This category may also include impersonal intention/prediction stance bundles, which are usually used to express logical or mathematical process (e.g., *is going to*). Finally, (4) ability bundles represent a small subset category of stance bundles that are used to express ability to accomplish something (e.g., *to be able to, to come up with*)

2.2.4.2 discourse organizing bundles. As their name implies, discourse organizing bundles “are word combinations used to express textual functions which are concerned with the meaning of the sentence as a message in relation to the surrounding discourse” (Cortes, 2004, p. 410). These bundles help to structure and organize the flow of ideas in a text and mark the relationship between parts of the text. This category can be divided into two subcategories: (1) topic introduction/focus, and (2) topic elaboration/clarification. Topic introduction/focus bundles are used to introduce new topics and often “occur with both first and second person pronouns” (Biber et al., 2004, p. 392), as in *I want to talk about* or *if you look at*. Meanwhile, topic elaboration/clarification bundles are used to provide detail and clarification of the topic, as in *on the other hand, it has to do with, or as well as*.

2.2.4.3 referential bundles. Referential bundles signify a large number of lexical bundles that “generally identify an entity or single out some particular attribute of an entity as especially important” (Biber et al., 2004, p. 393). These bundles fall into four major subcategories: (1) identification/focus, (2) imprecision indicators, (3) specification of attributes, and (4) time/place/text reference. Looking at these more closely, identification/focus bundles are usually used to identify what immediately follows the bundles (i.e., the noun phrase after the bundle) and

are often used in classroom teaching, as in: “for *those of you who* came late I have the uh, the quiz” (classroom teaching, Biber et al., 2004, p. 395). Identification/focus bundles might also hold a discourse organizing function, particularly when they are used to introduce a topic or emphasize or summarize the main point that precede these bundles, as seen in the following examples:

One of the things they stress in parenting is to be consistent and particularly with parents um some parents are inconsistent between siblings. Uh fathers are notorious for letting their little darling girls get away with what they swat the boys about... (Biber et al., 2004a, p. 394)

On the other hand, as the name indicates, imprecision bundles (the second subcategory of referential bundles) can “either indicate that a specified reference is not necessarily exact, or to indicate that there are additional references of the same type that could be provided” (Biber et al., 2004, p. 394), such as *or something like that* and *and things like that*.

The third subcategory of referential bundles, specification of attributes bundles, identifies certain qualities of the head noun that immediately follows the bundle. The bundles of this subcategory can be divided according to their functions: (1) quantity specification bundles that deal with quantity or amounts (e.g., *have a lot of, a little bit of, a little bit about*); (2) tangible framing attributes bundles that are used to describe the size or form of the noun (e.g., *the size of the, in the form of*); and (3) intangible framing attributes bundles that are used to identify abstract qualities (e.g., *the way in which, the nature of the, in terms of*).

The last subcategory of referential bundles includes bundles that refer to place, time, and text (Biber et al., 2004, p. 396). This subcategory can be also further divided into: (1) place reference bundles (e.g., *in the United States, of the United States*); (2) time reference bundles (e.g., *at the same time, at the time of*); (3) text deixis (e.g., *as shown in figure, shown in figure*

N); and (4) multi-functional reference bundles (e.g., *the end of the, at the end of, in the middle of*). **Table 1**, below, summarizes Biber's et al. (2004a) functional taxonomy of lexical bundles.

Table 1 The Functional Taxonomy of Lexical Bundles (Biber et al., 2004a)

Category	Examples
1. Stance Expressions	
A. Epistemic stance	
Personal	<i>I think it was, I don't know if</i>
Impersonal	<i>are more likely to, the fact that the</i>
B. Attitudinal/Modality stance	
B1) Desire	<i>if you want to, I would like to, what I want to do</i>
B2) Obligation/Directive	
Personal	<i>you look at the, you need to know, you have to do</i>
Impersonal	<i>it is necessary to</i>
B3) Intention/Prediction	
Personal	<i>what we are going to</i>
Impersonal	<i>is going to be</i>
B4) Ability	
Personal	<i>to be able to</i>
Impersonal	<i>it is possible to</i>
2. Discourse Organizers	
A. Topic introduction/focus	<i>With regard to the, in this chapter we</i>
B. Topic elaboration/clarification	<i>on the other hand, as well as the</i>
3. Referential Expressions	
A. Identification/Focus	<i>is one of the, one of the most</i>
B. Imprecision	<i>and things like that</i>

C. Specification of attributes

C1) Quantity specification	<i>the majority of the, a lot of people</i>
C2) Tangible framing attributes	<i>in the form of, in the size of</i>
C3) Intangible framing attributes	<i>in the case of, on the bases of</i>

D. Time/Place/Text reference

D1) Place reference	<i>in the United States</i>
D2) Time reference	<i>at the same time</i>
D3) Text deixis	<i>as shown in Figure N</i>
D4) Multi-functional reference	<i>at the end of</i>
-

Another remarkable attempt to classify lexical bundles according to their functions was the classification developed by Hyland (2008, 2008b) based on Halliday's (1993) three major metafunctions of language: ideational, textual, and interpersonal. In his study, Hyland (2008) examined the forms, structures, and functions of clusters (four-word bundles) in a large corpus (3.5 million words) of research articles, doctoral dissertations, and Master's theses in four disciplines. He classified lexical bundles into three major groups: (1) research-oriented bundles (reflect ideational functions); (2) text-oriented bundles (reflect textual functions); and (3) participant-oriented (reflect interpersonal functions). Each group was further classified into several sub-categories that reflect various issues in research writing. This functional taxonomy is introduced in Hyland (2008, p. 13-14) as follows:

(1) **Research-oriented clusters** that "help writers to structure their activities and experiences of real world" (p. 13) include:

- Location, i.e., time/place (e.g., *at the beginning of, at the same time, in the present study*);

- Procedure (e.g., *the use of the, the role of the, the purpose of the, the operation of the*);
- Qualification (e.g., *the magnitude of the, a wide range of, one of the most*) ;
- Description (e.g., *the structure of the, the size of the, the surface of the*); and
- Topic (e.g., *in the Hong Kong, the currency board system*).

(2) **Text-oriented clusters** “concerned with the organization of text and its meaning as a message or argument” (p. 13) include:

- Transition signals (e.g., *on the other hand, in addition to the, in contrast to the*);
- Resultative signals (e.g., *as a result of, it was found that, these results suggest that*);
- Structuring signals (e.g., *in the present study, in the next section, as shown in figure*); and
- Framing signals (e.g., *in the case of, with respect to the, on the basis of, in the presence of, with the exception of*).

(3) **Participant-oriented clusters** “focused on the writer or reader of the text” (p. 14) include:

- Stance features (e.g., *are likely to be, may be due to, it is possible that*); and
- Engagement features (e.g., *it should be noted that, as can be seen*)

2.3 Corpus Linguistics

2.3.1 Introduction to corpus linguistics. Computer-based corpora have been around in some form or the other since the 1960s (Kennedy, 1998), but only over the past three decades with the emergence of computer-based software, comes the development of more qualitative and

quantitative corpus-based studies (Conrad, 1996; O’Keeffe, McCarthy & Carter, 2007). The first mention of a computer-based corpus dates back to Francis and Kucera’s classic work developed in the 1960s: The Brown University Standard Corpus of Present-Day American English or simply the Brown Corpus (a one-million-word corpus). The corpus was compiled from randomly published materials in the United States and covers a wide range of genres. Since then, further development of corpus linguistics, defined as “the study of language on the basis of text corpora” (Aijmer & Altenberg, 1991, p.1), has helped overcome the challenges of manually compiling hard copies of authentic samples of writing or speech in order to examine them (Sinclair, 1991; Wray, 2002).

Corpus linguistics is considered a reliable source of authentic information about language. It provides a rapid method of processing and searching language data, as well as offer direct access to and the ability to manage huge amounts of data (Kennedy, 1998). Accordingly, the use of corpus linguistics has made it feasible for researchers to access and explore hundreds or thousands of texts encompassing hundreds of millions of words (Tribble & Scott, 2006) in order to examine the structure and use of language.

2.3.2 Types of corpora. The term “corpus” has been variously defined. For example, Francis (1982) defines a 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). Likewise, Kennedy (1998) refers to a corpus as “a body of written text or transcribed speech which can serve as a basis for linguistic research” (p. 1). O’Keeffe et al. (2007), meanwhile, define corpus as “a principled collection of electronic texts usually stored on a computer, available for qualitative and quantitative analysis” (p. 2).

Regardless of terminology, the use of corpora has been an extremely powerful way to investigate different kinds of multi-word units, such as lexical bundles, clusters, collocations, phrasal verbs, speech formulae, etc. Corpora are established to serve different needs across registers (Chafe, Du Bois, & Thompson, 2014), and can be of many different types, including specialized, general, parallel, comparable, pedagogic, historical or diachronic, and learner corpora (Hunston 2002, pp. 14-16). Below is a brief description of these seven types.

2.3.2.1 specialized corpora. Specialized corpora, such as the one intended for this current study, are “used to investigate a particular type of language” (Hunston 2002, p. 14). This type of corpus is usually relatively small and not intended to be representative of the language as a whole; rather, it “aims to be representative of a given type of text” (Hunston, 2002, p. 14). It is often collected by researchers independently and attempts to represent target language use in specific areas (O’Keefe et al., 2007). Specialized corpora vary in their size according to the breadth of the language use domain that these corpora represent, such as the Michigan Corpus of Academic Spoken English (MICASE) that is approximately 1.8 million words or the TOEFL 2000 Speaking and Writing Academic Language (T2KSWAL) corpus of nearly 2.8 million words. These two corpora have been widely used in lexical, grammatical, and lexico-grammatical research in specific registers. For example, in academic speech, Simpson (2004) explored the MICASE and identified numerous formulaic expressions, most of which functioned as discourse organizers (e.g., *and in fact*) and interactional markers (e.g., *and you can see*).

2.3.2.2 general corpora. Hunston (2002) describes a general corpus as “A corpus of texts of many types. It may include written or spoken language, or both, and may include texts produced in one country or many” (p. 14). Unlike specialized corpora, general corpora are intended to represent the state of language as a whole; thus, these corpora include a wide range

of texts from many different language situations. The London-Lund corpus, the Bank of English, the British National Corpus (BNC), the Corpus of Contemporary American English (COCA), and the Longman Spoken and Written English corpus (LSWE) are all examples of general corpora. These types of corpora are made up of hundreds of millions of words from a variety of sources such as news, conversations, books, movies, journal articles, and more. For example, the LSWE is a 40-million-word corpus developed by Biber et al. (1999) to identify grammatical variations across four primary registers: conversation, fiction, news, and academic prose. Meanwhile, the COCA contains 560 million words of texts from different levels of formality: spoken language, fiction, magazines, newspapers, and academic journals. A general corpus is sometimes referred to as reference corpus as “it is often used as a baseline in comparison with more specialized corpora” (Hunston, 2002, p. 15).

2.3.2.3 comparable corpora. Comparable corpora are described by Hunston (2000, p. 15) as “two (or more) corpora in different languages (e.g., English and Spanish) or in different varieties of language (e.g., Indian English and Canadian English)” that have been designed in a similar way so that language features can be compared and contrasted across the languages/varieties. The International Corpus of English (ICE) is an example of comparable corpora made up of different varieties of spoken and written English (e.g., dialogues, monologues, printed, and non-printed texts).

2.3.2.4 parallel corpora. Parallel corpora are closely related to comparable corpora. These corpora are defined by Hunston (2002) as:

Two (or more) corpora in different languages, each containing texts that have been translated from one language into the other (e.g., a novel in English that has been translated into Spanish, and one in Spanish that has been translated into

English) or texts that have been produced simultaneously in two or more languages (e.g., European Union regulations, which are published in all the official languages of the EU (p. 15).

Parallel corpora can be either mono-directional (i.e., the translation is from one language to another) or bi-directional (i.e., the translation is reversible from one language to the other), as illustrated in the above definition. These corpora seem to play an effective role in investigating the differences between languages. Examples of the parallel corpora include the Intersect Parallel Corpus (English-French), the International Telecommunications Union Corpus (English-Spanish), The Multilingual Parallel Corpus (Danish, English, French, German, Greek, Italian, Finish, Portuguese, Spanish, Swedish texts), and the English-German Translation Corpus.

2.3.2.5 learner corpora. A learner corpus is a collection of authentic texts (e.g., essays) produced by language learners (Hunston, 2002, p. 15). Learner corpora can be used either to highlight discrepancies in language use among L2 learners or to identify differences between the language of L2 learners and that of native speakers. Similar to a general corpus, this kind of corpus plays an invaluable role in providing an overall view of typical (frequently occurring) mistakes in learner language of certain background (e.g., Chinese learners of English) or proficiency level (e.g., beginner learners vs. advanced ones). The International Corpus of Learner English (ICLE) is an example of the learner corpus.

2.3.2.6 pedagogic corpora. A pedagogic corpus, as defined by Hunston (2002) refers to all the language that learners encounter during their studies (e.g., textbook materials, readings, tapes, etc.). This kind of corpus is useful in raising learners' awareness by collecting "all instances of a word or phrase... [learners] have come across in different contexts" Hunston

(2002, p. 15). Building on this description, the corpus introduced in this study, FYETC, is also a pedagogic corpus since it deals with the language that learners encounter in textbooks.

2.3.2.7 historical or diachronic corpora. Finally, historical or diachronic corpora refer to a collection of texts assembled from different periods of time, and which “is used to trace the development of aspects of language over time” (Hunston, 2002, p. 15). This type of corpus helps researchers explore changes in language in long diachrony such as the historical corpus of English: Helsinki Corpus that includes periodically organized texts from Old, Middle, and Early Modern English (i.e., from 700 to 1700) and made up of 1.5 million words.

All in all, corpus studies have proven beneficial in almost all the fields that relate to applied linguistics and discourse studies like learning and teaching languages, translation, forensic linguistics, and more. Many remarkable achievements in different fields were based on the use of corpora due to their ability to provide us with data and knowledge, which is not attainable through other means such as human observation and native speaker intuition. For example, corpora are used in electronic translations that many of us find useful for online communication. They are also made use of in developing teaching materials for special courses. The applications of using corpora are countless and valuable.

To summarize, this chapter includes a review of some of the literature written about three important areas of research. First, it reviews the theoretical background of formulaic language and its nature and characteristics, significance, how it is mentally processed, and the methods of identifying formulaic sequences. Next, the chapter sheds light on lexical bundles, as the main focus of this study. It reviews the research written about lexical bundles, how to identify them, and how to classify them structurally and functionally. Finally, the chapter briefly highlights the

importance of corpus linguistics, explaining different types of corpora and how they can be useful.

3. Methodology

Prior to discussing how this study attempts to achieve its goals and before reporting the methodology steps and procedures in detail, I hereby provide a reminder of the research questions of this study. The questions are:

1. What are the most frequent lexical bundles in first year Economics textbooks that are used in Canadian universities?
2. What functions do these lexical bundles serve?

To answer these questions, a corpus-based approach was adopted to pinpoint the target lexical bundles following the functional taxonomy suggested by Biber et al. (2004a) as a theoretical and analytical framework. The following sections explain in detail the procedures I followed regarding the choice of every aspect of this study, namely:

- 3.1. the choice of the setting of this study,
- 3.2. the choice of specific textbooks,
- 3.3. the choice of WordSmith Tools 6.0 and building the corpus, and
- 3.4. the identification and classification criteria of lexical bundles.

3.1 Choice of the Setting

As language use and lexical bundles preferences can differ from region to region, I thought it would be better to focus on the Canadian universities as the setting of my study, especially the universities in Ontario. The reason this study focuses on Ontario universities is that just as Ontario is home to almost 40% of the population of Canada, it also attracts around 40% of the

total number of international students who study in Canada, according to *Citizenship and Immigration Canada's Facts and Figures 2011* (2011) [See **Figure 1** below].

Distribution of International Students in Canada by Province and Territory, 2004 to 2010							
	2004	2005	2006	2007	2008	2009	2010
Newfoundland and Labrador	0.6%	0.6%	0.7%	0.8%	0.9%	0.8%	0.8%
Prince Edward Island	0.2%	0.2%	0.2%	0.3%	0.3%	0.3%	0.3%
Nova Scotia	3.3%	3.3%	3.2%	3.1%	3.3%	3.4%	3.6%
New Brunswick	1.9%	1.9%	1.9%	1.8%	1.8%	1.7%	1.6%
Quebec	15.5%	15.1%	15.2%	15.4%	15.7%	15.6%	14.8%
Ontario	37.8%	37.6%	37.5%	37.2%	36.9%	37.2%	39.1%
Manitoba	2.9%	3.0%	3.0%	3.0%	2.7%	2.7%	2.6%
Saskatchewan	2.2%	2.2%	2.1%	2.0%	2.1%	2.0%	2.0%
Alberta	7.3%	7.4%	7.6%	7.9%	8.1%	7.9%	7.5%
British Columbia	28.3%	28.6%	28.5%	28.5%	28.2%	28.3%	27.7%
Yukon, Northwest Territories and Nunavut	0.1%	0.1%	0.0%	0.1%	0.0%	0.0%	0.0%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Source: Citizenship and Immigration Canada

Figure 1 Distribution of inte'l students in Canada by province and territory, 2004-2010

Therefore, focusing on Ontario universities is an attempt to address the needs of as many students as possible; especially the ones in first year who, hypothetically, need help the most. It is, also, worth-mentioning that the number of international students in Canada is on a constant rise throughout the past two decades. For example, “in December 2015[,] there were 353,000 international students with a valid study permit in Canada, up from 84,000 in December 1995”, according to *Statistics Canada* (2017). Even more, some researchers and writers suggest that this number is expected to increase as a result of some changes in the political environment in the United States after the latest USA presidential elections. These changes could probably encourage more international students to choose Canada as their academic destination, instead of the United States. For example, Kathleen Harris, a senior writer in the CBC’s Parliament Hill bureau, states that, “the steady upswing in foreign [university] applicants began several years

ago, then started to spike after the U.S. presidential election in 2016” (“Foreign Students Flock to Canada”, 2017). This, of course, is in addition to the numbers of refugees and immigrants who might need EAP classes that are tailored to match their needs and skills to the needs of the local Canadian labour market.

More importantly, almost half of the foreign or international students in Canadian universities are studying business, management, and public administration (27%) and architecture, engineering, etc. (19%), according to *Statistics Canada* (2017) [See **Figure 2** below].

Source: Statistics Canada, Post-secondary Student Information System, 2004-2005 to 2013-2014.

Figure 2 Top 5 fields of study for inte'l university students, Canada, 2004-2005 to 2013-2014

However, despite all these facts and statistics about business being the most popular field of study among international students, and despite the importance of economics as a core subject of business; no research that we know of has focused on Economics textbooks in Canadian universities and how formulaic language studies and lexical bundles in particular can play a

crucial role in promoting and speeding up the learning process for international students whose first language is not English. So, this study focuses on economics in particular, as an attempt to fill this gap in the research body on formulaic language studies and lexical bundles

3.2 Choice of the Textbooks

As for choosing the textbooks for this study, there are some criteria that were carefully employed in order to ensure a good quality of results. In brief, the first criterion was that the textbooks had to be used in first year (FY) and/or introductory courses. By this, I believe the product of this study would be better aimed to help the target group of students and/or any researcher or curriculum developer who may choose to use it in their work. The second criterion was to ensure representativeness and commonality. That is, the chosen textbooks ought to be commonly used and popular among many universities and courses of Economics. For example, *Principles of Microeconomics* (Mankiw, 2014) is used as a required textbook in both (ECON 1000 A) course at Carleton University and (ECON 100 Y-L0101) course at the University of Toronto. Also, *Principles of Macroeconomics* (Mankiw & Kneebone, 2013) is used in (ECON 1000 A) course at Carleton University and (ECO 1102) course at the University of Ottawa.

Applying these criteria of textbook selection narrowed down the choices to the following textbooks:

1. Principles of Microeconomics 5th Ed. (Mankiw, 2014),
2. Principles of Macroeconomics 6th Ed. (Mankiw & Kneebone, 2013),
3. Economics is Everywhere (Hamermesh, 2014),
4. Macroeconomics 10th Ed. (Parkin, 2011),
5. Microeconomics 10th Ed. (Parkin, 2011),
6. Foundations of Microeconomics 6th Ed. (Bade & Parkin, 2012),

7. Principles of Macroeconomics (Bouman, 2001), and
8. Essential Foundations of Economics 6th Ed. (Bade & Parkin, 2013).

These textbooks are required for courses in a number of Ontario universities, i.e. The University of Toronto, The University of Waterloo, The University of Western Ontario, McMaster University, and Carleton University. They are also commonly used in other universities across Canada and other parts of the world. These textbooks seem to be highly renowned and foundational in Economics.

Combined together, the eight textbooks made up a corpus size of exactly 1,346,658 words in total, which can be regarded as a reasonable corpus size that is specialized in one branch of business, i.e. Economics. However, in spite of the reasonable corpus size; the limited number of texts had implications on the minimum range adopted in this study, as shown in the following section.

After following this corpus-based approach, I chose an efficient software that would help identify word sequences prior to further analysis. So, I chose *WordSmith Tools 6.0* (Scott, 2007).

3.3 WordSmith Tools 6.0 and Building the Corpus

WordSmith Tools 6.0 (Scott, 2007) is an easy-to-use and user-friendly software that is used in this study for finding the recurrent word patterns and sequences in the created corpus. *WordSmith Tools 6.0* comes with a very detailed user's manual that ushers and enables its users to find not only frequently used word stings; but also, it enables them to find out their repetition frequencies, distribution range, and concordances.

As for the steps I went through when using *WordSmith Tools 6.0* as a text analysis software of choice, I first needed to convert all the PDF textbooks to text files with the extension (.txt), since this is the format that is allows *WordSmith Tools 6.0* to process texts. This step

included eliminating textbooks prefaces, reference pages, instructional language -if any-, etc. Second, I created a *WordList Index*, which records the positions of all the words that exist in all the text files [See **Figure 3** below].

The screenshot shows a software window titled "First Year Economics Textbook Corpus (FYETC) index_file". The window has a menu bar with "File", "Edit", "View", "Compute", "Settings", "Windows", and "Help". Below the menu bar is a table with the following columns: "N", "Word", "Freq.", "%", "Texts", and "%". The table contains 15 rows of data, with the first row highlighted in yellow.

N	Word	Freq.	%	Texts	%
1	THE	92,082	6.84	8	100.00
2	#	86,378	6.41	8	100.00
3	OF	44,440	3.30	8	100.00
4	AND	37,341	2.77	8	100.00
5	A	35,188	2.61	8	100.00
6	TO	26,986	2.00	8	100.00
7	IN	26,968	2.00	8	100.00
8	IS	22,011	1.63	8	100.00
9	THAT	13,775	1.02	8	100.00
10	PRICE	12,365	0.92	8	100.00
11	FOR	11,688	0.87	8	100.00
12	S	8,435	0.63	8	100.00
13	ARE	8,118	0.60	8	100.00
14	ON	7,370	0.55	8	100.00
15	DEMAND	6,900	0.51	8	100.00

Figure 3 A *WordList Index* file is created as the first step of building the FYETC

The creation of a wordlist index speeds up the software ability to access these words whenever needed, e.g. in the case of looking for concordances and/or word clusters. After creating a wordlist index, the third step is to open this index and choose *Compute* and then the *Clusters* option from the drop-down menu [See **Figure 4**].

Figure 4 Computing the clusters

At this point, the software allows its users to set their preferred cluster size and minimum frequency cutoff. For example, some researchers choose to set their cluster size to 4, which means that they will get a list of all the clusters that consist of four words. The minimum frequency indicates the number of occurrences of a certain cluster or a word string in the text. So, if a researcher was looking for all the clusters that consist of four words and appeared at least five times in the text, he/she needs to set the cluster size to 4 and the minimum frequency to 5. There are recommended minimum frequencies, e.g. in Biber (2004a) the minimum frequency cutoff was 40/million words. However, the minimum frequency depends on the size of the corpus. That is, the bigger the corpus is, the higher the frequency can be; and a lower minimum frequency can be used with smaller corpora. This is in line with Wood (2015) as he states, “This frequency-based method is most appropriate for large corpora of hundreds of thousands of words, if not millions, taken from specific registers of language and/or academic disciplines. It has many limitations for use with small data sets, as the standard minimum cutoffs for frequency established in the field may not be met in such case” (p. 21). In this study, since the corpus size is 1,346,658 words, the minimum frequency cutoff is set to 40 times per million words. Also, the cluster size is set to 3-8-word clusters [See *Figure 5* below]. The reason behind choosing this range of cluster size is that it is by definition that the minimum size of a lexical bundle is three; and I also wanted to include as many clusters as possible.

Figure 5 Cluster choices

This allowed me to find all the clusters (the term *clusters* is used in *WordSmith Tools* software to refer to recurrent word sequences) that consist of three to eight words and have a high frequency of at least 40 times per million words. The list was then saved as an Excel spreadsheet file for easier access and editing, as *WordSmith Tools 6.0* allows this option [See **Figure 6** below].

Figure 6 Saving results as Excel spreadsheet for easier access and editing

As expected, the results obtained from *WordSmith Tools 6.0* or any similar software are only quantitative and most likely contain unneeded data. In other words, this led to a huge list of recurrent clusters that contained a lot of unwanted items and it then needed to be refined in order to determine which of them was really a lexical bundle and which was not. For example, in **Figure 7** below, the third most frequent word cluster is ‘The United States’. This word cluster appeared in the list depending only on the quantitative criterion of frequency; but it does not qualify as a lexical bundle since it does not match any of the functions listed in the theoretical framework that is followed in this study i.e. the lexical bundles functional taxonomy of Biber et al (2004a).

	A	B	C	E
1	N	Word	Freq.	Texts
2	43737	THE PRICE OF	2119	7
3	44368	THE QUANTITY OF	1560	8
4	45749	THE UNITED STATES	1353	8
5	16857	GOODS AND SERVICES	1104	7
6	40703	THE DEMAND FOR	1098	7
7	31956	PRICE OF A	832	7
8	40624	THE DEMAND CURVE	777	7
9	44269	THE QUANTITY DEMANDED	731	7
10	43738	THE PRICE OF A	703	7

Figure 7 Raw results before expert/human judgment contain irrelevant data

So, the essential point is that these raw results could not be taken for granted and that they needed to be examined and selected manually by the researcher. This entails that further qualitative analysis and human judgment are indispensable in order to refine and finalize a list of lexical bundles. To this end, I used my judgment, as a proficient English language speaker, to manually refine and prepare a list of the multi-word constructions that I believe to be lexical bundles. This list was, then, shown to three judges who are proficient speakers of English, study and/or have studied applied linguistics, familiar with formulaic language and lexical bundles, and hold a master's degree in the field or a higher degree. The valuable comments I received from the judges are discussed in detail in the Discussion section of this paper. After that, the list was discussed with my M.A. thesis supervisor, Dr. David Wood, who approved the LBs list.

In the list, the LBs were organized according to their functions in a descending order from the ones with the highest frequency to the least frequent ones. A class code was given to each bundle for easier processing depending on the function(s) it serves. For example, the class code of a lexical bundle such as '*the fact that*' is (1A). The number (1) refers to the first type of

lexical bundles in Biber et al. (2004a) functional taxonomy i.e. stance expressions; and the letter (A) refers to the epistemic stance. Also, the class code of a lexical bundle such as ‘*the rest of*’ is (3C1), which means that it belongs to referential expressions (3), specification of attributes (C), and quantity specification (1). The classification of those lexical bundles had to be done not only through human judgment, but also according to many criteria that are discussed in detail in the following section.

3.4 The Identification Criteria of Lexical Bundles

As mentioned above, the results obtained from any text analysis software need to be refined manually and through using the educated judgment of individuals who are specialized in formulaic language and lexical bundles, or at least have an adequate knowledge of them. The criteria according to which the lexical bundles were chosen to appear in the FYETC LBs list are in regard to: 1) the frequency, 2) the range, and 3) the function(s). These three criteria are briefly explained below.

1) The frequency criterion, as far as we know, is the first and foremost criterion that the vast majority of researchers seem to have a consensus upon; since Altenberg (1993, 1998) first highlighted the importance of frequency as a tool to decide on the formulaicity of *recurrent word combination*. In this study, frequency also was the first criterion to be employed in order to decide whether a word cluster can qualify to be a bundle. The minimum frequency was set to 40 occurrences per million words, as mentioned earlier in this paper, to be a gatekeeper that allows any word cluster into the list.

2) The range is the second criterion that helps bolster the frequency method of identifying lexical bundles. As Biber et al. (2004a) state, “a sequence must be used in at least five different texts to be counted as a lexical bundle, to guard against idiosyncratic uses by individual speakers

or authors” (p. 376). That is, many speakers/ authors tend to repeatedly use specific phrases in their speech/ writing. In order to avoid the over-representation of those sequences, the range criterion is used to guarantee that those sequences are widely common and used by more than one person. However, since the number of texts used to create the corpus in this study is limited to eight texts; following the guideline recommended by Biber et al. (2004a), i.e. five texts, would have been impractical. Instead, it was more practical and reasonable to follow the minimum range suggested by Hyland (2008), which is 10% of the total number of texts.

In any case, there was only one particular bundle ‘*in the foreign exchange market*’ that had this minimum range. This bundle has three important qualities. First, it is one of the rare lexical bundles that can serve three functions. That is, it can be a topic introducer (2A); an intangible framing bundle (3C3); and/or if we may consider virtual places, a place reference (3D1). The second quality of this particular bundle is that it is a clear case and example of how lexical bundles are generally discipline-bound. The third quality is that it has a high frequency of 81, which could not be overlooked since it is the third highest frequency in its main class i.e. place reference (3D1) after ‘*in the market*’ and ‘*in the world*’.

3) The function was the third criterion that allowed for the qualification of any word cluster found in the FYETC to be considered a lexical bundle. Only the clusters that matched the functions listed in Biber et al. (2004a) functional taxonomy, i.e. stance expressions, discourse organizers, and/or referential expressions, were considered as lexical bundles and included in the FYETC LBs final list. It is worth mentioning here that many of the lexical bundles were found to serve more than one function. For example, the bundle ‘*in this chapter*’ was found to serve not only a function of topic introduction/ focus (2A), but it is also a text deixis bundle (3D3). This phenomenon of serving more than one function is very common and normal, and was referred to

by many other scholars and researchers. For instance, Biber et al. (2004a) state that, “in some cases, a single bundle has multiple functions even in a single occurrence. For example, bundles like *take a look at* and *let’s have a look* function as both directives and topic introducers” (p. 383). Thus, in this study, any lexical bundle that has more than one function was listed under the function that it is more likely to serve and a comment was written next to it identifying the other functions that it could serve. An asterisk next to its class code also highlights this.

Applying the aforementioned identification criteria resulted in a list of 165 lexical bundles that matched the functions suggested in Biber et al. (2004a) functional taxonomy. The final FYETC LBs list is presented and discussed in the following chapter of the results and discussion.

4. Results and Discussion

In this chapter, I first report the results of this study after following all the steps mentioned in the Methodology Chapter above. The results are shown in the form of tables. The first column shows the lexical bundles that were found in FYETC. The second and third columns show the frequency and the range, respectively. Then, the fourth column shows the function(s) of the lexical bundles. The functions are referred to in the form of a class code. For example, the Stance Expressions>Attitudinal/ Modality Stance>Ability bundles are referred to by the class code (1B4). Also, the Referential Expressions>Specification of Attributes>Quantity Specification bundles are referred to by the class code (3C1). The fifth column contains some comments that are worth noting regarding the other functions a lexical bundle can serve, irregular frequencies and ranges, etc. As mentioned earlier, the bundles that can serve more than one function were marked by an asterisk next to their class codes.

This chapter is divided into two main sections: 1) The Results, where I only report the results of this study in the form of a table after applying the identification criteria and consulting the expert judges and my thesis supervisor. The next section, 2) The Discussion, is where I discuss the results of this study in detail and highlight the significant LBs.

4.1 Results

In this section, a list of the lexical bundles that were found in FYETC is presented in the form of tables as follows:

Table 2 Stance Bundles in FYETC LBs List

1. Stance LBs					
A. Epistemic Stance			1A		
A1. Personal			1A1		
N	Lexical Bundle	Freq	Tex	Class Code	Comments
1	do you think	117	6	1A1	
A2. Impersonal			1A2		
2	more likely to	71	8	1A2	
3	the fact that	54	6	1A2	
4	it is possible	49	7	1A2	
B. Attitudinal/Modality			1B		
B1. Desire			1B1		
5	are willing to	319	6	1B1	
6	we need to	84	6	1B1	
7	have an incentive to	56	6	1B1	
B2. Obligation/Directive			1B2		
8	suppose that the	136	6	1B2	
9	look at the	115	7	1B2	
B3. Intention/Prediction			1B3		
10	it would be	67	7	1B3	Prediction and/or Directive (1B3/1B2)
11	is expected to	64	7	1B3*	
12	now going to	52	4	1B3	
B4. Ability			1B4		
13	be able to	181	7	1B4	Ability and /or Identification/focus (1B4/3A)
14	you can see	162	6	1B4*	
15	be used to	57	7	1B4	

Table 3 Discourse Organizing Bundles in FYETC LBs List

2. Discourse Organizing LBs					
A. Topic Introduction/Focus				2A	
N	Lexical Bundle	Freq	Tex	Class Code	Comments
1	to work problems	397	4	2A	Topic Introduction and/or Identification Focus (2A/3A)
2	there is a	333	7	2A*	
3	in part A	225	5	2A*	Topic Introduction and/or Text Deixis (2A/3D3)
4	in part B	221	5	2A*	Topic Introduction and/or Text Deixis (2A/3D3)
5	in this chapter	152	6	2A*	Topic Introduction and/or Text Deixis (2A/3D3)
6	when the price of a	84	6	2A*	Topic Introduction and/or Time Reference (2A/ 3D2)
7	a closer look at markets	72	2	2A	Low range+ Discipline specific
8	to illustrate the	71	6	2A*	Topic Introduction and/or Topic clarification (2A/2B)
9	there are two	67	7	2A*	Topic Introduction and/or Identification Focus (2A/3A)
10	if the price is	66	7	2A	<i>Suggested:</i> Conditional
11	to calculate the	64	6	2A	
12	according to the	63	7	2A	
13	to show the	61	6	2A	
14	by looking at	56	5	2A	
15	to increase the	54	7	2A	
16	in the first	54	6	2A*	Topic Introduction and/or Multifunctional (2A/3D4). <i>Suggested:</i> Sequencing
17	there are many	51	7	2A*	Topic Introduction, Identification/Focus, and/or Quantity (2A/3A/3C1)
B. Topic Elaboration/Clarification				2B	
18	the relationship between	337	7	2B	<i>Suggested:</i> Cause-effect/Dependency Relationship
19	what happens to	165	7	2B	
20	depends on the	160	7	2B	<i>Suggested:</i> Cause-effect/Dependency Relationship
21	in which the	153	7	2B	
22	an example of	137	7	2B	
23	as well as	100	8	2B	
24	at which the	97	7	2B	
25	is that the	93	7	2B	

26	in other words	91	7	2B	
27	to see how	89	5	2B	
28	that results from	87	6	2B	<i>Suggested: Cause-effect/Dependency Relationship</i>
29	that it is	86	7	2B	
30	such as the	81	7	2B	
31	which is the	78	7	2B	
32	on the demand	76	6	2B	
33	to do so	73	7	2B	
34	because of the	73	7	2B	<i>Suggested: Cause-effect/Dependency Relationship</i>
35	to make a	73	6	2B	
36	based on the	72	7	2B	<i>Suggested: Cause-effect/Dependency Relationship</i>
37	is called a	65	6	2B	
38	in order to	60	6	2B	
39	other than the	60	5	2B	
40	in response to	58	7	2B	<i>Suggested: Cause-effect/Dependency Relationship</i>
41	results from a	52	5	2B	<i>Suggested: Cause-effect/Dependency Relationship</i>
42	relative to the	50	6	2B	
43	that there are	49	7	2B	
44	is based on	48	7	2B	<i>Suggested: Cause-effect/Dependency Relationship</i>
45	that raises from	48	5	2B	<i>Suggested: Cause-effect/Dependency Relationship</i>

Table 4 Referential Bundles in FYETC LBs List

3. Referential LBs					
A. Identification/Focus			3A		
N	Lexical Bundle	Freq	Tex	Class Code	Comments
1	there is a	333	7	3A*	Topic Introduction and/or Identification/Focus (2A/3A)
2	one of the	215	7	3A	
3	some of the	180	6	3A	
4	part of the	173	7	3A	
5	of the market	143	7	3A	
6	of the following	112	6	3A	
7	each of the	89	7	3A	
8	which of the	84	5	3A	
9	a market in which	72	5	3A	
10	there are two	67	7	3A*	Topic Introduction and/or Identification/Focus (2A/3A)
11	a number of	65	7	3A	
12	this type of	56	7	3A	
13	there are many	51	7	3A*	Topic Introduction and/or Quantity (2A/3C1)
B. Imprecision			3B		
14	and so on	70	6	3B	
C. Specification of Attributes			3C		
C1. Quantity Specification			3C1		
15	an increase in	602	7	3C1	<i>Suggested: Value>Change</i>
16	a change in	490	7	3C1	<i>Suggested: Value>Change</i>
17	the change in	418	7	3C1	<i>Suggested: Value>Change</i>
18	is less than	265	7	3C1	<i>Suggested: Value>Change</i>
19	the percentage change in	260	6	3C1	<i>Suggested: Value>Change</i>
20	the increase in	245	7	3C1	<i>Suggested: Value>Change</i>
21	a decrease in	239	6	3C1	<i>Suggested: Value>Change</i>
22	the value of the	237	8	3C1	
23	the rest of	221	7	3C1	
24	the sum of	219	7	3C1	
25	a change in the	200	6	3C1	<i>Suggested: Value>Change</i>
26	percent of the	183	6	3C1	
27	remain the same	180	6	3C1	
28	rise in the	176	7	3C1	<i>Suggested: Value>Change</i>
29	the rest of the world	167	5	3C1	

30	in the quantity of	151	7	3C1*	Quantity and/or Intangible framing(3C1/3C3)
31	decrease in the	148	6	3C1	<i>Suggested: Value>Change</i>
32	A rise in	146	7	3C1	<i>Suggested: Value>Change</i>
33	equal to the	139	6	3C1	<i>Suggested: Value>Change</i>
34	the change in the	134	7	3C1	<i>Suggested: Value>Change</i>
35	the same as	129	7	3C1	<i>Suggested: Value>Comparison, Not limited to quantity</i>
36	most of the	126	7	3C1	
37	a fall in	114	5	3C1	<i>Suggested: Value>Change</i>
38	rise in the price	103	6	3C1	<i>Suggested: Value>Change</i>
39	a rise in the	102	6	3C1	<i>Suggested: Value>Change</i>
40	no change in	96	7	3C1	<i>Suggested: Value>Change</i>
41	A lot of	87	7	3C1	
42	is equal to	83	6	3C1	<i>Suggested: Value>Change</i>
43	the decrease in	83	6	3C1	<i>Suggested: Value>Change</i>
44	a fall in the	82	5	3C1	<i>Suggested: Value>Change</i>
45	is greater than	78	7	3C1	<i>Suggested: Value>Change</i>
46	by more than	77	7	3C1	<i>Suggested: Value>Change</i>
47	rise in the price of	77	6	3C1	<i>Suggested: Value>Change</i>
48	the same amount	71	7	3C1	<i>Suggested: Value>Change</i>
49	fall in the price	71	5	3C1	<i>Suggested: Value>Change</i>
50	a change in the quantity	70	5	3C1	<i>Suggested: Value>Change</i>
51	a decrease in the	69	6	3C1	<i>Suggested: Value>Change</i>
52	an increase in supply	66	6	3C1	<i>Suggested: Value>Change</i>
53	a change in the price	65	6	3C1	<i>Suggested: Value>Change</i>
54	the increase in the	64	6	3C1	<i>Suggested: Value>Change</i>
55	increase in the quantity of	61	6	3C1	<i>Suggested: Value>Change</i>
56	the fall in	61	5	3C1	<i>Suggested: Value>Change</i>
57	quantity of a	59	7	3C1	
58	a percentage of	58	5	3C1	
59	as a percentage of	56	5	3C1	
60	a rise in the price	56	5	3C1	<i>Suggested: Value>Change</i>
61	amount of the	55	7	3C1	
62	all the other	54	6	3C1	
63	fall in the price of	54	5	3C1	<i>Suggested: Value>Change</i>
64	much of the	52	7	3C1	
65	is the same as	52	7	3C1	<i>Suggested: Value>Comparison</i>
66	with no change in	52	4	3C1	<i>Suggested: Value>Change</i>
67	change in the value of	51	5	3C1	<i>Suggested: Value>Change</i>
68	as much as	50	7	3C1	<i>Suggested: Value>Comparison, Not limited to quantity</i>

C2. Tangible Framing				3C2	
69	the size of the	95	7	3C2	
70	the form of	62	7	3C2	
71	in the form of	57	7	3C2	
C3. Intangible Framing				3C3	
72	in the long run	329	8	3C3*	Intangible Framing and/or Time Reference (3C3/3D2)
73	in the short run	282	8	3C3*	Intangible Framing and/or Time Reference (3C3/3D2)
74	as a result	166	8	3C3	
75	at a price of	140	7	3C3	
76	in terms of	117	7	3C3	
77	in monopolistic competition	91	4	3C3	Discipline-specific
78	as a whole	78	7	3C3	
79	in the value of	74	8	3C3	
80	a result of	57	8	3C3	
81	in the case of	55	7	3C3	
82	as a result of	52	8	3C3	
83	of the total	52	7	3C3	
84	in the global economy	51	5	3C3	Discipline-specific
D. Place/Time/Text Reference				3D	
D1. Place Reference				3D1	
85	in the market	337	7	3D1*	Place and/or Intangible Framing (3D1/3C3), Discipline-specific
86	in the world	100	7	3D1	
87	in the foreign exchange market	81	1	3D1	Place and/or Intangible Framing (3D1/3C3), Low rang+ Discipline-specific
88	in other countries	69	6	3D1	
89	in the United States and	58	6	3D1	
D2. Time Reference				3D2	
90	in the future	89	7	3D2	
91	at the same time	79	7	3D2	
92	when the price is	67	7	3D2	
D3. Text Deixis				3D3	
93	shown by the	116	4	3D3	
94	in the table	96	6	3D3	
95	At point A	94	5	3D3	
96	on the X	88	5	3D3	

97	on the Y	84	5	3D3
98	curve shows the	83	5	3D3
99	at point C	71	5	3D3
100	in the figure	68	4	3D3
101	At point B	61	5	3D3
102	at the margin	48	6	3D3

D4. Multi-Functional Reference				3D4
103	the end of	135	7	3D4
104	at the end	106	7	3D4
105	at the end of	105	7	3D4

4.2 Discussion

In this section, I discuss the results of this study in detail. This discussion includes displaying statistics and charts to help visualize the results and the distribution of lexical bundles among the three types of functions; i.e. stance expressions, discourse organizers, and referential expressions. It also includes highlighting some particular LBs that show a degree of irregularity in terms of their frequencies, ranges, functions, and/or the extent to which they are discipline-specific. Examples of irregular frequency and/ or range include LBs with high frequency and low range, and LBs that can be common in a wide range of registers but have low frequency in FYETC. Another group of LBs that were subject to controversy among the judges and the researcher of this study are also discussed. For example, some judges had different views from those of the researcher on how to classify specific LBs and which function(s) these LBs served. Then, based on my experience in this study, I provide a brief description of some of the difficulties that may face many researchers during the application of the functional taxonomy of lexical bundles of Biber et al. (2004a). This is followed by some suggestions that may help better represent the functions some LBs serve, e.g. creating new subcategories.

Before discussing the distribution of lexical bundles in FYETC, it is necessary to highlight that many lexical bundles serve more than one function even in a single position. In fact, this is one of the common difficulties that face researchers when categorizing lexical bundles according to their function(s). For example, a lexical bundle such as ‘(You) *are expected to* (do this)’, which is a *prediction* lexical bundle; shows in this example that it has a *directive* function. This issue was discussed in a number of research papers and articles, e.g. Biber et al. (2004a), Biber and Barbieri (2007), and Chen (2008). Biber et al. (2004a) state that, “in some cases, a single bundle has multiple functions even in a single occurrence. For example, bundles

like *take a look at* and *let's have a look* function as both directives and topic introducers". They add, "in other cases, a single bundle serves different functions depending on the context" (pp. 383-384). Also, Chen (2008) indicates that one of the practical issues that usually faces researchers when classifying LBs according to the functional taxonomy is "the possibility of multiple functions of some lexical bundles, which makes it hard to place lexical bundles in one specific functional subcategory" (p. 118). In addition, Biber and Barbieri (2007) comment on the findings of their study and highlight that "In most cases, identification/focus bundles have a discourse organizing function" (p. 271). They further explain that a lexical bundle such as '*if you look at*' helps and "directs students' attention to the prop, indirectly introducing a new topic by reference to it" (p. 277). They reiterate the notion that stance bundles can be used as directives through stating that those bundles "more often have directive functions" (p. 281). Also, they add that "discourse organizing bundles are also used for topic elaboration or clarification" (p. 277). In addition, they indicate that some of the intention/prediction bundles can be topic introducers (p. 281). Furthermore, they explain that the function(s) of any given lexical bundle changes according to the way a writer or a speaker uses it. For example, Biber and Barbieri (2007, p. 277) explain that some lexical bundles that had been classified as referential identification/focus bundles in Biber et al. (2004a) were classified as discourse organizers in Biber and Barbieri (2007) "because they usually serve a discourse organizing function". Also, a bundle like '*do you think*' was found to be a personal epistemic stance lexical bundle in Biber and Barbieri (2007, p. 276); and was found to serve a discourse organizing topic introduction/ focus function in Sykes (2010) as it is used to share ideas and elicit opinions in academic interactive contexts, e.g. seminars and discussion groups (p. 46). However, it can also be used as an indirect suggestive or

directive bundle, e.g. *‘Do you think it would be better if...?’*; or in polite requests like *‘Do you think you can do...?’*.

Thus, we can see that many lexical bundles can serve more than one function even in one occurrence and that their classification can change depending on the situation and use. Therefore, it is assumed normal and inevitable to find some slight differences in the way researchers categorize their lexical bundles given that the function depends to a great extent on the situation and context in which a lexical bundle occurs.

Likewise, in this current study many lexical bundles (marked with an * in the table) have been found to serve more than one function even in their single occurrences, for example:

- *‘is expected to’* serves both the intention/ prediction (1B3) and the directive (1B2) functions;
- *‘there is a’*, *‘there are two’*, and *‘there are many’* are found in this study to serve the functions of topic introduction (2A) and referential identification/ focus (3A), in addition to quantity reference (3C1) as a third function for *‘there are many’*;
- *‘in part A’*, *‘in part B’*, and *‘in this chapter’* serve as both topic introducers (2A) and text deixis (3D3);
- *‘in the first’* can serve as a topic introducer (2A) and a multi-functional bundle (3D4);
- *‘when the price of a’* can work as a topic introducer (2A) and/or a time reference (3D2);
- *‘to illustrate the’* can be a topic introducer (2A) and/or a topic elaboration/ clarification bundle (2B);
- *‘in the quantity of’* can be a reference bundle that refers to quantity (3C1) and/ or intangible framing (3C3);

- '*in the long run*' and '*in the short run*' can be seen as intangible framing (3C3) and/ or time reference (3D2); and
- '*in the market*' and '*in the foreign exchange market*' can be intangible framing (3C3) or place reference (3D1), if we consider virtual places.

This multi-functionality of lexical bundles probably enables them to play a vital role as discourse building blocks in maintaining the text coherence and cohesion, bridging phrases and linking ideas, clarifying thoughts and concepts, allowing for smooth transitions, easing communication, and many more. Being able to serve more than one function indicates how essential and powerful these bundles are.

After showing how a lexical bundle can have many functions and those functions depend on the context, situation, and/ or the intention of the person who uses it; it is pivotal at this point to display and comment on the statistics and numbers of lexical bundles that are found in FYETC.

As for the distribution of the FYETC LBs across their functions, it is so obvious that the referential expressions have the largest proportion. The referential category has more lexical bundles than the stance expressions and the discourse organizers combined. That is, while the referential expressions cover 64% of all the bundles in FYETC; the stance expressions represent only 9%, and the discourse organizing bundles represent 27% of them [see **Figure 8** below].

Figure 8 The distribution of FYETC LBs according to function

As can be seen from the pie chart in *Figure 8* above, the referential expressions represent the biggest portion of FYETC LBs. This is in line with what Biber and Barbieri (2007) concluded in their study as they reported, “referential functions are the dominant use of lexical bundles in the written university registers” (p. 282).

In addition, the specification of attributes was the dominant subcategory of the referential bundles, containing 70 bundles that refer to quantity, tangible framing, or intangible framing attributes [See *Figure 9* below].

Figure 9 The dominance of specification of attributes as part of referential expressions

Out of the 70 bundles that refer to specification of attributes, the quantity specification was the dominant subcategory with 54 bundles, compared to only 3 tangible framing and 13 intangible framing bundles [See *Figure 10* below].

Figure 10 The dominance of quantity bundles as part of the specification of attributes

As it represents more than 32% of the total number of LBs in FYETC, the quantity specification is by far the dominant. This dominance of quantity lexical bundles may be due to the fact that FYETC was created out of Economics textbooks. Those textbooks belong to a discipline that may focus on values in general and quantity in particular as part of discussing the dynamics of markets and economy.

On the other hand, the least frequently used lexical bundles found in FYETC were both the personal epistemic stance bundles (1A1) and the imprecision referential bundles (3B). Both of these classes, (1A1) and (3B), had only one lexical bundle each. The personal epistemic stance bundle found in FYETC is '*do you think*' with a high frequency of 117 and appeared in 6 texts out of 8. The imprecision referential bundle is '*and so on*' with a frequency of 70 and a range of 6. Each of them represents 0.6% of the total number of lexical bundles found in FYETC. Presumptively, this can be due to the nature of academic writing prose and conventions where authors sometimes prefer not to address readers directly; instead, they tend to state what they see as facts and avoid refutation through hedging their claims. **Figure 11** below show all FYETC lexical bundles distributed across the different functional categories and subcategories of the taxonomy.

Figure 11 The functional distribution of all LBs in FYETC

Also, some LBs that are commonly used in everyday language have low frequency in FYETC LBs list. For example, the impersonal epistemic stance bundle ‘*it is possible*’ is common in real-life usage and is used in many cases “to hedge claims” (Biber & Barbieri, 2007, p. 276); yet, it has a frequency of 49 in FYETC. This can be seen as a low frequency if compared to other bundles in the same class (1A2) e.g. ‘*more likely to*’ with a frequency of 71. Other examples of the same sort include ‘*that there are*’ and ‘*is based on*’ that function as topic elaboration/ clarification bundles (2B) and have frequencies of 49 and 48, respectively. This can be an indication of how the usage of lexical bundles varies from one discipline to the other. Apparently, there is something that relates to economics as a discipline that discourages writers from using this type of LBs. It could be that economics is more about numbers, facts, and/ or quantities that leave little or no space for hedging. It could also be that this observation is only a

result of the comparison between one sort of science like economics to other sorts like humanities or everyday language. I believe that an in-depth qualitative research is needed to illuminate this point.

Interestingly, there are also some lexical bundles that seem unique to a discipline like Economics in the FYETC LBs list. These bundles can be hardly seen in other academic disciplines; and at the same time, they have higher frequencies than other common bundles that are used in real-life situations. For example, bundles like '*in monopolistic competition*', '*in the foreign exchange market*', '*a closer look at markets*', and '*in the global economy*' (with frequencies of 91, 81, 72, and 51, respectively) have higher frequencies than other bundles that are deemed common such as '*as much as*', '*it is possible*', '*that there are*', and '*is based on*' (with frequencies of 50, 49, 49, and 48, respectively). To all appearances, this can be considered an evidence of how lexical bundles are discipline-bound and that the frequency depends on the characteristics of that discipline.

Furthermore, some bundles have high frequencies in FYETC and low ranges. There are two specific cases that this description applies to:

- 1) '*a closer look at markets*' with a frequency of 81 and a range of 1 text, and
- 2) '*in the foreign exchange market*' with a frequency of 72 and a range of 2 texts.

These two lexical bundles seem to exhibit how author preferences and discipline requirements play vital roles in the frequency and use of lexical bundles and their likelihood to appear in any frequency-based research. However, due to the limited number of texts in this study i.e. eight texts, these bundles have been accepted in the FYETC LBs list depending on a minimum percentage of 10% of the total number of texts as recommended by Hyland (2008).

When the final list was shown to three different judges, they provided the researcher with valuable feedback and observations regarding the classification of FYETC lexical bundles. As an example of these observations, the bundle '*do you think*' was seen by one of the judges as an invitation from speakers/ writers to the listeners/ readers to share their opinions about matters. In other words, '*do you think*' can be used to introduce topics (2A) in order to draw the attention of listeners/ readers to a new matter, while it is classified as a personal epistemic stance bundle (1A1) in FYETC LBs list. Also, the same judge and the researcher shared the same view that '*now going to*' was difficult to be seen inseparable from verb 'to be'.

In addition, there were different views than that of the researcher regarding specific lexical bundles and their functional classification. For example, the bundle '*it is possible*' that is classified in this study as an impersonal epistemic stance bundle (1A2) was seen by one of the judges to be an ability bundle (1B4) that belongs to attitudinal/ modality stance bundles. Most likely, this disagreement stems from the judge's preference to seek guidance from how the bundle is actually classified in Biber et al. (2004a); while the researcher's preference was to apply the class description provided in the same source. That is, Biber et al. (2004a) state that "epistemic stance bundles comment on the knowledge status of the information in the following propositions: certain, uncertain, or probable/ possible... [and]... impersonal epistemic stance bundles express degrees of certainty rather than uncertainty" (p. 389). Therefore, the bundle '*it is possible*' is classified in this study as an impersonal epistemic stance (1A2).

The lexical bundle '*you can see*', which is classified in this study as an ability stance expression (1B4) and identification/ focus bundle (3A), was seen by one of the judges to serve as a topic introducer discourse organizer (2A). In fact, this bundle is one of a very small group of bundles found in FYETC that show how writers aim their speech directly to readers in order to

engage them with what they read. In some cases, this bundle is preceded by other bundles like '*if you look at X*' or '*by looking at X*'; in this case, it can be argued that it serves as a topic introducer. However, the bundles preceding it are the ones that serve the topic introduction function; while the bundle '*you can see*' itself does not. Instead, it serves either the ability and/or the identification/ focus function.

Also, '*the relationship between*' is classified in this study as a topic elaboration/ clarification bundle (2B) since it helps provide additional clarification and explanation regarding how different entities relate to each other. This bundle was seen by one of the judges to serve a tangible framing function (3C2) as it refers to some sort of a frame.

Other bundles that captured the attention of the judges were '*in the long run*' and '*in the short run*'. One of the judges argued that these two bundles can also function as time reference (3D2), while they were seen by others to serve both the time reference (3D2) and the intangible framing function (3C3). So, the two LBs were accepted in FYETC list for their ability to have more than one function, and an asterisk marked this in the list.

In addition, the bundle '*at a price of*' was seen by one of the judges to serve a tangible framing referential function (3C2). Yet, this bundle is classified in this study as an intangible framing bundle (3C3). The divergent views regarding the classification of this bundle could have stemmed from how differently each person conceives the concept of 'price' e.g. tactile/ material vs abstract/ immaterial value.

Finally, the lexical bundle '*shown by the*' was argued by one of the judges to be a referential bundle used for identification/ focus (3A). This bundle is classified in this study as a text deixis bundle (3D3), which is used by writers to refer their readers to specific parts of the text (i.e. figures, tables, charts, sign, etc.) for better illustration or visualization of data.

All in all, the generous feedback of the judges of this study were all welcomed and taken into consideration since they provided the researcher with useful information, insight, and eye-opening hints. The divergence in views on how to functionally classify LBs is deemed normal, regarding a number of inevitable factors such as: 1) the judgment depends to a great extent on the individual intuition, which does not seem to be totally objective; 2) context plays an important role in deciding the pragmatic function of lexical bundles; and 3) even a single text can have many interpretations, which would eventually lead to different views on how language/utterances are used, what they mean, and/ or what pragmatic functions they serve.

Disagreement about matching lexical bundles to their functions seems to be inevitable and displays some of the difficulties researchers may face during the process. That is, researchers sometimes struggle to match specific lexical bundles to their subcategories in the taxonomy. This issue was pointed out by other researchers as well. For example, Chen (2008) provides a critique of the functional taxonomy in order to demonstrate some of these difficulties. She indicates that the taxonomy as it is, despite being a good tool, “may fail to capture multiple functions that some lexical bundles have in the original context” (p. 117). This notion can be valid especially that Biber et al. (2004, p. 383) point out that the functional taxonomy proposed in their paper is preliminary and that their approach to developing this taxonomy is mainly “inductive”. This means that the taxonomy is more generated from the data they had than created in a top-down manner. Thus, it entails that there should be continuous development of the functional taxonomy in order to help reflect the language functions that exist in a wide range of disciplines. Also, it entails that researchers who are interested in different disciplines should share their ideas and suggestions in order to improve the taxonomy and ensure more inclusion of different disciplines.

There is no doubt that the functional taxonomy provided a solid platform for researchers who wish to classify the LBs they find in their corpora. However, since it is a preliminary taxonomy, I believe that any attempt to improve its performance would be welcomed, e.g. adding new subcategories, renaming existing categories to better reflect its contents, etc. For example:

- ‘Value’ should be a category that refers to values in general by any means, concrete or abstract; which can comprise some referential subcategories like quantity, change (increase/decrease), comparison between/ among values, etc. This category should include lexical bundles such as *‘the amount of the’*, *‘much of the’*, *‘a lot of’*, *‘some of the’*, *‘the majority of’*, *‘a percentage of’*, *‘an increase in’*, *‘a fall in the’*, *‘a change in’*, *‘is less than’*, *‘equal to the’*, etc.

- ‘Conditional Probability’, should be a subcategory; for example, *‘in the case of/ in case of’*, *‘in this case’*, *‘when the price is’*, *‘if this happens’*, etc.

- ‘Cause-Effect/ Dependency Relationship’ can also be a distinctive subcategory of the topic elaboration/ clarification discourse organizers. It can include lexical bundles such as *‘as a result (of)’*, *‘depends on the’*, *‘that results from’*, *‘which leads to’*, *‘in order to’*, *‘is based on’*, *‘that arises from’*, *‘based on the’*, etc.

- ‘Sequencing’ can include lexical bundles such as *‘in the first’*, *‘first of all’*, *‘last but not least’*, *‘at the end of’*, etc.

One more observation that draws attention is that despite the fact that lexical bundles have been normally defined as combinations of three or more words (Biber & Conrad, 1999, p. 188; Cortes, 2004, p. 400; Wood, 2010, p. 92; Longman Grammar of Spoken and Written English, 1999), many researchers tend to set the minimum cluster size of LBs in their studies to exclusively four-word constructions. Wood (2015) states, “a general consensus has emerged that four words is the most productive unit to study”. He adds, “three-word sequences are quite

common in discourse and tend to be difficult to categorize as bundles, especially in terms of functions” (p. 132). Therefore, this tendency to limit and focus research on lexical bundles of at least four words, I believe, deprives us from really useful and widely used LBs such as: *as soon as*, *as well as*, *and so on*, *in order to*, *first and foremost*, *in case of*, *in other words*, *is expected to*, *by and large*, etc. In addition, I believe that many two-word strings can be considered lexical bundles if we disregard the three-word condition; since they have the same qualities that qualify a word string to be a lexical bundle, i.e. frequency, range, and function. For example; combinations like *such as*, *for example*, *as well*, *generally speaking*, *that is*, *in brief*, *in particular*, *for instance*, etc. are all highly frequent, widely used in different texts, and have LBs functions such as discourse organizing or referential functions.

Regarding the limitations in this study, the most noteworthy limitation is that despite having a corpus size of 1,346,658 words in total; the number of textbooks that were used to create the corpus was limited. This limited number of texts did not allow for further investigation of the range and usage of lexical bundles. It did not allow for comparisons among similar lexical bundles in terms of how widely they were used.

Another limitation in this study that is worth mentioning is that none of the formulaic sequences that can be described as frames with fillable slots, i.e. concgrams, appeared in FYETC LBs list. The words that compose a concgram are not adjacent to each other in a form of a sequence. Instead, they are separated by fillable slots that writers/ speakers fill with content words, phrases, or clauses according to what they want to communicate. For example, in a frame like ‘*not only..., but also...*’, the four words composing the bundle do not share borders. This type of sequences has not been possible for text analysis software to detect for some time, but with the development of technology this issue has been resolved. For the purpose of this study,

which is identifying the most frequent lexical bundles in Economic textbooks and classifying them in line with Biber et al. (2004a) functional taxonomy, the researcher decided to focus only the contiguous bundles in this discipline.

Yet, if this study were to be repeated, two limitations should be overcome through:

1) increasing the number of textbooks, which would allow for a meaningful comparison of range and usage of lexical bundles. It would also allow, as a result, for a wider variety of authors and a bigger size of corpus. This would eventually lead to a larger number of lexical bundles in the final list.

2) investigating concgrams that exist in the corpus, which would provide more insight into how this type of formulaic sequences work in context. This would be useful in deepening our understanding of how different types of lexical bundles serve the cohesion of a text. It could also open our eyes towards new categories and subcategories of functions that lexical bundles serve.

In addition, if this study were to be repeated, I would ensure the exclusivity of genuine components. That is, each component of the bundle has to be an authentic constituent. In many cases in the literature written on lexical bundles, some clusters are named 4-word LBs while at least one of its components is not really a genuine part of it. For example, in some studies on 4-word lexical bundles, authors list a 3-word LB followed by another word that does not belong to the bundle. This could be because they wanted to fulfill the cluster size requirements that were chosen for their study. For instance, in a bundle such as '*in case of the*', the definite article '*the*' is not a genuine component of the bundle. In fact, it is there to define what follows the bundle. So, I believe that calling '*in case of the*' a 4-word lexical bundle is a misnomer. What may support this point of view, beside the fact that omitting '*the*' would not affect the LB itself

whatsoever neither pragmatically nor semantically, is that ‘*in case of*’ is widely used in other instances without ‘*the*’, e.g. ‘*in case of fire*’. In the Corpus of Contemporary American English (COCA), ‘*in case of the*’ has a frequency of 33 per million words; while ‘*in case of*’ has a high frequency of 1604 per million words mainly in the academic register [See **Figure 12** below]. This can mean that ‘*in case of*’ is the lexical bundle with all its genuine components, while ‘*the*’ is added in some cases only to match what follows the bundle.

The screenshot shows the COCA interface with the 'FREQUENCY' tab selected. It displays two search results for the phrase 'in case of'. The first result is 'IN CASE OF THE' with a frequency of 33. The second result is 'IN CASE OF' with a frequency of 1604. Each result has a 'CONTEXT' button and a 'FREQ' column. A 'COMPARE' button is visible in the top right corner of the results area.

	CONTEXT	FREQ
1	IN CASE OF THE	33
1	IN CASE OF	1604

Figure 12 The frequencies of ‘*in case of the*’ and ‘*in case of*’ in the COCA

Finally, as a normal result of using text analysis software, a huge amount of overlapping appears in the word clusters lists. The degree of overlapping among LBs refers to the case(s) when two or more of the lexical bundles intersect and/or extend over each other. For example, a word cluster like ‘*is based on the*’ appeared in the raw results three times as ‘*is based on*’, ‘*based on the*’, and ‘*is based on the*’. This overlapping is a constant issue that was repeatedly indicated in the work of many researchers and scholars e.g. Wood and Appel (2014), Hyland (2008), Chen (2010), Cortes (2004), Biber et al. (2004a), and Biber and Barbieri (2007). In addition, it is also important to note that two different forms of overlapping were noticed in the raw results of this study:

- 1) Overlapping among a group of bundles of the same root

For example, *'you will be able to'* appeared in the raw results of this study several times as *'be able to'*, *'will be able'*, *'will be able to'*, *'you will be able'*, and *'you will be able to'*. So, if this study were to be conducted again, I would include only the core bundle with its genuine components to the FYETC LBs list. For instance, in the example above only *'be able to'* would be in the FYETC LBs list, since its three components form its core and are indispensable.

2) Overlapping between two consecutive LBs

For example, *'is willing to'* and *'to pay for'* share the preposition *'to'* in *'is willing to pay for'*, where *'to'* is part of both the bundles. So, I believe that the longer string of *'is willing to pay for'* should not be counted as one bundle; instead, they should be dealt with as two different overlapping bundles.

However, in some other cases, LBs seem to overlap and only one of the overlapping LBs should be included in the list. But, it is not always true that they are the same despite their similarity. For example, it could be argued that *'as a result'* and *'as a result of'* are the same depending on their resemblance and the fact that they are derived from the same root. But, by referring to the frequency and range of each of them we can see that they have different frequencies, which means that they occur in different occasions. To be specific, *'as a result'* recorded a high frequency of 166, while *'as a result of'* recorded only 52; which means that the former bundle has been used in different instances than the latter. Also, these two bundles are used differently in terms of their position in a sentence and their relationship with what precedes and what follows. For instance, *'as a result'* can be used at the end of a sentence or at the beginning of a sentence as a topic introducer; while *'as a result of'* has to be followed by a noun or a phrase.

5. Conclusion

There is a fast-growing body of research on the different types of formulaic language e.g. idioms, collocations, sentence-building frameworks, lexical bundles, etc. One of the active areas of research has been lexical bundles for the importance they represent. Many studies have been conducted on the usage of lexical bundles in different registers and disciplines, e.g. Biber et al. (1999, 2004a), Biber and Conrad (1999), Cortes (2004), Scott and Tribble (2006), Chen (2008), Wood and Appel (2014), etc. With the expansion in research on lexical bundles came the need for a theoretical framework that would help researchers investigate the communicative and pragmatic functions that lexical bundles serve. The functional taxonomy of lexical bundles, suggested in Biber et al. (2004a), gained favour and has been the platform that many researchers use and depend on in their studies, including the current study. This taxonomy describes three main categories of functions for the lexical bundles i.e. stance expressions, discourse organizing, and referential expressions. It divides these categories into subcategories that help reaching more detailed functions. For example, the discourse organizing category is divided into two subcategories i.e. topic introduction/ focus, and topic elaboration/ clarification.

The advancement in technology brought significantly helpful tools that are used in the field of applied linguistics and discourse studies. These tools include the text analysis software that allowed for rapid and advanced corpus analysis. The text analysis software employed in this study is *WordSmith Tools 6.0* (Scott, 2007). This software is user-friendly and allows researchers to set their preferred cluster size, minimum frequency, and the range of texts in which the clusters appear.

This study is a corpus-based study on lexical bundles. It aims to cover a gap in research by investigating the lexical bundles that exist in first year Economics textbooks, hence came the name of the corpus, i.e. FYETC. The research questions in this study are: 1) What are the most frequent lexical bundles in first year Economics textbooks that are used in Canadian universities? and 2) What functions do these lexical bundles serve?

In preparation for this study, the researcher chose the setting to be Canadian universities, especially those in Ontario. In brief, the reason for this is that Ontario is home to around 40% of the Canadian population, an academic destination to around 40% of the international students, and home to the biggest number of educational institutions across Canada.

The textbooks used as the basis for creating the corpus were chosen according to some criteria including being foundational in the field of Economics, being used for first year students, and being shared across as many universities as possible. The application of these criteria resulted in narrowing the number of textbooks to eight. These books were then used to create the First Year Economics Textbooks Corpus (FYETC). In total, the eight textbooks created a corpus that consisted of 1,346,658 words. *WordSmith Tools 6.0* was used in order to pinpoint the most frequently used word clusters. The cluster size was set to range from three to eight words per cluster to allow for as many clusters to appear as possible. The frequency was set to 40 occurrences per million words. The range of texts was set to a minimum of 10%, as suggested by Hyland (2008). Since the total number of texts was eight, a single text in this study would exceed this percentage to reach 12.5%; so, the minimum range accepted was one text.

After obtaining the raw results, the lexical bundles identification criteria were applied. The criteria application was the tool that qualified a word cluster to be considered a lexical bundle and be incorporated in the FYETC LBs list. These criteria included that the cluster

should: 1) have the minimum frequency, 2) have the minimum range, 3) match at least one function of Biber et al. (2004a) functional taxonomy of lexical bundles. The application of these criteria resulted in a list of lexical bundles that were then classified according to their functions and organized in an ascending order from the most to the least frequent bundles. Some bundles were marked by an asterisk to indicate that they could serve more than one function. The list was, then, given to three knowledgeable judges for revision and providing feedback prior to seeking the approval of my thesis supervisor.

In terms of the results of the study, the total number in FYETC LBs list was 165 lexical bundles serving all three categories: stance, discourse organizing, and referential functions. The distribution statistics show that the referential expressions are the dominant category among the three, recording nearly 64% of all the bundles in FYETC (105 bundles out of 165). Stance expressions form only 9% of the lexical bundles, while the discourse organizers were 27%. Within the referential expressions, the bundles that refer to the specification of attributes (quantity, tangible framing, and intangible framing) were the majority. The quantity attribute formed the biggest portion of all recording 54 out of the 70 bundles of the specification of attributes, and 32% of all the bundles in FYETC. On the other hand, the personal epistemic stance expressions and the imprecision referential bundles were the least used ones; each recorded 0.6% of the total number of lexical bundles in FYETC.

In addition, some bundles showed irregularities in terms of frequency and range. For example, some bundles like '*in the foreign exchange market*' had high frequencies and low ranges, which can be an indication of the author preference and/ or the topic requirements. Other bundles like '*in monopolistic competition*' had notable high frequencies but seemed to be rarely used outside this discipline, which can signify how the frequency and usage of LBs are

discipline-bound. A third group of bundles like '*is based on*' and '*it is possible*' seemed very common but had low frequencies in FYETC, which is again an indication on how the discipline dictates the frequency and usage of lexical bundles.

For the most part, the judges agreed with the researcher of this study on how the lexical bundles are classified. However, there were some controversial lexical bundles, which the judges had different views on. This controversy is deemed normal regarding the characteristics and nature of LBs that allow them to serve more than one function even in a single situation. For example, '*there is a*' can be a topic introducer and at the same time an identification/ focus referential bundle. Serving more than one function can pose a difficulty for researchers who wish to classify the LBs they find in their corpora. Also, sometimes it seems that there are not enough subcategories in the functional taxonomy to reflect the functions that LBs serve. In this regard, the researcher suggests that some subcategories can be added to the functional taxonomy, e.g. conditional probability, cause-effect relationship, comparison, etc.

Should this study be re-conducted, the researcher would increase the number of textbooks used to create the FYETC. The limited number of texts, i.e. eight textbooks, made it difficult to further investigate the range and usage of LBs, and to run comparisons among similar LBs to see how widely they are used. Also, I would investigate the concgrams in order to see how they behave in context and how they serve the cohesion of texts. Finally, I would further focus on the core constructions of lexical bundles and list only those with genuine components. This would ensure that only the lexical bundles with no other additions are in the FYETC list.

This study resulted in the creation of a list of lexical bundles that are frequently used in Economics textbooks. The study and the FYETC LBs list can be of benefit for many groups that are involved in applied linguistics and education. For example, they can be used by researchers

in applied linguistics in two different ways: 1) Since the list is the result of a specialized corpus, it can be used in comparison with other corpora to identify differences in language use across disciplines; and 2) Since the corpus is compiled from Economics textbooks, which makes it a pedagogic corpus, the list can be used as a tool to evaluate the degree to which EAP textbook language caters to the needs of business students.

For EAP students, the list can provide a rich source of lexical bundles and discourse functions that they can use in their writing. This can be especially evident regarding the fact that the FYETC LBs list was created from first year Economics textbooks and that the majority of EAP students are the novices in the field. Also, it is assumed that novice students mainly depend on textbooks rather than academic articles; therefore, the list can prove useful in familiarizing them with the discipline and allow them to be “more effectively grasping these constructions in their writing” (Liu, 2012, p. 33). In addition, it can help them be more cognitively engaged in the reading material that they encounter.

Also, EAP teachers, material developers, and curriculum designers can use the FYETC LBs list as a source of discipline-specific language to help promote students’ learning and language acquisition. This can be done through developing classroom activities that raise students’ awareness of the most recurrent expressions in their field and familiarize them with the lexical bundles they are more likely to encounter in their discipline. These activities can include not only the functions of LBs, but also how they connect with the discourse that precedes and follows them; and what parts of speech can do that. Teachers can also use the list to evaluate the degree to which their students’ writing reflects their knowledge of the field/ discipline.

References

- Aijmer, K., & Altenberg, B. (2014). *English Corpus Linguistics*. Routledge.
- AlHassan, L., & Wood, D. (2015). The effectiveness of focused instruction of formulaic sequences in augmenting L2 learners' academic writing skills: A quantitative research study. *Journal of English for Academic Purposes*, 17, 51–62.
- Altenberg, B. (1990). Speech as linear composition. In G. Caie, K. Haastrup, A.L. Jakobsen, J.E. Nielsen, J. Sevaldsen, H. Specht and A. Zettersten (eds.) *Proceedings from the Fourth Nordic Conference for English Studies*, Helsingor, May 11-13 1989. Copenhagen University, Department of English. 133-143.
- Altenberg, B. (1993). Recurrent verb-complement constructions in the London-Lund Corpus. In J. Aarts, P. de Hahn, & N. Oostdijk (Eds.), *English language corpora: Design, analysis and exploitation* (pp. 227–245). Amsterdam: Rodopi.
- Baayen, H., & Lieber, R. (1991). Productivity and English derivation: A corpus-based study. *Linguistics*, 29(5), 801–844.
- Barkema, H. (1993). Idiomaticity in English NPs. In J. Aarts, P. de Hahn, & N. Oostdijk (Eds.), *English language corpora: Design, analysis and exploitation* (pp. 257–278). Amsterdam: Rodopi.
- Becker, J. D. (1975). *The phrasal lexicon*. In *Proceedings from TINLAP '75: The 1975 Workshop on Theoretical Issues in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written*

- registers*. Philadelphia, PA: John Benjamins.
- Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26(3), 263–286.
- Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. *Language and Computers*, 26, 181–190.
- Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: An initial taxonomy. In A. Wilson, P. Rayson, & T. McEnery (Eds.), *Corpus linguistics by the Lune: a festschrift for Geoffrey Leech* (pp. 71–93). Frankfurt, DE: Peter Lang.
- 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., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, 36(1), 9–48.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow, UK: Pearson.
- Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, H. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research*, 10, 245–261.
- Bybee, J. (2002). Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition*, 24, 215–221.
- Chafe, W. L., Du Bois, J. W., & Thompson, S. A. (1991). Towards a new corpus of spoken American English. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 64–82). London: Longman.

- Chafe, W. L., Du Bois, J. W., & Thompson, S. A. (2014). Towards a new corpus of spoken American English. In *English Corpus Linguistics*, 76–94. Routledge.
- Chen, L. (2008). An investigation of lexical bundles in electrical engineering introductory textbooks and ESP textbooks (Master's thesis). Carleton University, Ottawa, Ontario.
- Chen, L. (2010). An investigation of lexical bundles in electrical engineering introductory textbooks and ESP textbooks. In D. Wood (Ed.), *Perspectives on formulaic language* (pp. 107–125). New York: Continuum.
- Chen, X. (2009, December). *The role of formulaic sequence in speech fluency development in ESL*. Paper presented at the International Conference on Computational Intelligence and Software Engineering, Wuhan, China.
- Cieslicka, A. (2006). Literal salience in on-line processing of idiomatic expressions by second language learners. *Second Language Research*, 22, 115–144.
- Conklin, K. & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and non-native speakers? *Applied Linguistics*, 29, 72–89.
- Conklin, K. & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, 32, 45–61.
- Conrad, S. (1996). Investigating academic texts with corpus-based techniques: An example from biology. *Linguistics and Education*, 8, 299–326.
- Cortes, V. (2002). Lexical bundles in freshmen composition. In R. Reppen, S. M. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 131–145). Amsterdam: John Benjamins.

- Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23(4), 397–423.
- Cortes, V. (2007). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education*, 17(4), 391–406.
- Cortes, V. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora*, 3(1), 43–57.
- Cortes, V., Jones, J., & Stoller, F. (2002, April). *Lexical bundles in ESP reading and writing*. Paper presented at the TESOL Conference, Salt Lake City, UT.
- Coulmas, F. (1979). On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics*, 3(3/4), 239–266.
- Coulmas, F. (Ed.). (1981). *Conversational routines*. The Hague: Mouton.
- Cowie, A. P. (1992). Multiword lexical units and communicative language teaching. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 1–12). Basingstoke, UK: Macmillan.
- Csomay, E., & Cortes, V. (2009). Lexical bundle distribution in university classroom talk. *Language and Computers*, 71(1), 153–168.
- Erman, B. & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, 20(1), 29–62.
- Foster, P. (2001) Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic task: Second language learning, teaching and testing* (pp. 75–93). Harlow, UK: Longman.
- Francis, W. N. (1982). *Problems of assembling and computerizing large corpora*. In S.

- Johansson (Ed.), *Computer corpora in English language research* (pp. 7–24). Bergen, Norway: Norwegian Computing Centre for the Humanities.
- Gibbs, R., & Gonzales, G. (1985). Syntactic frozenness in processing and remembering idioms. *Cognition*, 20, 243–259.
- Gibbs Jr, R. W., Nayak, N. P., & Cutting, C. (1989). How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of Memory and Language*, 28(5), 576–593.
- Glucksberg, S. (1989). Metaphors in conversation: How are they understood? Why are they used? *Metaphor and Symbol*, 4(3), 125–143.
- Government of Canada, Global Affairs Canada, International Education, International Education Promotion. (2016, August 29). Global Affairs Canada. Retrieved from http://www.international.gc.ca/education/report-rapport/economic-impact-economique/sec_4.aspx?lang=eng
- Government of Canada (2016, October 20). *International students in Canadian universities, 2004-2005 to 2013-2014*. Retrieved from <http://www.statcan.gc.ca/pub/81-599-x/81-599-x2016011-eng.htm>
- Government of Canada (2017, August 22). *International students, immigration and earnings growth*. Retrieved from <http://www.statcan.gc.ca/daily-quotidien/170822/dq170822c-eng.htm>
- Halliday, M. A. K., & Hasan, R. (1989). *Language, context, and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.
- Halliday, M. A. (1993). Towards a language-based theory of learning. *Linguistics and Education*, 5(2), 93–116.

- Harris, K. (2017, September 3). *Foreign students flock to Canada as government struggles to get grads to stay*. Retrieved from <http://www.cbc.ca/news/politics/international-students-jump-1.4268786>
- Hickey, T. (1993). Identifying formulas in first language acquisition. *Journal of Child Language*, 20, 27–41.
- Hoey, M. (2005). *Lexical priming: A new theory of word and language*. New York: Routledge.
- Howarth, P. A. (1996). *Phraseology in English academic writing: Some implications for language learning and dictionary making*. Tübingen, Germany: Niemeyer.
- Hunston, S. (2002). *Corpora in applied linguistics*. Stuttgart, Germany: Ernst Klett Sprachen.
- Hunston, S. & Laviosa, S. (2000). *Corpus linguistics*. Birmingham, UK: Centre for English Language Studies, The University of Birmingham.
- Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4–21.
- Hyland, K. (2008b). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, 18(1), 41–62.
- Jiang, N. A., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, 91(3), 433–445.
- Kennedy, G. (1998). *An Introduction to corpus Linguistics*. New York: Longman.
- Khodadady, E., & Shamsaei, S. (2012). Formulaic sequences and their relationship with speaking and listening abilities. *English Language Teaching*, 5(2), 39–49.
- Kjellmer, G. (1984). Some thoughts on collocational distinctiveness. In J. Aarts & W.

- Meijs (Eds.), *Corpus linguistics: Recent developments in the use of computer corpora in English language research* (pp. 163–171). Amsterdam: Rodopi.
- Levy, S. A. (2003). Lexical bundles in professional and student writing (Doctoral dissertation). University of the Pacific, Stockton, California.
- Lin, P. (2010). The phonology of formulaic sequences: A review. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 174–193). New York: Continuum.
- Lin, P. M. S. & Adolphs, S. (2009). Sound evidence: Phraseological units in spoken corpora. In A. Barfield & H. Gyllstad (Eds.), *Collocating in another language: Multiple interpretations*. Basingstoke, UK: Palgrave MacMillan.
- Liu, D. (2012). The most frequent multiword constructions in academic written English: A multi-corpus study. *English for Specific Purposes*, 31(1), 25–35.
- Meinardi, M. (2009). Speed bumps for authentic listening material. *ReCALL*, 21(3), 302–318.
- Millar, N. (2010). The processing of malformed formulaic language. *Applied Linguistics*, 32(2), 129–148.
- Moon, R. (1997). Vocabulary connections: Multi-word items in English. *Vocabulary: Description, Acquisition and Pedagogy*, 40–63.
- Moon, R. (1998a). *Fixed expressions and idioms in English*. Clarendon, UK: Oxford University Press.
- Moon, R. (1998b). Frequencies and forms of phrasal lexemes in English. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 79–100). Clarendon, UK: Oxford University Press.

- Morley, J. (2001). Aural comprehension instruction: Principles and practices. *Teaching English as a second or Foreign Language*, 3, 69–85.
- Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: 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. M. (2009). English L1 and L2 Speakers' knowledge of lexical bundles. *Language Learning*, 59(3), 647–686.
- O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge, UK: Cambridge University Press.
- Partington, A. (1998). *Patterns and meanings*. Philadelphia, PA: John Benjamins.
- Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication* (pp. 191-226). New York, NY: Longman.
- Peters, A. M. (1977). Language learning strategies: Does the whole equal the sum of the parts? *Language*, 53(3), 560–573.
- Peters, A. M. (1983). *The units of language acquisition*. Cambridge, UK: Cambridge University Press.
- Read, J., & Nation, P. (2004). Measurement of formulaic sequences. In N. Schmitt (Eds.), *Formulaic sequences: Acquisition, processing, and use*, 9, 23-35.
- Renouf, A. (1992). What do you think of that: A pilot study of the phraseology of the core words of English. In G. Leitner (Ed.), *New directions in English language*

- corpora: Methodology, results, software developments*. Mouton de Gruyter, Berlin, 301-317.
- Renouf, A., & Sinclair, J. (1991). Collocational frameworks in English. *English Corpus Linguistics*, 128–143.
- Schmitt, N. (Ed.). (2004). *Formulaic sequences: Acquisition, processing, and use* (Vol. 9). Philadelphia, PA: John Benjamins.
- Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing, and use* (pp. 127–151). Philadelphia, PA: John Benjamins.
- Scott, M. (2007). *Oxford Wordsmith Tools: Version 5.0*. Released June 2007 from <http://www.lexically.net>.
- Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education* (Vol. 22). Philadelphia, PA: John Benjamins.
- Simpson, R. (2004). Stylistic features of academic speech: The role of formulaic expressions. In T. Upton & U. Connor (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 37–64). Amsterdam: John Benjamins.
- Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31, 487-512.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, 27, 1–22.

- Swinney, D., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behaviour*, 18, 523–534.
- Taguchi, N. (2007). Chunk learning and the development of spoken discourse in a Japanese as a foreign language classroom. *Language Teaching Research*, 11, 433–457.
- Tremblay, A., & Baayen, H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 151–172). New York: Continuum.
- Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye- movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: acquisition, processing, and use* (pp. 153–172). Philadelphia, PA: John Benjamins.
- Van Lancker, D., Canter, G., & Terbeek, D. (1981). Disambiguation of ditropic sentences: Acoustic and phonetic cues. *Journal of Speech and Hearing Research*, 24, 330–335.
- Van Lancker-Sidtis, D. (2003). Auditory recognition of idioms by first and second language speakers of English. *Applied Psycholinguistics*, 24, 45–57.
- Weinert, R. (1995). The role of formulaic language in second language acquisition: A review. *Applied Linguistics*, 16(2), 180–205.
- Wood, D. (2001). In search of fluency: What is it and how can we teach it? *Canadian Modern Language Review*, 57(4), 573–589.

- Wood, D. (2002). Formulaic language acquisition and production: Implications for teaching. *TESL Canada Journal*, 20(1), 1–15.
- Wood, D. (2006). Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *Canadian Modern Language Review*, 63(1), 13–33.
- Wood, D. (2009a). Preparing ESP learners for workplace placement. *ELT Journal*, 63(4), 323–331.
- Wood, D. (2009b). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *The Canadian Journal of Applied Linguistics*, 12(1), 39–57.
- Wood, D. (2010a). *Formulaic language and second language speech fluency: Background, evidence, and classroom applications*. New York: Bloomsbury.
- Wood, D. (2010b). Lexical clusters in an EAP textbook corpus. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 8-106). New York: Continuum.
- Wood, D. C., & Appel, R. (2014). Multiword constructions in first year university engineering and business textbooks and in EAP textbooks. *Journal of English for Academic Purposes*, 15, 1–13.
- Wood, D. (2015). *Fundamentals of formulaic language: An introduction*. New York: Bloomsbury.
- Wray, A. (1999). Formulaic sequences in second language teaching: Principles and practice. *Applied Linguistics*, 21(4), 463–489.

- Wray, A. (2002). *Formulaic language and the lexicon*. New York: Cambridge University Press.
- Wray, A. (2004). 'Here's one I prepared earlier'. *Formulaic sequences: Acquisition, processing, and Use*, 9, 249-268.
- Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.
- Wray, A. (2009). Identifying formulaic language: Persistent challenges and new opportunities. *Formulaic Language*, 1, 27–51.
- Wray, A. & Fitzpatrick, T. (2008). Why can't you just leave it alone? Deviations from memorized language as a gauge of nativelike competence. In F. Meunier & S. Granger (Eds.), *Phraseology in language learning and teaching* (pp. 123–148). Amsterdam: John Benjamins.
- Wray, A., & Namba, K. (2003). Use of formulaic language by a Japanese-English bilingual child: A practical approach to data analysis. *Japanese Journal for Multilingualism and Multiculturalism*, 9 (1), 24–51.
- Wray, A. & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication*, 20(1), 1–28.
- Yorio, C. (1980). Conventionalized language forms and the development of communicative competence. *TESOL Quarterly*, 16(4), 433–442.