THESIS

LEXICAL BUNDLES IN MASTER'S-LEVEL FINANCE RESEARCH ARTICLES

Submitted by

Alhassane Ali Drouhamane

Department of English

In partial fulfillment of the requirements

For the degree of Master of Arts

Colorado State University

Fort Collins, Colorado

Spring 2016

Master's Committee

Advisor: Tatiana Nekrasova-Beker

Anthony Becker
Vickie Bajtelsmit

ABSTRACT

LEXICAL BUNDLES IN MASTER’S LEVEL FINANACE RESAECR ARTICLES

Lexical bundles are a type of formulaic sequences mainly identified on the basis of their frequencies and ranges. They have been found to consistently serve important discourse functions in academic prose, where, for example, they are used to evaluate or to refer to the size of something (Hyland, 2008a). Their forms, functions and uses were also found to be different in different academic disciplines. The present study extends this line of investigation by directly investigating the extent to which the four-word lexical bundles relied upon in master’s-level finance research articles differ from or are similar to those used in other academic disciplines, including business texts. Analyzing a corpus of 1,034, 587 words, the researcher found that more than 60% of lexical bundles in master’s-level finance research articles were identified in earlier studies on lexical bundles used in academic prose. However, 33 lexical bundles identified in the current study were not identified in previous literature. Structurally, like in previous literature, most bundles were found to be noun phrase and prepositional phrase fragments. Functionally, most bundles analyzed in the present study include research-oriented and text-oriented bundles, like in previous literature. They, however, differ from the bundles identified in the business studies sub-corpus of Hyland (2008a) by including more research-oriented bundles.

ACKNOWLEDGMENTS

I am very thankful to the Fulbright Program and Colorado State University for giving me the opportunity to study in the USA. I am also grateful to all the staff of the English department for their support, especially Marnie Leonard.

I am particularly thankful to my advisor, Dr. Tatiana Nekrasova-Beker for directing my thesis and for her support, guidance throughout my studies in the M.A. TESL/TEFL program. I am also thankful to Dr. Anthony Becker for being my other thesis committee member, and for his support. I also thank Dr. Gerald Delahunty for his support and his guidance. I would also like to extend my gratitude to Dr. Vickie Bajtelsmit for being my outside committee member and for guiding me to the right journals for sampling my texts.

DEDICATION

I dedicate this thesis to my mother, without whom I would not be here. I am forever thankful to her for her sacrifice and dedication to her children.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

MFRAC: Master of Finance Research Articles Corpus

EAP: English for Academic Purposes

ESL: English as a Second Language

EFL: English as a Foreign Language

L1: First Language

L2: Second/Additional Language

CHAPTER ONE: INTRODUCTION

Over the past few decades, there has been a growing body of research into recurrent multi-word units. As a result, it has now been established that recurrent multi-word units are important in language use and learning (Nattinger & Decarrio, 1992; Wray, 2002; Schmitt & Carter, 2004), due, among others, to their pervasiveness in both spoken and written natural language (Hyland, 2008a), and also to the fact that both children and second language learners start using unanalyzed chunks of language before being able to analyze them into their constituent parts (Nattinger & Decarrio, 1992; Schmitt & Carter, 2004).

Multi-word sequences can be defined as frequently-occurring word strings, some of which are fixed. In Wray and Perkins (2000)'s words, a multi-word sequence is: "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). They are intermediary units between lexis and grammar of varying lengths, and they are used to perform discourse functions such as showing relationships between ideas or expressing time based on a language users' pragmatic competence (Nattinger & Decarrio, 1992).

It should, however, be noted that a comprehensive and complete definition is for the moment difficult to reach, the above definition reflecting only characteristics typical of formulaic sequences (Schmitt & Carter, 2004). This is among others due to the diversity in the lengths, functions, and variability versus fixedness of multiword sequences (Schmitt & Carter, 2004). This diversity resulted in different terms to refer to the phenomenon of formulaicity, including idioms, clichés, p-frames, fossilized forms, or formulaic language (Schmitt & Carter, 2004; Wray & Perkins, 2000; Wray, 2002, Biber, Johansson, Leech, Conrad, & Finegan, 1999; Nattinger & Decarrio, 1992).

More recently, mainly following Biber et al. (1999), a growing number of studies have adopted what Biber and Barbieri (2007) considered a complementary approach to the study of formulaic sequences by describing frequent sequences in various discourse types. The sequences which are the focus of this line of investigation, as well as the present thesis, are referred to as lexical bundles. The term "lexical bundles" was first used in Biber et al. (1999), and they are defined as the most frequent word strings occurring in a corpus, and which are familiar to users of a language. They tend not to be complete structural or grammatical units, and not to be idiomatic in meaning (Biber et al., 1999). They include two or more orthographic words, which extend across structural or grammatical units (Biber et al., 1999), and which have certain discourse functions (Hyland, 2008a).

The study of lexical bundles is useful pedagogically because it helps ESL/EFL learners and teachers to focus on sets of frequent and relevant bundles in various discourse types. These are the ones that learners are likely to encounter in their university/academic studies, and thus, help learners and teachers to maximize the return for the learning effort. Research following this line of inquiry examined bundles in conversation (Biber, Conrad & Cortes, 2004) in applied linguistics, business studies, biology and electrical engineering (Hyland, 2008a). They showed differences in the patterns of uses of lexical bundles. For example, Biber et al. (1999) reported that academic prose uses more noun and prepositional phrase fragments than conversation. They also showed that different discourse types use different sets of lexical bundles linked to the typical communicative purposes of these different discourse types, where lexical bundles, among others, are used to structure and organize texts and their meanings (e.g. *on the other hand*) or serve as time markers (e.g. *at the time of*) in texts.

In identifying lexical bundles, studies use different frequency cut-offs and ranges, which means that there is no agreement among researchers about specific frequency cut-offs and distribution criteria. The frequency criteria range from at least 10 occurrences per million words (Biber et al., 1999) to 20 times (Cortes, 2004; Hyland 2008a, 2008b), and 40 times per million words (Biber & Barbieri, 2007). These frequencies are referred to as normalized frequencies, and they indicate how many times particular bundles occur in every one million words. After the frequency-based retrieval of bundles, a number of studies, including Chen and Baker (2010) manually removed overlapping and content-dependent bundles.

In the literature, four-word bundles are the most studied, and after they are identified, they are classified structurally and functionally, using the functional taxonomy developed in Biber et al. (2004), and expanded in later works by Biber and colleagues. The taxonomy consists of the three broad categories of referential, discourse-organizing and stance bundles, which in turn include sub-categories such as clarification or identification. Another set of functional categories is the one developed and used in Hyland (2005, 2008a, 2008b). These include the three categories of research-oriented, text-oriented and participant-oriented bundles, which in turn are sub-divided into sub-categories such as description and transition signals.

Studies following this line of inquiry have revealed differences between spoken and written texts (e.g. Biber et al., 1999; Biber et al., 2004), L1 and L2 written texts (Rica-Peromingo, 2012; Chen & Baker, 2010), and among written academic disciplines (Cortes, 2004; Hyland, 2008a). For example, spoken English exemplified by conversation was found to generally rely on more lexical bundles than written English exemplified by academic prose

(Biber et al., 1999). Structurally, written English tends to use more noun phrase and prepositional phrase fragments than spoken English. Conversely, spoken English tends to use more clausal lexical bundles than written texts (Biber et al., 2004). Functionally, referential expressions and discourse organizers or research-oriented and text-oriented bundles were found to be used in written English, especially academic prose, more than in spoken English (e.g. Biber et al., 1999; Biber et al., 2004). In contrast, spoken English tends to use more participant-oriented or stance bundles than written English (e.g. Biber et al., 1999; Biber et al., 2004).

In academic prose, a number of bundles were found to be among the most frequent bundles across different corpora of academic texts, and these include phrases like *on the other hand* or *in the case of* (Biber et al., 1999). Furthermore, the majority of bundles in academic prose were found to be fragments of noun and prepositional phrases, and most bundles in academic texts are referential expressions and text-organizers or research-oriented and text-oriented bundles (Hyland, 2008a, 2008b; Cortes, 2004).

At the same time, different academic disciplines rely on different sets of lexical bundles in terms of distribution, frequency, forms and functions (Hyland, 2008a, 2008b; Cortes, 2004). For example, Hyland (2008a) found that electrical engineering texts rely on more different lexical bundles than applied linguistics. Structurally, electrical engineering uses slightly more anticipatory *it* structure (e.g. *it is important to, it is possible to*) than applied linguistics. On the other hand, applied linguistics uses more noun phrase + *of*-phrases (e.g. *the nature of the, the sum of the*) than electrical engineering.

The characteristics of lexical bundles in academic disciplines, as very briefly explained above, show the great contribution of previous studies regarding the behaviors of lexical bundles in different academic disciplines. The present study, to some degree, extends earlier studies of lexical bundles by focusing only on one genre, research articles, in Master's-level texts and in a specific sub-discipline, finance, given that no studies on lexical bundles including studies on bundles in business texts have directly investigated the extent to which a sub-discipline uses similar sets of bundles to the discipline in which it is. Specifically, using a corpus of 1,034,587 words, the present study investigates the structural and functional characteristics of frequent bundles in master's-level finance research articles, which are commonly used as reading assignments in the finance and real estate department at Colorado State University.

This is mainly intended to provide master's-level finance students with a restricted list of bundles specific to their sub-field instead of a list related to business studies as a whole, part of which may not be useful to finance students. As such, research articles are suitable because tremendous scholarship is disseminated through them (Hyland, 2008a), and they represent models of writing from which students can learn.

Results suggest that master's-level finance research articles, while having many bundles in common with other academic disciplines (e.g. *on the other hand, in the case of*), including business studies, also rely on a number of lexical bundles that were not identified in earlier studies on lexical bundles in academic disciplines (e.g. *as a proxy for*, *of the underlying asset*), including business studies. Functionally, the bundles identified in the present study to some extent serve different functions from the ones identified in the business studies sub-corpus of Hyland (2008a). For example, the bundles identified in the current study include more research-oriented bundles than the ones identified in the business studies sub-corpus used in Hyland (2008a).

CHAPTER TWO: LITERATURE REVIEW

This section will present an introduction to previous research on formulaic language. Further, main criteria used to identify formulaic language will be discussed before discussing lexical bundles, one of the specific types of formulaic language. The sections on lexical bundles will cover the criteria used in identifying them, along with their structural and functional classification. Finally, studies of the structural and functional characteristics of lexical bundles in different texts will be reviewed. These studies examined lexical bundles in spoken and written English, in L1 and L2 English writing, and in different disciplines.

2.1. Introduction to Formulaic Sequences

Over the past few decades, there has been a growing body of research into recurrent multi-word units. As a result, it has now been established that recurrent multi-word units are important in language use and learning (Nattinger & Decarrio, 1992; Wray, 2002; Schmitt & Carter, 2004), due to, among other reasons, their pervasiveness in both spoken and written natural language (Hyland, 2008a). Also, both children and second language learners start using unanalyzed chunks of language before being able to analyze them into their constituent parts (Nattinger & Decarrio, 1992; Schmitt & Carter, 2004). Specifically, prefabricated expressions are important for both first and second language acquirers and learners in their early stages of learning to converse with proficient speakers of the language. In addition, language learning, according to one of the more current theories in L2 acquisition, proceeds from learning items of the language to learning the system: the rules of the language (Matthews, Theakston & Tomasello, 2005; Tomasello, 2003; Ortega, 2009).

Multi-word sequences can be defined as frequently-occurring word strings, some of which are fixed, others not. In Wray and Perkins's (2000) words, a multi-word sequence is: "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). They are intermediary units of varying lengths between lexis and grammar, and they are used to perform discourse functions such as showing relationships between ideas or expressing time based on a language users' pragmatic competence (Nattinger & Decarrio, 1992). They are learned first as fixed expressions before being analyzed into their components and being replaced by certain words and phrases in specific slots with semantically-appropriate and constrained language (Schmitt & Carter, 2004). Formulaic sequences have also been found to have processing advantages over creatively generated, equivalent sequences by being retrieved more quickly in receptive and productive uses (Wray, 2002; Nattinger & Decarrico, 1992).

It should, however, be noted that a comprehensive and complete definition is for the moment difficult to reach, the above definition reflecting only characteristics typical of formulaic sequences (Schmitt & Carter, 2004). This is, among other reasons, due to the diversity in the lengths, functions, and variability versus fixedness of multiword sequences (Schmitt & Carter, 2004). This diversity resulted in different terms to refer to the phenomenon of formulaicity. Multi-word units are variously referred to, among other reasons, as idioms, clichés, p-frames, fossilized forms, conventionalized forms, holophrases, ready-made utterances, prefabricated routines and patterns, routinized formulas, formulaic language, amalgams, unanalyzed chunks of speech (Schmitt & Carter, 2004; Wray & Perkins, 2000; Wray, 2002, Biber et al., 1999; Nattinger & Decarrio, 1992). These are identified using different criteria which, to a great extent, depend on the investigator’s research purposes. It is then important that a researcher thoughtfully and appropriately use a definition that effectively helps to identify multi-word units which are suitable for their study’s purposes (Wray, 2008).

2.2. Identifying Formulaic Sequences

Research into multi-word units tends to point to fixedness, idiomaticity, frequency, length of sequence, structural/syntactic completeness, recognition based on native speakers’ intuition, semantics or pragmatics (Conrad & Biber, 2004). Studies give differential primacy to the above characteristics based on the research focus.

Wray (2008) grouped or categorized methods or ways used to identify multi-word units into the ones based on intuition, phonological features, idiosyncrasy, form, spelling and frequency. Studies that use phonological indicators, Wray (2008) explained, take phonetic reductions as criteria for formulaicity. High-frequency sequences tend to be pronounced using phonological reductions more than medium- and low-frequency ones without affecting their communicative effectiveness because the forms of very frequent sequences are predictable by hearers. Even low-frequency sequences such as idioms have been found to be pronounced with fewer pauses and shorter content words than non-idioms. She also referred to a study taking liaison in French -- which is the pronunciation of a word-final silent consonant when it precedes a word-initial vowel -- as an indicator of formulaicity, with more common instances of liaison being more likely to continue to be used by speakers of French.

Formal criteria for identifying multi-word units include considering as formulaic any sequence with two or more words. The problem, Wray (2008) argued, is that it becomes difficult to deal with issues related to functional similarities. That is, expressions such as *thank you* or *good bye* would be considered as formulaic whereas words such as *thanks* and *hello* would not, even though the latter two words serve similar functions to *thank you* or *good bye*. Furthermore, she explained, a number of two-word expressions vary in the ways they are written; they are either written as two-word expressions or one-word expressions. For example, *no one* and *a lot* can be written as *noone* and *alot*. Another formal criterion for determining formulaicity consists of checking that changing a member of an expression with a synonym results in a change of the expression meaning, function or idiomaticity. A problem with this criterion, Wray (2008) argued, is that idioms that are considered fixed have been found to vary in their form.

Another criterion used in identifying recurrent multi-word units consists of looking for idiosyncrasies in the linguistic production of L1 and L2 language learners or preschoolers. Specifically, researchers determine the formulaicity of expressions by determining whether the expression is beyond the developmental level of the learner or child. Expressions that are unusually long and complex with regard to a learner's developmental level are considered as formulaic. This methodology involves having enough information about the individual's knowledge of the language in order to reliably determine whether the expressions can be produced creatively or generatively based on their level. With preschoolers, during first language acquisition, words that tend to be used together, not independently of each other, are considered as formulaic. Peters (1983), Wray (2008) and Myles (2005) explained that this method of identifying recurrent multi-word sequences is based on the assumption that formulaic sequences are only temporarily frozen, and that the members of the unit are ultimately used independently of each other by the child or outside the sequences.

Intuition is yet another means for identifying recurrent multi-word units. Researchers either use their own intuition about the formulaicity of an expression (Wray, 2008) or use other informants, which include judges, who are native speakers or are near natives. To make an intuition-based identification more reliable, it is often combined with other criteria, which may include fixedness, non-compositionality and syntactic sophistication. Although intuition may not always be a reliable means of identifying formulaicity, it is used in literally all studies, even those using frequency in corpora as a criterion. In fact, Wray (2008) added, outside corpus-based identification, it seems difficult to dispense with intuition in identifying multi-word units.

A frequency-based criterion for identification consists of using computer programs to search language corpora for recurrent multi-word units. The speedy and seemingly clean character of using computer programs to analyze corpora makes this method attractive, even though infrequent formulaic expressions may be left out (Wray, 2008). At the same time, patterns found in frequent units can subsequently be detected in less frequent material in order to determine formulaicity. Another problem with a frequency-based method is that depending on the criteria for the length and number of occurrences, and the variation in the sequences, researchers can obtain widely different results (Wray, 2008). That is why Wray (2008) pointed to the importance of not using frequency as an easy means for identifying multi-word units, but for some justified purpose or motivation.

More recently, mainly following Biber et al. (1999), a growing number of studies have adopted what Biber and Barbieri (2007) considered a complementary approach to the study of formulaic sequences by describing frequent sequences in various discourse types. The sequences which are targeted in this body of research are referred to as lexical bundles.

2.3. Lexical Bundles

The term “lexical bundles” was first used in Biber et al. (1999), and they are defined as the most frequent word strings occurring in a corpus, and which are familiar to users of a language. They tend not to be complete structural or grammatical units, and not to be idiomatic in meaning (Biber et al., 1999). They include two or more orthographic words, which extend across structural or grammatical units (Biber et al., 1999), and which have certain discourse functions (Hyland, 2008a).

Lexical bundles are different from other multi-word sequences by being more frequent, by having transparent meanings and being structurally incomplete (Biber and Barbieri, 2007). Further, the studies that examined them, among others, focused on their structures and functions in different text/discourse types of authentic language uses. Identifying lexical bundles mainly based on frequency and distribution has the advantage of using a clear and rather straightforward methodology. Also, this line of investigation, like some others on formulaic language, is useful pedagogically by helping ESL/EFL learners and teachers to focus on sets of frequent and relevant bundles in various discourse types. These are the ones that learners are likely to encounter in their university/academic studies, for example, and thus, help learners and teachers to maximize the return for the learning effort. Frequency-based identification/selection is also useful because native speakers’ intuitions – often used in identifying other types of formulaic sequences -- about the frequencies of particular language features are often inaccurate (Biber & Conrad, 2001).

The studies following this line of inquiry (e.g. Biber et al., 1999; Biber et al., 2004; Biber & Barbieri, 2007) revealed that lexical bundles are consistently functional in both spoken and written English. That is why Biber and Barbieri (2007) argued that the high frequency of lexical bundles is indicative of their formulaicity.

Research following this line of inquiry examined bundles in conversation (Biber et al., 2004) in applied linguistics, business studies, biology and electrical engineering (Hyland, 2008a). They showed differences in the patterns of uses of lexical bundles. For example, Biber et al. (1999) reported that academic prose uses more noun and prepositional phrase fragments than conversation. They also showed that different discourse types use different sets of lexical bundles linked to the typical communicative purposes of these different discourse types, where lexical bundles, among other functions, are used to structure and organize texts and their meanings (e.g. *on the other hand*) or serve as time markers (e.g. *at the time of*) in texts.

2.3.1. Identifying lexical bundles

Studies investigating lexical bundles use corpora of different sizes. Those studying more broad or general language tend to use larger corpora than those studying specialized language. For example, comparing academic prose and conversation, Biber et al. (1999) used a 5,331,800-word sub-corpus for academic prose and 6,410,300-word corpus for conversation. On the other hand, Cortes (2004), studying bundles in biology, history, and research articles texts, used 966,187 and 1,026,344 words for history and biology research articles, respectively.

In identifying lexical bundles, studies use different frequency cut-offs and ranges. In other words, there is no agreement among researchers about specific frequency cut-offs and distribution criteria. The frequency criteria range from at least 10 occurrences per million words (Biber et al., 1999) to between 20 times (Cortes, 2004; Hyland 2008a, 2008b) and 40 times per million words (Biber & Barbieri, 2007). These frequencies are referred to as normalized frequencies, and they indicate how many times particular bundles occur in every one million words. Normalizing frequencies for identifying clusters is important because the corpora and sub-corpora studied and compared are of different sizes. For example, Biber and Barbieri (2007) studied lexical bundles across different sub-corpora ranging from 1,248,811 words for language teaching to 760,619 words and even 151, 500 words. The criteria for distribution range in occurrence from 5 different texts (Biber et al., 1999; Conrad & Cortes, 2004) to 10% of texts (Hyland, 2008a, 2008b).

It should be pointed out that intuition is, to some extent, also used in selecting lexical bundles. In other words, after the automated, frequency-driven, retrieval of lexical bundles, researchers manually remove bundles that, intuitively, do not seem formulaic to them. These include mostly content words considered to be text- or context-dependent (Hyland, 2012). These are often strings that are closely related to the topics of the texts that make up the corpora under consideration. But ultimately, some degree of intuition is used in deciding what is content-dependent and what is not. For example, Chen and Baker (2010) removed clusters such as *financial and non financial* and *the Second World War*. Such intuition-based selections may be criticized as being subjective (Hyland, 2012). That is why some researchers, such as Simpson-Vlach & Ellis (2010) also used corpus statistics, namely the MI score. The MI score calculates the strength of the collocations or the associations of strings, and it is employed because it seems to indicate phrasal coherence, which, seemingly, correspond to distinctive functions and meanings. The problem, Hyland (2012) argued, is that the MI score, being conceived for 2-word collocations may not be reliable for longer strings. Additionally, Biber (2009) pointed out that it tends to select low-frequency items, and does not take into account the orders of the words forming the collocations.

Other strings of words often removed manually include overlapping clusters that inflate the numbers of items (Hyland, 2012). These include, for example, two three-word bundles in a single four-word bundle, or two four-word bundles in a single five words (Biber, Johansson, Leech, Conrad, & Finegan, 1999). In such cases, the overlapping bundles are combined into one (Chen & Baker, 2010).

The frequency and distribution criteria discussed are mostly used for the identification of four-word bundles, which have received more attention than three-, five- and six-word bundles in terms of the study of their structural and functional characteristics. For example, while Biber et. al. (1999) used a frequency cut-off of at least 10 times per million words for four-word bundles, they used a lower cut-off of 5 times per million words for five- and six-word bundles. This is because the longer a string of words, the less frequent it is in authentic language uses. For example, in their sub-corpora of British and America conversation and academic prose totaling close to 12 million words, Biber et. al. (1999) found that three-word bundles occur about ten times more frequently than four-word bundles. Likewise, four-word bundles occur almost 10 times more frequently than five-word bundles.

Specifically, four-word bundles have been more studied than other strings because five and six-word bundles are much less frequent, and often include shorter strings, and three-word bundles are much more frequent and more collocational in nature (Biber et. al., 1999), in addition to not being interesting for structural and functional analysis (Hyland, 2012). Additionally, four-word lexical bundles "offer a wider variety of structures and functions to analyze" (Hyland, 2012).

2.3.2. Classifying/Categorizing lexical bundles

Nattinger and Decarrico (1992) preceded Biber, Johansson, Leech, Conrad, & Finegan (1999) in classifying formulaic sequences for pedagogical and description purposes. Formally/structurally, they categorized sequences or "lexical phrases" into "polywords", "institutionalized expressions", "phrasal constraints" and "sentence builders." (pp. 38-45). They used four criteria to differentiate each of the four categories from the three others. The categories include length and grammatical status, canonical (those with the typical English sentence structure) versus non-canonical, variable versus fixed and continuous versus non-continuous. They stressed the necessity of thinking in terms of continuum in applying the criteria and distinguishing the four categories, rather than in terms of separate categories. For example, polywords, according to Nattinger and Decarrico (1992), are in terms of length and grammatical status short phrases used like individual lexical words. They can be canonical or not. They are mostly invariable and continuous, and they include expressions such as *in a nutshell* or *so long*. Similarly, phrasal constraints are said to include short and medium-length sequences, to be canonical and non-canonical, of including a variety of lexical and phrasal categories such as nouns, verbs, noun and verb phrases, and of being mostly continuous. Examples of phrasal constraints include *to wrap this up, a year ago,* or *let me start by.*

Functionally, Nattinger and Decarrico (1992) used 3 categories: social interactions, necessary topics and discourse devices. Social interactions are sub-categorized into conversational purpose and conversational maintenance. Conversational maintenance, for example, is related to how conversation begins, continues and ends. Sequences in this sub-category include those for summoning such as *pardon/excuse me*. Necessary topics refer to topics used in daily conversations, and they include expressions such as *my name is,* *how much/big* or *I like/enjoy*. The last broad category includes discourse devices with sub-categories such as logical connectors (e.g. *as a result, nevertheless*), evaluators (e.g. *as far as I know, there is no doubt*).

However, Nattinger and Decarrico (1992) did not use frequency as a criterion to identify their sequences. Specifically, they identified the sequences that are perceptually salient, recognized based on native speaker's intuition and which serve discourse functions based on pragmatic competence. The problem is that one does not know how frequent the studied sequences are in normal language use. Furthermore, structurally, while lexical bundles studies focus on grammatical/syntactic structures (phrases and clauses), Nattinger and Decarrico (1992) also examined lengths, variability, and the interrupted versus uninterrupted characteristics of sequences.

A greater and more direct influence on the line of inquiry used in the analyses of lexical bundles in the past comes from Altenberg (1998). He identified sequences -- which he referred to as "recurrent word-combination" -- of varying lengths and frequencies in the London-Lund Corpus of Spoken English, a corpus of nearly half a million running words. He found 68,000 different sequences, making up 201,000 sequences in total. But for practical reasons, he only focused on strings that are at least three-word long, and which occurred at least 10 times in the corpus. His resulting sequences were 6,692 tokens and 470 different sequences.

He categorized them into three broad structural categories, each with sub-categories, and studied the functions that these categories served in spoken English. His structural categories included full clauses, which comprise independent and dependent clauses, clause constituents, which comprise multiple clause constituents and single clause constituents, and incomplete phrases. A set of discourse functions is associated with each category.

Full clauses serve functions such as responses (e.g. *thanks very much*), epistemic tags (e.g. *I' m not sure*) with independent clauses, or express comments (e.g. *as it were*), indirect conditions (e.g. *if I may*) or mark apposition (e.g. *that is to say*) with dependent clauses (pp. 104-109). Multiple clause constituents (e.g. *and you know, and then I*) -- in the clause constituents category -- act mainly as frames, onsets or stems depending on their positions in the clause. Frames are language in pre-subject position; onsets are thematic elements including subjects and preceding the finite verb. Stems are formed with any preceding thematic elements, a subject and a verb. The three form 97% of recurrent sequences used to express old/background information preceding new information. Single clause constituents – the other sub-category of clause constituents – are usually complete phrases including vagueness tags (e.g. *or something like that, that sort of thing*), qualifying expressions (e.g. *more or less*) and connectors (e.g. *first of all, in other words*). Finally, incomplete phrases are various kinds of strings generally lacking the head or the postmodifier of the phrase. Many of them have slots in which sets of lexically and grammatically-related words occur, some of which have become more or less fixed

However, more recent lexical bundles studies differ from earlier studies by studying almost exclusively structurally incomplete units -- mostly composed of the fragments of two adjacent phrases or clauses -- and using, to some extent, different categorizing terms or descriptors in classifying bundles functionally. The functional categories used in classifying lexical bundles mainly emerge or come from Biber et al. (1999) and Biber et al. (2004), among others. The three categories are stance bundles, discourse organizing and referential bundles. Stance bundles, according to Conrad and Biber (2004), are used to express assessments and attitudes providing the frame for interpreting the idea that follows the stance expression bundle.

Stance expression bundles or clusters include *I don't know if* and *it is necessary to*. Stance expressions were further broken down into epistemic stance, attitudinal stance (desire, directives, intention/prediction and ability).

Discourse organizing lexical bundles include two sub-categories, which are topic introduction and focus lexical bundles, and topic elaboration and clarification lexical bundles. The introduction and focus lexical bundles signal that a topic is being introduced or focused on. In classroom teaching, for example, topic introduction lexical bundles include *what I want to do is*, *if you look at*, etc. In textbooks, the very few topic introduction bundles include *in this chapter we*. Topic elaboration bundles are used to make a previously stated information more clear, ask for clarification about it, or add information to it. In conversation, clarification bundles include *what do you mean*, *nothing to do with* (Biber & Barbieri, 2004). In academic prose, clarification bundles include *on the other hand, as well as the,* etc.

Referential bundles, Conrad and Biber (2004) explain, are used to refer to entities, physical and abstract, or to the textual context. They are divided into four sub-categories (Biber et al., 1999; Conrad & Biber, 2004). Referential identification/focus bundles identify physical and abstract entities in a way that signals them as noteworthy (e.g. *one of the most*). Imprecision bundles communicate the imprecision of a previous sentence, paragraph or larger text (e.g. *or something like that*). Referential specification bundles bring focus to particular attributes of entities. These are used to specify such attributes as quantities, tangible and intangible attributes of entities (e.g. *per cent of the*). Lexical bundles in the fourth sub-category of referential lexical bundle are used as time, place or text references (e.g. *the end of the*). Bundles used to serve special conversation functions are those used for politeness routines, simple inquiry and as fragments of reporting clauses.

However, even though many subsequent studies have used functional categories developed from the three stance, text-organizing, and referential categories, they, to some extent, often extend or modify those to include other categories or sub-categories. This may be that the structural and functional classification being inductive and exploratory, researchers had to create subcategories that properly describe the bundles that they identified. For example, in the referential category, Cortes (2004) created a quantifying sub-category, which introduces quantities and amounts. Similarly, Hyland (2008a, 2008b) used different categories to classify his bundles functionally. He categorized them into research-oriented, text-oriented and participant-oriented. Research-oriented bundles, according to Hyland (2008a, 2008b), help writers in structuring real-world activities and experiences. Research-oriented bundles were further classified into location (e.g. *at the beginning of*), procedure (e.g. *the operation of the),* quantification (e.g. *the magnitude of the*), description (e.g. *the structure of the*) and topic (e.g. *in the Hong Kong*) (p. 13). Text-oriented clusters are used to organize a text and its meaning into an argument or a message. They include transition signals (e.g. *on the other hand*), resultative signals (e.g. *as a result of*), structuring signals (e.g. *in the present study*), and framing signals (e.g. *in the case of*). Finally, participant oriented clusters focus on writers and readers by conveying their attitudes, for example. They include stance features (e.g. *it is possible that*), engagement features (e.g. *It should be noted*).

The study of structural and functional categories across different text types has revealed that different registers or text types use different sets of structural and functional categories reflecting different communicative purposes. In the following sections of the current review, I will discuss the structural and functional characteristics of bundles used in spoken and written English first, then in L1 and L2 writing, and in different academic disciplines.

2.3.3. Lexical bundles in spoken and written English

My discussion of lexical bundles in spoken and written English will focus on conversation, classroom teaching, classroom management, office hours, study groups and service encounters for oral registers (Biber et al., 1999; Biber et al., 2004; Biber & Barbieri, 2007). For written registers, I will focus on academic prose, textbooks, course management and institutional writing (Biber et al., 1999; Biber et al., 2004; Biber & Barbieri, 2007). Specifically, I will compare the lexical bundles used in spoken registers to those used in written registers. I will also discuss differences among the spoken registers, as well as the written registers.

The first difference between spoken and written registers is that spoken registers tend to use more bundles than written registers both in terms of different bundles and quantity. For example, Biber et al. (1999) found 5000 occurrences of lexical bundles per million words in academic prose and 8500 instances in conversation. Similarly, Biber et al. (2004) found that classroom teaching, a spoken register, uses more bundles, both in terms of quantity and different types, than textbooks and academic prose. This may be due to the spontaneous nature of oral interaction where, under time pressure, speakers choose from recurrent and relatively limited and expected/predictable sets of prefabricated expressions to convey their ideas. However, while this may usually be the case with regard to the quantity of bundles, it is not always the case regarding the use of various or different types of bundles. For example, Biber and Barbieri (2007) found that course management and institutional writing, both written registers, use more different types of bundles than classroom teaching in their corpus, with over 120 different types of bundles used in course management, and fewer than 90 in classroom teaching, for example.

Research into the structural patterns of four-word lexical bundles in spoken English revealed that bundles in spoken English registers tend to be composed of parts or fragments of questions and declarative clauses or verb phrases more than in academic written registers (Biber et al., 1999; Conrad and Biber, 2004). In conversation, Biber et al. (1999) grouped bundles into 14 structural categories. These are fragments of clauses comprising, for example, personal pronouns + lexical verb phrases (e.g. *I don't know what*), pronoun or noun phrase + copula be (e.g. It was in the), questions (e.g. *Can I have a*) (Biber et al., 1999).

In academic prose and textbooks the majority of clusters are parts of noun phrases and prepositional phrases, but also anticipatory *it* + verb or adjective phrase + complement clause (Biber et al., 1999; Biber, Conrad & Cortes, 2004). Biber et al. (1999) grouped bundles in academic prose into 12 major structural categories, which include parts of noun phrases containing post-modifiers (e.g. *the nature of the*), preposition + noun phrase (e.g. *as a result of*) (Biber et al., 1999). But Biber and Barbieri (2007) found that classroom teaching relies both on clusters incorporating declarative and interrogative clause fragments prevalent in conversation (e.g. *you don't have to, I want you to*), and fragments of noun and prepositional phrases (e.g. *one of the things, the end of the*) prevalent in academic prose and textbooks.

Functionally, spoken and written registers tend to rely on relatively different categories. Spoken registers tend to use more stance (e.g. *I don't know how, I want you to*) bundles than written registers. In contrast, written registers, that is, textbooks, academic prose, course management and institutional writing tend to use more referential expressions than spoken registers. For example, in conversation, stance expressions are used the most, followed by discourse organizers (e.g. *what do you think*) (Conrad & Biber, 2004). Conversation uses far fewer referential expressions than stance bundles. In stance expression bundles, personal epistemic stance expressions are the most used (e.g. *I think it was*). The patterns of uses of clusters in conversation reflect the priority given to conveying personal thoughts and attitudes, the concern for not imposing on people and for politeness. You have relatively similar patterns in other spoken registers. For example, all of the five university spoken registers investigated in Biber and Barbieri (2007) (service encounters, classroom teaching, class management, office hours and study groups) were found to use more stance bundles than discourse organizing and referential bundles. Classroom teaching relies on referential bundles (e.g. *that's one of the, and things like that, how many of you*) more than all the other university spoken genres, followed by class management.

However, even though the use of stance bundles is a general characteristic of spoken registers, different spoken registers rely on different sub-categories of stance bundles. For example, study groups rely on epistemic stance lexical bundles more than all the other university spoken genres. These stance bundles (e.g. *I don't know what, I don't think that*) are usually used in study groups situations to make claims tentatively, instead of making assertions (Biber & Barbieri, 2007). Likewise, office hours rely more on ability stance bundles (e.g. *I don't know if, I don't think so*) than any other spoken university register. These are used in office hour situations to indirectly give orders to students to do an assigned task, by emphasizing that students have the ability to perform the task. Also, classroom teaching, which uses the most discourse organizing bundles in spoken registers, uses them mainly to introduce or focus on a topic, or elaborate or clarify a topic (e.g. *I want to talk about, if we look at*). Classroom teaching relies on relatively many stance and referential bundles, according to Biber et al. (2004), because it involves at the same time involved spoken discourse and informational written discourse.

Of the written registers Biber and Barbieri (2007) found that only course management does not rely on more referential bundles (e.g. *in the case of*) than discourse-organizing and stance bundles. For example, in academic prose, most of referential bundles are used for attributes specification (Conrad & Biber, 2004). The few stance clusters used in academic prose are three common four-word bundles, which include *the fact that the* and *it is necessary to,* all of which are impersonal stance bundles. In addition, contrary to epistemic stance bundles used in conversation, which usually show uncertainty, the most common epistemic stance cluster in academic prose, *the fact that the,* expresses certainty (Conrad & Biber, 2004). In academic prose *the fact that the* is often used to present a concept as accepted information or established. Of the few discourse organizing bundles used in academic prose, *on the other hand* is the most common, and it is used to show contrast explicitly. The patterns of uses of lexical bundles in academic prose reflect the emphasis on conveying information (Conrad & Biber, 2004).

Written registers, to some extent, also differ in their uses of the different functional categories of bundles. Institutional writing relies on referential bundles much more than the other written genres or registers. For example, Biber and Barbieri (2007) found that institutional writing uses over 60 different referential bundles while course management, the closest to it, uses fewer than 40 different referential lexical bundles. Textbooks rely on even fewer, with fewer than 20 referential bundles. In institutional writing, place references are the most common in the referential bundles because, according to Biber and Barbieri (2007), people need to refer to places on campus. These bundles include *in the college of, from the office of.*

Course management, which is “Written course management includes 10 syllabi ‘text’ files (196 syllabi totaling ca. 34,000 words) and 11 course assignment ‘text’ files (162 individual assignments totaling ca. 18,500 words)” (p.264) relies on stance bundles the most with over 70 different stance bundles (Biber & Barbieri, 2007). In fact, unlike the other three written registers, course management relies more on stance bundles than referential bundles. It also uses more discourse organizing bundles than all the other written registers. Textbooks use fewer different types of stance, discourse-organizing and referential bundles than course management and institutional writing, possibly because textbooks use far fewer different bundles than the two overall. However, academic prose uses even fewer stance, discourse-organizing and referential bundles. The referential expressions that textbooks rely on mainly refer to tangible and intangible framing attributes. These bundles include *the size of the, the nature of the*.

However, as Biber and Barbieri (2007) admitted, one should be cautious in considering these findings, given that office hours, class management, which “occurs at the beginning and end of class sessions, to discuss course requirements, expectations, and past student performance,” (p. 264) and course management, only 50,400-words, 39,255-words and 52,410-words, respectively, are small for investigating lexical bundles. More generally, it seems that larger sub-corpora, at least in Biber and Barbieri (2007) yielded fewer different types of bundles than smaller sub-corpora. More research seems to be needed to see how studying lexical bundles across corpora and sub-corpora of similar sizes compare to studying lexical bundles across corpora and sub-corpora of significantly different sizes. In the specific case of lexical bundles analysis in different registers, studying the same registers as Biber, Conrad and Cortes (2004) or Biber and Barbieri (2007) using sub-corpora of similar sizes may produce different findings.

Beside the study and comparison of bundles across spoken and written texts, a fair amount of studies compared and contrasted the lexical bundles frequently used in L1 and L2 English academic writing in terms of their frequencies, varieties, as well as their structural and functional characteristics. This is because of the pedagogical implications of findings from such research (Perez-Llantada, 2014). In the next section, I shall review some of the research into the use of lexical bundles by English first and second language writing.

2.3.4. Lexical Bundles in L1 and L2 English Writing

There has been a growing number of studies into L1 and L2 English academic writing, most of which seem to have focused on university-level students' and expert writings. This is partly due to the recognition that the analysis and comparison of L1 and L2 English written production is useful to identify the overuse and underuse of particular bundles in non-natives' writings, and ultimately apply those findings to the teaching of ESL/EFL learners (Chen & Baker, 2010). Some studies investigated and compared the uses of four-word lexical bundles in novice L1 and L2 English writings, novice and expert writings or expert L1 and L2 English writings (Cortes, 2004; Chen & Baker, 2010). Others focused on specific or limited numbers of lexical bundles, studying and comparing their uses across novice and expert L1 and L2 English writings (Rica-Peromingo, 2012; Hassan et al., 2009). For example, Cortes (2004) compared the uses of lexical bundles by expert writers in history and biology to those of students in these disciplines. Similarly, Chen and Baker (2010) compared the uses of lexical bundles in a corpus of Chinese L2 English essay writers, L1 English essay writers and expert/published academic writing from academic research articles and textbooks extracts. Jalali, Rasekh and Rizi (2009) focused on a sub-type of extraposed structure involving anticipatory *it +is* followed by predicative adjective (e.g. important) + infinitival *to* or conjunctive *that.*

Findings pointed to the fact that expert and native writers tend to use more different bundles than non-natives, with expert writers using the most different types of bundles. For example, Adel and Erman (2012) found 120 different bundles in the corpus of native writings and 60 in the corpus of non-native writings (p. 85). Similar findings were reported in Cortes (2004), even though Hyland (2008a) found that Master's theses contained more different lexical bundles than the more proficient PhD theses and published research articles, arguing that less proficient writers tend to rely on more prefabricated expressions. Chen and Baker (2010) attributed the discrepancies in the findings to the fact that Hyland (2008a) did not remove overlapping and context-dependent bundles while they did. They also cautioned against comparing findings across studies using corpora of different sizes, made up of different text types.

It should, however, be noted that there are consistent differences in the uses of lexical bundles by either novice L1 and L2 writers, novice and expert writers, or expert L1 and L2 writers. For example, both novice L1 and L2 English writers tend to overuse certain lexical bundles and underuse bundles frequently used by expert writers (Cortes, 2004). Cortes (2004) reported that students in both history and biology seldom or never used many of the bundles used by published authors, 29 out of the 54 bundles identified in the published writing in the case of history students. Even the referential, time-marking, bundles (e.g. *the beginning of the, the end of the*) used by students with higher frequency were found to be different from the ones typical of published history writing (e.g. *at the turn of, in the course of*).

Structurally, also, Chen and Baker (2010) found that student/novice writers tended to differ from expert writers in their uses of lexical bundles, with expert writers using more noun phrase-based and preposition-based phrases than novice L1 and L2 English writers. In contrast, they found that novice writers, both L1 and L2 English writers, used more verb phrase-based bundles than expert writers. For example, noun phrase-based and preposition-based phrases make up 68.5% of bundles in the published, expert writing corpus used in their study, 44.2% in the L1 English corpus and 57.5% in the corpus of L2 English writing. From the percentages, one can see that Chinese L2 English used more noun phrase-based and prepositional phrase-based bundles than L1 English students, being closer to expert writers in this respect. However, further analyzing the noun phrase-based bundles, they found that novice L1 English writers used most of the noun phrase-based bundles not followed by *of* while L2 English writers used none of this kind of bundle.

Functionally, Chen and Baker (2010) found that expert writers used more referential bundles, 60% of types of bundles in expert writing, than novice L1 and L2 English writers, 37% and 41% respectively. On the other hand, novice writers' corpora contained more stance bundles than expert writers, with stance bundles representing 42% of the types of bundles used in L2 English corpus and 37% in the L1 English corpus. The significant use of referential bundles appears to be indicative of mature academic writing. In the corpus of published/expert academic writings, a type of referential bundle, a type of quantifying bundle (degree/extent modifiers), for example *the extent to which*, is also found in novice L1 English writing while it was not found in novice L2 English writing. In that novice native speakers of English are closer to expert writers in their uses of lexical bundles. Chen and Baker (2010) also found that L1 English essay writers exhibited similar uses of lexical bundles by having more control of hedging devices than L2 English essay. L2 English writers tendency to use significantly fewer hedging devices, according to Chen and Baker (2010), reflects the tendency in immature, second language writing to overgeneralize and be categorical in expressing ideas.

Even in expert writing, differences have been reported across L1 and L2 English writing. In a 5.7 million-word corpus, of L1 English, L2 English and L1 Spanish research articles by expert/mature writing, Perez-Llantada (2014) found that L2 English and L1 Spanish writers used more different lexical bundles than L1 English writers. For example, while L2 English expert writers used 77 different bundles, L1 English expert writers used 54 different bundles. But he found that 36 lexical bundles were used as core or overlapping bundles in the three corpora, even though L2 English and L1 Spanish expert writers were found to use more noun-phrase and prepositional-based lexical bundles than L1 English research articles. Conversely, L1 English writers were reported to use more verb-based lexical bundles than the other two groups. In the verb phrase-based bundles used by L1 English writers, anticipatory *it* followed by verb *be* + adjective + clause fragment is the most frequent structure. This structure is characteristic of academic texts where such a structure is used for hedging (e.g. *it is possible to*) and for emphasis (e.g. *it is clear that*), for example.

Functionally, L1 Spanish and L2 English were found to use more referential bundles than L1 English. In contrast, L1 English were reported to use more stance bundles than the other two groups. Overall, Perez-Llantada (2014) concluded that even among expert L2 writers only few ever acquire and use all the range of lexical bundles used by native expert writers, even though this difference is not significant. He argued that these small differences are to some extent due to transfer from their first language.

So far, I have discussed lexical bundles in the spoken and written English modes, as well as in English first and second language. In order to provide a fuller description of research into lexical bundles in English, I shall discuss, in the next section, some of the research into the uses of lexical bundles in different academic English disciplines.

2.3.5. Lexical bundles in different academic disciplines

Research into the lexical bundles used in academic disciplines focuses on various disciplines, including history and biology (Cortes, 2004), electrical engineering, business, applied linguistics and biology (Hyland, 2008a). The differences in the use of text types across studies investigating lexical bundles in different disciplines reflect differences in purposes. For instance, Cortes (2004) chose research articles because professors in the field considered by Cortes considered research articles as models of good writing from which students were to learn. Additionally, she was comparing students' writings to expert writing.

The results in such studies revealed differences, but also similarities, in the patterns of uses of bundles across different disciplines. For example, in terms of structure, most of the bundles identified across studies relate to academic prose. In other words, most of, or at least, the bundles identified in these studies are fragments of noun and prepositional phrases. For instance, over 60% of bundles in biology and history research articles are noun phrase with "*of*" phrase fragments (e.g. *a function of the, both sides of the*), noun phrases with post nominal clause fragments (e.g. *the degree to which, the ways in which*), prepositional phrases with embedded "*of*" phrase (e.g. *as a consequence, at the time of*) (Cortes, 2004).

However, some disciplines use more bundles than others. For example, Hyland (2008a) found that electrical engineering, with 213 different bundles, uses more bundles than the other three disciplines, followed by business studies with 144 bundles. Hyland (2008a) speculated that this, among others, that "it could be a consequence of the relatively abstract and graphical nature of technical communication…the dependence of Engineering…on visual and numeric representation" (pp. 9-10).

Structurally, also, academic English shows disciplinary variation. These pertain, among others, to the distribution of the bundles of different structures across different disciplines. For example, Hyland (2008a) found that business studies use noun phrase + of fragments (e.g. *the end of the, the price of the*) the most with 28.5% of all bundles in business studies being parts of this structure. In comparison, 22.9 of all clusters in applied linguistics are parts of noun phrase + of structure, 22.3% in electrical engineering, and 23.7% in biology (p. 10). In contrast, business studies use four-word lexical bundles incorporating anticipatory it structure the least with only 4.5% of all bundles incorporating this structure. Electrical engineering uses this structure the most (e.g. *it can be observed, it was found that*). In comparison, only 6.3% of bundles in biology incorporate anticipatory *it* structure, 5.6% in applied linguistics (p. 10). But generally, anticipatory it tends to be used less than other structures both in novice and expert writing in different disciplines. For instance, in Hyland (2008a), this structure makes up only 2.5% of all bundles used in all the four disciplines (i.e. biology, electrical engineering, applied linguistics, business studies). But, even though clausal in structure, extraposed structures such as those involving anticipatory *it* + *Vbe* + *adjective* + *clause fragment* (e.g. *it is important to, it is clear that*) are characteristic of written English academic discourse (Biber et al., 1999).

Functionally, academic disciplines have in common their reliance on large proportions of referential bundles, which is characteristic of academic prose, even though different disciplines tend to rely on varying amounts proportions of referential bundles. But generally, referential and text-organizing bundles tend to be used more than stance bundles in academic prose. Like for structure, lexical bundles have different patterns of uses functionally across different disciplines. Hyland (2008a) found that soft sciences (represented by business and applied linguistics) use different categories from hard sciences (represented by biology and electrical engineering) (p. 14). While soft sciences use more text-oriented bundles than any other functional categories, hard sciences use more research-oriented bundles than text- and participant-oriented bundles. 49.5% of bundles in applied linguistics are text-oriented, 31.2% of bundles are research-oriented, and 18.6% are participant-oriented. In business studies, 48.4% of bundles are text-oriented, 36% research-oriented, and 16.6% are participant-oriented (p. 14). In biology, 48.1% of bundles are research-oriented, 43.5% are text-oriented, and 8.45 are participant-oriented. Research-oriented bundles make up 49.4% of bundles in electrical engineering, text-oriented bundles make up 40.4% of bundles, and 9.2% are participant-oriented. Hyland (2008a) argued that these patterns of uses are due to the tendency in hard sciences to communicate real-world and laboratory-focused sense, giving specific descriptions of research objects and contexts (e.g. *the structure of the, the base of the*). On the other hand, arguments in soft sciences are more discursive and evaluative with more interpretation in researchers' attempts to persuade.

Findings in Cortes (2004), to some extent, echo those in Hyland (2008a). For example, in history research articles, she found that close to half of bundles are text-organizers. In biology, there are approximately equal numbers of referential and text-organizing bundles. But in her biology corpus, unlike in history research articles, which have no stance bundles, there are relatively many stance bundles, albeit far fewer than referential and text-organizing bundles. In both history and biology, most of the many referential bundles are quantifying bundles (e.g. *one of the most, a large number of*).

The studies of bundles in different discourse/text types, mostly English texts, has provided some insight into the sets of lexical bundles relied on by spoken and written modes or in different disciplines. We now know, for example, that novice L2 English writers use fewer hedging devices than novice L1 English writers or that written academic English uses more referential bundles than stance bundles. However, some of the findings are not always consistent across different studies. For example, while Adel and Erman (2012) found that more proficient, L1 English, students used more different lexical bundles, Hyland (2008a) found that the less proficient Master's theses writers relied on more prefabricated strings than the expert research article writers. These discrepancies, as Chen and Baker (2010) suggested, may be due to the use of corpora of different sizes and text types across different studies.

In Hyland (2008a) the corpus being made up of Master's theses and PhD dissertations by second language learners, in addition to research articles, it may not be an accurate representation of expert writing in business studies, as well as biology, applied linguistics and electrical engineering. Also, his comparison of the distribution of his functional categories across the four disciplines is rather broad. In other words, functionally, he compared the uses of bundles in soft sciences (business studies and applied linguistics) versus hard sciences (electrical engineering and biology) instead of comparing their uses across the individual four disciplines.

2.4. The present study

The current study is different from earlier studies of lexical bundles in business by focusing only on one genre, research articles, in Master's-level texts and in a specific sub-discipline, finance, given that many studies have tended to study lexical bundles in business or business studies, not in its specific sub-disciplines. This is intended to provide master's-level finance students with a restricted list of bundles specific to their sub-field instead of a list related to business studies as a whole, part of which may not be useful to finance students. My objective in doing that is to maximize the return for the learning effort by focusing on sets of bundles that Master's-level students will most likely encounter in their readings, and will need when writing in their sub-discipline. Lexical bundles, among others, provide frames for interpreting the following, larger, phrases and clauses, and the developing discourse as a whole. They are also used to structure and organize academic texts. As such, understanding the meanings, uses and functions of frequent lexical bundles in Master's-level finance research articles is essential for students to read and understand Master's-level finance texts. For writing, by identifying them I will provide master's-level finance students, both first and second language learners, with prefabricated expressions, among others, to structure their writings and show the relationships between their ideas in a discipline-appropriate way in order to better communicate the meanings of their texts.

I chose research articles because tremendous scholarship is disseminated through them (Hyland, 2008a), and they represent models of writing from which students can learn. In order to study lexical bundles in master's-level finance research articles, I will attempt to answer the following research questions:

What are the most frequent lexical bundles in Master's level finance research articles?

What are their structural characteristics?

What are their functional characteristics?

Based on findings from research into expert academic written texts, it can be hypothesized that, structurally, a large proportion of the lexical bundles identified in the corpus will be noun-based and prepositional-based phrases. Functionally, it can be hypothesized that a large proportion of the identified bundles will include referential or research-oriented bundles. If there are verb-based bundles, they will probably include extraposed structures such as *it* + Vbe/V+ adjective+ *to*/that.

CHAPTER THREE: METHODOLOGY

This sections presents the three main steps involved in conducting the study. In other words, it presents information about how the corpus was collected, how the lexical bundles were identified, and how the retrieved bundles were categorized structurally and functionally. Specifically, the first step includes choosing the texts comprising the corpus, cleaning the chosen texts and converting the texts in machine-readable format (plain text format). The second step mainly consists of identifying and retrieving the lexical bundles using a computer program, and specific frequency and distribution/range criteria. The third step is mostly about the structural and functional categorization of the identified bundles by displaying the textual contexts in which the target bundles occur in the corpus.

3.1. Corpus collection

In order to create a corpus that is representative of finance research articles as used at the master's-level, six professors in the finance and real estate department at Colorado State University were consulted. The department has two undergraduate programs: real estate and finance; and one graduate program: finance. The six professors were asked to give the titles of journals from which they took articles or would take articles as reading assignments for their master's-level students. In total, 17 journals were identified (See Appendix A for a full list of the journals suggested). Four journals, which were identified by all or most professors, were targeted, including *Journal of Corporate Finance*, *Financial Analysts Journal*, *Journal of Portfolio Management* and *Journal of Derivatives.*

Research articles from each of the four journals formed about 250, 000 words, resulting in 1,034,587 words. Specifically, I downloaded about 50 articles from each of the four journals and converted them into word doc using the tools pane of Adobe Acrobat 11 Pro. Following Cortes (2004) I removed the references, graphs, tables, scientific formulae, headers, footers, captions from each article. After additional dates were removed from the articles, the number of words were counted, and all articles from a journal were placed in one file, converted into plain text with the name of the journal, and labelled after the name of the journal. It should be noted that I sampled whole articles, not only parts of them.

In the end, I arrived at a corpus of 1,034, 587 words, which I called MFRAC. It stands for Master's-Level Finance Research Articles Corpus. In terms of the size of the corpus, Biber (2006) suggested that a corpus be large enough to adequately represent the language features under investigation, as those features occur in the language. Generally, the study of the language features requires a larger corpus than the study of the features of specialized language (McEnery, Xiao & Tonio, 2006). In other words, general corpus is usually larger or bigger than specialized corpora. In this study, a one-million-word corpus was considered a suitable size because specialized language was being studied, and also because the corpora of published history writing and biology writing used in Cortes (2004) are 966,187 words and 1,026,344 words respectively. In other words, her two corpora are about 1,000,000 words.

Table 1, below, shows the composition of the corpus, namely, the number of words/tokens of the texts of each of the four journals whose articles form the corpus, and the total number of words of the corpus. It also presents information about the number of texts/articles from each of the four journals, and the total number of texts/articles that make up the corpus. Finally, table 1 shows the average number of words per text in the journals and the corpus as a whole.

Table 1

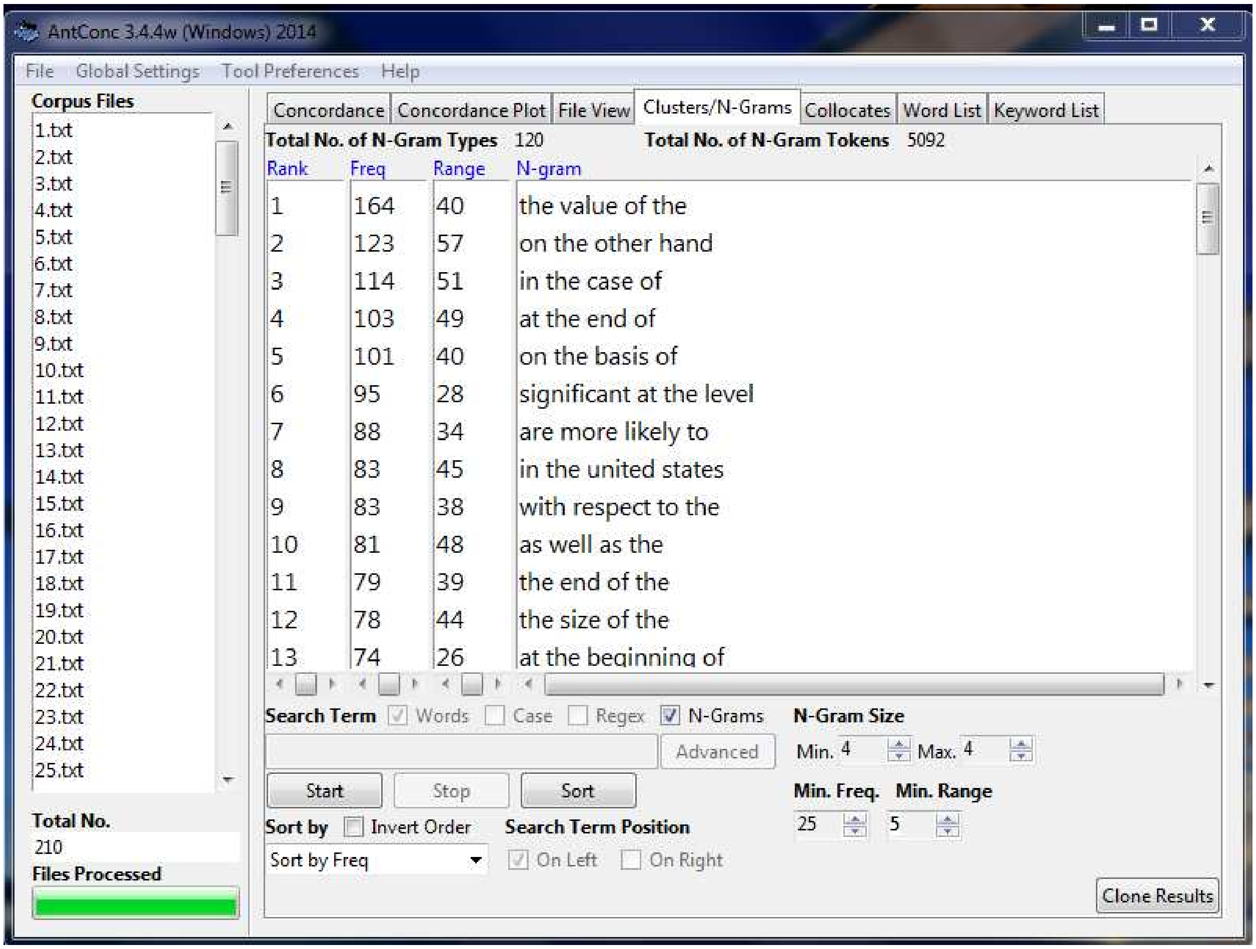
Corpus of master's-level finance research articles (MFRAC)

Journals	Number of words	Number of articles/texts	Average number of words per texts
Journal of Corporate Finance	275,374	31	8302.38
Financial Analysts Journal	259,649	45	5769.97
Journal of Portfolio Management	264,380	100	2643.8
Journal of Derivatives	253,521	35	7243.45
Corpus	1,034,587	210	

3.2. Identification/Selection of lexical bundles

As a reminder, lexical bundles are essentially strings of words that tend to co-occur together, identified on the basis of their frequent occurrence in texts, and which have certain discourse functions (Hyland, 2008). Also, they do not usually have idiomatic meanings, and are not complete structural units (Biber et al., 2004), usually bridging two phrases or clauses.

However, the frequency cut-offs for identifying lexical bundles are rather arbitrary and are different in different studies. In the current thesis, after the corpus had been created, the lexical bundles classified structurally and functionally were identified in four main steps. The researcher started by retrieving the four-word bundles occurring at 25 times per million words and in five different texts (articles). Following previous studies on lexical bundles (e.g. Cortes, 2004; Hyland, 2008; Chen & Baker, 2010; Herbel-Eisenmann, Wagner, & Cortes, 2010) four-word lexical bundles were targeted in order to have a manageable number of bundles for concordance checks and categorization, often about a hundred bundles. The uses of four-word bundles were also investigated in this study because, according to Hyland (2008a), they have clearer structural and functional differences among themselves than 3-word bundles. In order to retrieve the four-word bundles from MFRAC, the current study uses AntConc 3.4.4. Specifically, the cluster and N-grams function of AntConc 3.4.4 was used to identify the bundles that occurred at the minimum frequency and range set by the researcher. AntConc 3.4.4 was mainly used because it allows one to retrieve and extract strings of varying lengths, and to set desired or chosen minimum frequencies and ranges/distributions of occurrences for the strings rather easily. It also allows to display the concordance lines and wider textual contexts in which the target bundles occur for structural and functional classification. *Figure 1,* below, shows part of the four-word lexical bundles from the MFRAC as displayed by AntConc 3.4.4.



Total No. 210

Total No. of N-Gram Types 120 Total No. of N-Gram Tokens 5092

Rank	Freq	Range	N-gram
1	164	40	the value of the
2	123	57	on the other hand
3	114	51	in the case of
4	103	49	at the end of
5	101	40	on the basis of
6	95	28	significant at the level
7	88	34	are more likely to
8	83	45	in the united states
9	83	38	with respect to the
10	81	48	as well as the
11	79	39	the end of the
12	78	44	the size of the
13	74	26	at the beginning of

Search Term ☑ Words □ Case □ Regex ☑ N-Grams

N-Gram Size Min. 4 Max. 4

Min. Freq. 25 Min. Range 5

Sort by □ Invert Order

Sort by Freq

Search Term Position ☑ On Left □ On Right

Figure 1. Screenshot showing part of the MFRAC bundles

After the first step, which had consisted of retrieving the four-word bundles occurring at least 25 times per million words and in five different articles, the researcher displayed the concordance lines for the retrieved lexical bundles, and manually checked how many of them occurred in the texts of at least three of the four journals, which is one of the criteria used in identifying the lexical bundles. Twenty different lexical bundles did not occur in the research articles of at least 3 journals, and these, which include *and non family firms and panel a of table*, were discarded (See Appendix B for a complete list of the 20 lexical bundles removed).

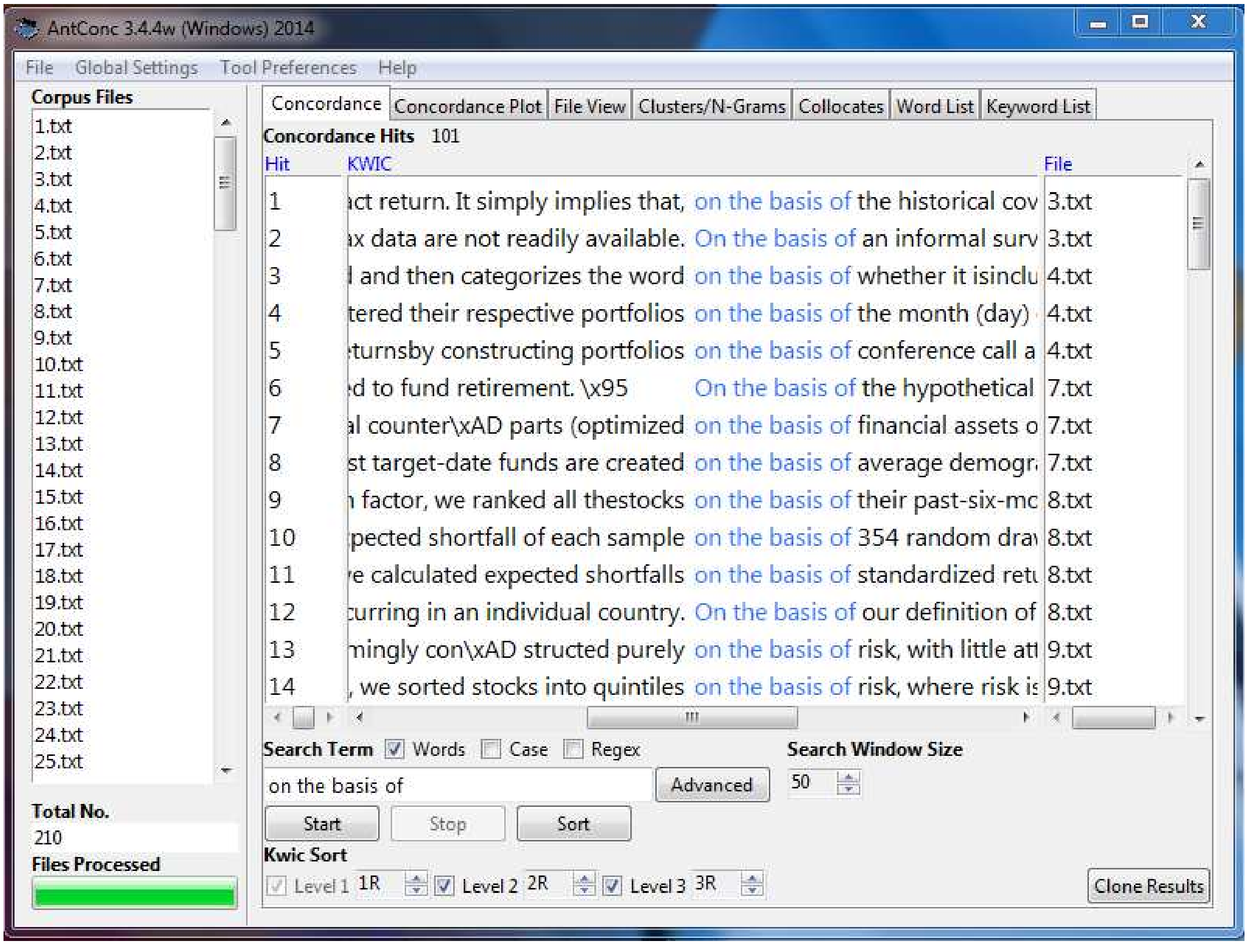
After the second step, the researcher examined/checked the remaining bundles for content-dependent bundles. Specifically, the remaining bundles were scanned in order to detect bundles that appeared to indicate or refer to topics or subjects related to finance or business. Four different types of lexical bundle seemed to be content-dependent, and these were *the global financial crisis, the s p index, of the s p* and *the Sharpe ratio of,* whose concordance lines were analyzed to verify that they refer to topics or information closely related to finance or business. *the global financial crisis* refers to the 2008 crisis which started in the United States; *the s p index* and *of the s p* refer to The Standard & Poor's 500, which is a stock market index using the market capitalization of 500 large companies commonly listed on the New York Stock Exchange and the NASDAQ. *the Sharpe ratio of* is a type of standard measure used in finance. Along with the only one content-dependent bundle, another bundle was removed from the remaining list of bundles. The bundle is *xad tion of the,* and it is mainly made up of the affix *ation* and *of the.*

In the fourth step, overlaps were detected, their concordance listings were checked to see whether each set of overlap performed the same or similar discourse functions. Overlapping bundles involve cases where two similar four-word strings occur or overlap in four-, five- or six-word strings. In total, 15 sets of overlaps were merged, and these sets include *as a result of* and *as a result the,* and *on the basis of and the basis of the.* It should be noted that part of the overlaps that were merged are used differently to express the same functions. For example, *as a result the* is used after the cause has been mentioned, and the effect or result is mentioned just after the bundle while in the case of *as a result of* both the cause and the effect are mentioned after the bundle (See Appendix C for a complete list of the 15 sets of overlaps).

After the initial list of bundles retrieved from AntConc 3.4.4 had been refined, the researcher classified the bundles structurally and functionally.

3.3. Structural and Functional classification/categorization

AntConc 3.4.4 was used to display the concordance lines for each of the remaining bundles to first classify them structurally (e.g. noun phrase fragments, complete phrases) and then functionally. AntConc 3.4.4 allows to display all the lines in which a target lexical bundle occurs in the corpus. These lines are referred to as concordance lines. *Figure 2*, below, presents part of the concordance lines of one of the bundles identified in the MFRAC. The target bundle appears in the middle of the lines.



Total No. 210

Concordance Hits 101

Hit	KWIC	File
1	ıct return. It simply implies that, on the basis of the historical cov	3.txt
2	ıx data are not readily available. On the basis of an informal surv	3.txt
3	l and then categorizes the word on the basis of whether it isinclu	4.txt
4	tered their respective portfolios on the basis of the month (day)	4.txt
5	turnsby constructing portfolios on the basis of conference call a	4.txt
6	d to fund retirement. \x95 On the basis of the hypothetical	7.txt
7	l counter\xAD parts (optimized on the basis of financial assets o	7.txt
8	st target-date funds are created on the basis of average demogr	7.txt
9	ı factor, we ranked all thestocks on the basis of their past-six-mc	8.txt
10	pected shortfall of each sample on the basis of 354 random dra	8.txt
11	e calculated expected shortfalls on the basis of standardized retı	8.txt
12	curring in an individual country. On the basis of our definition of	8.txt
13	ningly con\xAD structed purely on the basis of risk, with little att	9.txt
14	, we sorted stocks into quintiles on the basis of risk, where risk is	9.txt

Search Term ☑ Words ☐ Case ☐ Regex

on the basis of

Search Window Size 50

☑ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R

Figure 2. Screenshot showing the concordance listing for one of the MFRAC bundles

In total, 11 structural categories were used to classify the bundles of the MFRAC. eight of the original 12 major structures used in Biber et al. (1999) were used, two categories were adapted, and one new was adapted. Table 3, below, presents the structural categories used to classify the bundles of the MFRAC, and examples of the structures.

Table 2

The structural categories used to classify the bundles of the MFRAC

Structures	Examples
1. NP with *of*-phrase fragment	the value of the/a
2. NP with other post-modifier fragment	the extent to which
3. NP with pre-modifiers (created)	the risk free rate
4. Prepositional phrase with embedded *of*-phrase fragment	in the context of
5. Other prepositional phrase fragment	on the other hand
6. Anticipatory *it* + verb phrase/adjective phrase	it is important to
7. Passive verb + prepositional phrase fragment	is defined as the
8. (Pronoun/NP) + copula *be* + NP/adjective phrase (adapted)	the dependent variable is
9. (NP) (verb phrase +) *that*-clause fragment (adapted)	we find that the
10. (verb/adjective +) *to*-clause fragment (adapted)	to control for the
11. Other expressions	as well as the

In order to categorize the extracted bundles in terms of their functions in discourse, I used Hyland's (2008a, 2008b) three main categories of research-oriented, participant-oriented and text-oriented bundles. The aforementioned categories were chosen over Biber, Conrad & Cortes (2004) and Biber & Barbieri (2007) because, as Hyland (2008a) pointed out, the functional categories used by Biber and colleagues emerged from the analysis of more broader and general categories than his research-focused genres of research articles, master's theses or PhD dissertations.

Research-oriented bundles as used in Hyland (2008a, 2008b) include five categories, which are location, procedure, quantification, description and topic. But the current study used four of the five categories, that is, location, procedure, quantification and description. Location bundles are used to refer to the time and place of some activity, event, phenomena, etc.; procedure bundles are used to refer to steps or methods in carrying out some action; quantification bundles, among others, evaluate or refer to the size, value or amount of something, or are used to speak about something in relation to the size, value or amount of something. The description bundles are used to describe miscellaneous phenomena, processes, etc. Bundles in the topic sub-category, according to Hyland (2008), are "related to the field of research." (p. 13) Such bundles, in Hyland (2008a), include *in the Hong Kong, the currency board system*.

Participant-oriented bundles include stance features and engagement features, but only 1 sub-category, stance features, was used. Stance features indicate a writer's attitudes and evaluations of information, knowledge, being discussed. In Hyland (2008), for example, participant-oriented bundles include *may be due to, it is possible that it should be noted that*, etc.

Text-oriented bundles include 4 sub-categories, which are transition signals, resultative signals, structuring signals and framing signals. Transition signals are used to indicate addition as well as contrast relationships between elements of discourse. Resultative signals, according to Hyland (2008a) “mark inferential and causative relations between elements” (14). Structuring/referring/relating bundles are used to refer to or direct attention to specific discourse, author or work, or to organize texts. Framing bundles specify the conditions of the arguments. Table 3, below, presents the functional categories used to classify the bundles retrieved from the MFRAC.

Table 3

The functional categories used to classify the bundles of the MFRAC

Categories		Examples
Research-oriented		
	Location	over the sample period
	Procedure	to control for the
	Quantification	the price of the
	Description	the volatility of the
Text-oriented		
	Transition signals	on the one hand
	Resultative signals	as a result the
	Structuring/relating/referring signals (adapted)	in line with the
	Framing signal	with respect to the
Participant-oriented		
	Stance features	it is important to, we assume that the

When the functional category of a bundle could not be easily identified based on concordance lines, the extended textual context in which bundles occurred were examined to determine their discourse functions. The bundles that I found challenging, and whose wider textual contexts I had to examine, are mainly description and procedure lexical bundles. It may be because concordance lines do not always provide enough information to separate description bundles from procedure bundles because procedure bundles also involve a type of description. Some of the challenging bundles include *can be used to, to control for the* for procedure bundles, and *the correlation between the, the ratio of the* for description bundles.

In order to improve the reliability of my functional classification, I had another coder examine 10% of my bundles. The coder agreed with my functional classification of all the sample bundles, even though she asked me to slightly modify the terminology (i.e. use text-oriented location phrases instead) used to refer to text-oriented bundles in order to fully reflect the functions served by bundles in the text-oriented category. But I decided to use the original terminology because I thought the coder examined only 10% of bundles, and as such may not have had a full picture of the functions and uses of bundles in the text-oriented category.

CHAPTER FOUR: RESULTS AND DISCUSSION

This section presents and discusses the results of the analysis of the four-word bundles identified in the MFRAC. After giving an overview of the bundles retrieved from MFRAC, their structural and functional characteristics will be presented and discussed. In addition, the results of the analysis of the relationship between the structural and functional characteristics of identified bundles will be presented and discussed.

One hundred and twenty different types of lexical bundles were retrieved, using the frequency cut-off and range set by the researcher. After refinement (i.e., after bundles not occurring in at least three journals, non-formulaic bundles, content-dependent bundles were removed, and over-lapping bundles were merged), a final list of 80 different types of bundles was compiled for structural and functional categorization. In terms of the number of tokens, 5092 bundles were identified in total and before refinement. The words in the 5092 four-word lexical bundles make up 0.49% of the words in the corpus. This is less than the 2% of four-word bundles found by Biber et al. (1999). This may be attributed to the much higher frequency cut-off – 25 times per million words -- used in this study compared to Biber et al. (1999), who used a frequency cut-off of 10 times per million words.

After refinement, I arrived at 4366 tokens. More than 60% of the 82 different lexical bundles identified after refinement occur more than 30 times per million words in the MFRAC. These bundles include *in addition to the*, *the nature of the* and *in this case the*. The 20 most frequent bundles occur more than 50 times per million words. Almost all of the bundles occur in more than 10 different texts/articles, with the most widely distributed bundles occurring in more than 40 different texts, as indicated in the range column of table 5. Table 4, below, shows the number of different types of lexical bundles as well as the number of tokens in the MFRAC before and after refinement.

Table 4

Number of tokens and different types of lexical bundles before and after refinement

Before refinement		After refinement	
Number of types	120	Number of types	80
Number of tokens	5092	Number of tokens	4,281

The most frequent bundles include *the value of the/a, on the other hand, (in) the case of (the), (at) the end of (the)* and *as well as the*, and these were part of those identified in previous studies (e.g. Biber, Johansson, Leech, Conrad, & Finegan, 1999; Hyland, 2008a, 2008b) as the most frequent lexical bundles in academic prose. Table 5, below, presents the 20 most frequent different lexical bundles of the MFRAC along with information such as their frequencies and ranges.

Table 5

The 20 most frequent lexical bundles in the MFRAC

Lexical bundles	Raw frequency	Number of bundles per text	Range/distribution
the value of the/a	164	4.10	40
on the other hand	132	2.32	57
(in) the case of (the)	114	2.24	51
(at) the end of (the)	103	2.10	49
(on) the basis of (the)	101	2.53	40
significant at the level	95	3.39	28
(are) more likely to (be)	88	2.44	34
(in) the united states (and)	83	2.18	45
with respect to the	83	2.18	38
as well as the	81	1.68	48

the size of the	78	1.77	44
(at) the beginning of (the)	74	2.84	26
in the context of	72	1.6	45
we find that the	66	2.12	31
is/are consistent with the	62	1.72	36
the risk free rate	61	2.25	27
out of the money	60	3	20
at the same time	59	1.43	41
the difference between the	59	1.47	40
the standard deviation of	58	2.23	26

However, among the 80 different types of lexical bundles identified in this study, 22 were not identified in earlier studies on academic prose (e.g. Cortes, 2004; Hyland, 2008). The 22 different types of bundles specific to the MFRAC include bundles both among the most frequent bundles and the least frequent bundles. For example, in the 30 most frequent lexical bundles of the MFRAC, six were found not to have been identified in the previous literature (e.g. Biber et al., 1999; Cortes, 2004; Hyland, 2008a, 2008b) read by the researcher. These bundles, which occur more than 45 times per million words, include *significant at the level, we find that the, the risk free rate*. Table 6, below, presents the six most frequent different types of lexical bundles specific to the MFRAC.

Table 6

The 6 most frequent lexical bundles specific to the MFRAC

Lexical bundles	Raw frequency	Number of bundles per text	Range/Distribution
significant at the level	95	3.39	28
we find that the	66	2.18	31
the risk free rate	61	2.25	27
out of the money	60	3	20
in this article we	57	1.58	36
as a proxy for	47	1.80	26

In the less frequent lexical bundles of the MFRAC, that is the strings that occur less than 45 times per million words, items that were not identified in previous literature on academic prose include *is a dummy variable, over the sample period, as a measure of, (is) the ratio of (the).* Table 7, below, presents the 16 different types of lexical bundles occurring less than forty-five times, and which seem to occur specifically in the MFRAC, compared with earlier studies read by the researcher.

Table 7

The 16 less frequent lexical bundles specific to the MFRAC

Lexical bundles	Raw frequency	Number of bundles per text	Range
is a dummy variable	38	2.37	16
the dependent variable is	36	1.89	19

the volatility of the	36	1.89	19
over the sample period	35	2.18	16
as a measure of	33	2.06	16
the correlation between the	33	1.43	23
(is) the ratio of (the)	29	2.23	13
of the stock price	29	2.23	13
of the underlying asset	29	4.83	6
the present value of	29	1.93	15
the coefficient of the	27	1.8	15
the return on the	27	2.07	13
the market value of	26	1.36	19
to control for the	26	1.52	17
more than of the	25	1.78	14
the absolute value of	25	2.08	12

It is worth pointing out that 22 different types of bundles were not identified in the sub-corpus of business studies analyzed in Hyland (2008a). This seems to indicate that matser's-level finance research articles to some extent rely on different sets of lexical bundles than business studies in general. The results seem to reinforce the view of Hyland and Tse (2009), who argue for a restricted list of focused and specific items for students in specific disciplines. In view of the 22 different types of bundles that appear to occur specifically in the MFRAC, it seems that students in different sub-disciplines or sub-fields to some extent may rely on different sets of bundles. This, of course, warrants more research into bundles used in other sub-disciplines in other fields.

At this stage already it can be argued that the 22 different types of bundles not identified in previous literature reviewed above make a good list for a class of English for specific academic purposes. In other words, an EAP teacher may focus on the 22 different types of lexical bundles in a class composed of students preparing to take or taking master's-level finance classes. On the other hand, the rest of the lexical bundles, that is the items identified in earlier studies as well, may make a good list for classes of English for general academic purposes, where students in different disciplines are in the same class. It can be argued that this is even truer for the common most frequent bundles, such as *on the other hand,* or *in the case of*, identified in other academic texts as well.

After this preliminary overview, the following sections will present the results and discussion of the structural and functional analysis of the identified bundles. First, the structural characteristics of the identified bundles will be discussed.

4.1. Structural characteristics of the lexical bundles of the MFRAC

Like in previous studies on lexical bundles in academic prose (e.g. Biber et al., 1999; Cortes, 2004; Hyland, 2008a, 2008b), the strings identified in the MFRAC are not complete structural units, and they tend to bridge two structural units. This is illustrated by bundles such as

or *it is important to* and *of table shows that.*

However, the bundles can be grouped according to their structural correlates by analyzing the concordance listings of the identified bundles. But, like in previous literature, some of them had to be classified as *other expressions*, which, according to are strings "that do not fit neatly into any of the other categories" (p. 1024).

Adapting the structural categories identified in Biber et al. (1999), the lexical bundles in the MFRAC were found to have similar structural characteristics to bundles in academic prose, as reported in earlier studies. That is, most of the bundles identified in the MFRAC are phrase-based. Phrasal lexical bundles include noun phrase fragments, preposition phrase fragments, adjective phrase fragments and verb phrase fragments. In the MFRAC, noun phrase-based lexical bundles include *the size of the* or *the value of the*; prepositional phrase-based lexical bundles include *in the context of, (on) the basis of (the)*; verb phrase-based bundles include anticipatory *it* + verb phrase or adjective phrase (e.g. *it is important to*). Other verb phrase-based bundles are passive verb + prepositional phrase fragments exemplified by bundles such as *is defined as the, is based on the*.

Clausal bundles include bundles integrating *that*-clause fragments and *to*-clause fragments. In addition to the broad categories of phrasal and clausal bundles, there are bundles that fall within the *other expressions* category. *Other expressions* include *as well as the* or *more than of the*.

Of the 80 MFRAC different types of lexical bundles classified, 70 are phrasal, seven are clausal, and three were classified as *other expressions,* which shows the dominantly phrasal nature of the bundles in the MFRAC. These results are similar to findings in earlier studies on academic texts (e.g. Biber et al. 1999). Table 8, below, presents a complete list of the MFRAC 80 different types of lexical with their structural correlates. The bundles are divided into phrasal, clausal and other expressions.

Table 8

The lexical bundles of the MFRAC and their structural correlates, with the bundles specific to the MFRAC underlined.

Structures	Bundles
Phrasal	
1. Noun phrase with *of*-phrase fragment (18)	the size of the, the value of the/a, the nature of the, the results of/for the, the magnitude of the, the sum of the, the total number of, of the underlying asset the effect of the, the coefficient of the, the market value of, a large number of the absolute value of, the standard deviation of, the impact of the, the price of the the volatility of the, the present value of
2. Noun phrase with other post-modifier fragment (6)	the correlation between the, the extent to which, an increase in the, the return on the the relationship between the, the difference between the
3. Noun phrase with pre-modifiers (created) (1)	the risk free rate
4. Prepositional phrase with embedded *of*-phrase fragment (13)	(in) the case of (the), (on) the basis of (the), (at) the beginning of (the) as a measure of, in the context of, in terms of the, (as) a function of (the) for each of the, in the form of, (at) the end of (the), at the time of, as a result of/the (in) the presence of (a)
5. Other prepositional phrase fragment (22)	on the other hand, in the U S, (in) the united states (and), with respect to the over the sample period, in addition to the, in this case the, in line with the before and after the, for the united states, of the stock price, in the next section, on the one hand, in the long run, in our study we, in this paper we, out of the money at the same time, in this article we, in this section we, as a proxy for (to) the fact that (the)
6. Anticipatory *it* + verb phrase/adjective phrase (1)	it is important to
7. Passive verb + prepositional phrase fragment (2)	is defined as the, is based on the
8. (Pronoun/noun phrase) + copula *be* + noun phrase/adjective phrase (adapted) (8)	the dependent variable is, is/are consistent with, results are consistent with is the number of, is a dummy variable, (is) one of the (most), (is) the ratio of (the) significant at the level
	(continued)
Clausal bundles	
9. (noun phrase/pronoun) (verb phrase +) *that*-clause fragment (adapted) (3)	we assume that the, we find that the, that there is no
10. (verb/adjective +) *to*-clause fragment (4)	(are) more likely to (be), can be used to, to control for the, is likely to be
Other expressions (2)	

Structures	Bundles
11. as well as the, more than of the	

4.1.1. Phrasal bundles

Among the phrasal bundles, preposition phrase fragments are the most numerous (35 of the 80 different types of bundles classified structurally and functionally), followed by noun phrase fragments (25 of the 80 different types of lexical bundles classified). The rest of the phrasal bundles (11 of the 80 different types of phrasal lexical bundles) include other phrasal structures such as anticipatory *it* + verb phrase/adjective phrase fragment (one bundle), passive verb + preposition phrase fragments (two bundles), (noun phrase/pronoun) + copula *be* + noun phrase/adjective phrase (five bundles).

The great use of phrasal and nominal structures in finance research articles, as exemplified in the present study, reinforces the concept that academic writing, especially research articles, uses many more nouns than verbs, and more nominalization, compared to conversation, for example (Biber, 2006; Reppen, 2010) As such, the results of the MFRAC bundles structural analysis seem to be consistent with findings in Stoller and Robinson (2008), for example. This may reflect the trend which has consisted of the increased use of nominal and phrasal structures in written registers, such as academic research articles, with highly informational purposes where the researcher is to a great extent concerned with presenting the findings of their research to a specialized audience (Biber & Gray, 2013). Nominal/phrasal structures involve nouns derived from verbs or adjectives, or verbs converted to nouns, and they allow researchers to present information in a compact or compressed manner (Biber & Gray, 2013).

Around half of the 60 prepositional and noun phrase fragments, include some form of post-modification. The post-modification is realized with *of*-phrase fragments (e.g. *the effect of the, the coefficient of the*) as well as with other post-modifier fragments (e.g. *the extent to which, (on) the basis of (the)*).

The proportion of noun phrases involving post-modification (24 of the 25 different types of noun phrase fragments) is considerably higher than that of prepositional phrase fragments involving post-modification (about 14 of the 35 different types of prepositional phrase fragments). With more than half of the preposition and noun phrase fragments incorporating post-modifiers, the results in this study are in line with findings in earlier literature. That is why Cortes (2004) said “that academic writing […] is strongly marked for post-nominal modification, as in the case of genitive expressions or other prepositional phrases which are post-nominal modifiers” (p. 404), and since the majority of post-modifiers in the MFRAC are in noun phrase fragments, one may hypothesize that a great proportion of the post-modification in master’s-level finance research articles is realized in noun phrase fragments.

However, the majority of phrasal bundles do not involve post-modification. This may show that expert writers of master’s-level finance research articles also use many phrasal bundles not involving post-modification. A number of phrasal bundles not involving post-modification incorporate pre-modifiers. These bundles include *the risk free rate, the dependent variable is, on the one hand, on the other hand, in the next section, of the stock price*. Other bundles not involving any pre- and post-modification include *with respect to the, before and after the,* etc.

4.1.2. Clausal bundles

As indicated above, clausal bundles are far fewer than phrasal bundles, only seven of the MFRAC bundles classified, with three bundles incorporating *that*-clause fragments (e.g. *we assume that the*) and four *to*-clause fragments (e.g. *can be used* to). The use of fewer clausal and verbal fragments in the MFRAC may be attributable to the general historical trend which has consisted of reducing the use of verbal and clausal structures in professional academic research articles (Biber & Gray, 2013). Again, Biber and Gray (2013) attributes this trend to the "the combination of a highly specialized audience and a highly informational purpose dealing with technical information" (p. 25). This, in turn, according to Biber, Grieve and Iberri-Shea (2009), is due to the "pressure to communicate information as efficiently and economically as possible, resulting in compressed styles that depend heavily on tightly integrated noun phrase constructions" (p. 184).

4.1.3. Other expressions

The third broad category includes even fewer bundles than clausal bundles. Called *other expressions* by Biber et al. (1999), they are represented by only two bundles (*as well as the, more than of the*). Even though the two bundles do not belong to any of the 10 other structural categories used in this study, the analysis of their concordance listings reveals that they are more phrasal than clausal in nature.

Unlike Biber et al. (1999), the structural categories in the MFRAC do not include lexical bundles incorporating adverbial clause fragments and copula *be* + noun phrase/adjective phrase fragments. On the other hand, a new structural category (a noun phrase with a pre-modifier) was created in order to classify one bundle (*the risk free rate*) that does not belong to any of the structural categories identified in Biber et al. (1999). But overall the structural characteristics of the MFRAC bundles are similar to those found in the business studies sub-corpus in Hyland (2008). Also, a number of bundles exclusive to Hyland's (2008) business studies sub-corpus and applied linguistics were also present in the MFRAC. These are *on the basis of, in the context of, the relationship between the, it is important to*.

4.2. Functional characteristics of the lexical bundles of the MFRAC

As explained in the method section (chapter 3) the categories used to study the functional characteristics of the bundles of the MFRAC were developed by Hyland (2005, 2008a, 2008b). They include the three broad categories of research-oriented, text-oriented and participant-oriented. Each of the broad categories encompass sub-categories.

After the classification of the identified bundles, it was found that the distribution of the bundles in the functional categories is similar to those found in previous literature on lexical bundles in academic prose (Hyland, 2008a, 2008b). In other words, the bundles were found to be predominantly research-oriented and text-oriented, with very few participant oriented. This may, to some extent, reflect the impersonal and relatively objective nature of academic prose, compared to conversation, for example, which is more involved. Of the 80 different types of bundles 53 are research-oriented bundles -- more than half of the total number of bundles identified after refinement --, and 23 are text-oriented. Only four bundles are participant-oriented. Research-oriented bundles are to some extent comparable to referential expressions while text-oriented can be said to be to some extent comparable to the discourse organizers developed in Biber et al., (2004). Examples of research-oriented include *the coefficient of the, the effect of the*; text-oriented bundles include *on the other hand, in addition to the*. The few participant-oriented bundles are *(are) more likely to (be), is likely to be, it is important to,* and *we assume that the*.

Table 9, on pages 58, 59 and 60, shows the total number of bundles after refinement, and classified according to their functional categories.

When the three broad categories are broken down into their sub-categories, description sub-category has the most bundles, followed by the quantification, which in turn, are followed by the location bundles, and finally, only two procedure bundles. This may suggest the greater role of description in the master's-level research articles, compared to quantification or location. It may also be that quantification, or even procedures are realized using other devices than four-word lexical bundles.

In the text-oriented bundles, structuring/relating/referring and framing signals include roughly equal numbers of bundles, eight and nine bundles respectively. The bundles in these sub-categories include far more bundles than transition and resultative sub-categories, which include four and two bundles respectively. This may suggest that specifying the conditions under which arguments are valid, and referring to specific discourse or work, for example, are features of master's-level research articles that are more used than transition devices, for example. It may also be that transitions are realized using other devices than four-word lexical bundles. The predominance of structuring/relating/referring and framing signals are to some extent consistent with findings in Hyland (2008a), where framing bundles, which make up around 50% of the text-oriented, are used, among others, to specify the conditions under which arguments apply.

In the participant-oriented bundles, only bundles in the stance features sub-category were identified in the MFRAC using the identification criteria referred to in chapter 3. Unlike Hyland

(2008a), no bundle in the engagement feature sub-category was identified in the present study. This may suggest that writers in the MFRAC use other devices to engage the reader by directly addressing him, or that this feature of academic discourse has a limited use in master's-level finance research articles.

It is worth mentioning that unlike the present study, the business studies sub-corpus in Hyland (2008a) was found to have more text-oriented bundles than research-oriented bundles, and about four times more participant-oriented bundles than the bundles of MFRAC. This may also suggest that while a single list of lexical bundles in business texts might be useful and adequate for the students in all the sub-fields and sub-disciplines of business, more specialized lists might be more adequate and preferable.

Table 9

The lexical bundles of the MFRAC and their functional characteristics, with the bundles specific to the MFRAC underlined

Category	Sub-category	Bundles
Research-oriented		
	Location (10)	
		(at) the beginning of (the), (at) the end of (the), in the U S at the same time, at the time of, over the sample period (in) the united states (and), for the united states, in the long run before and after the
	Procedure (2)	
		can be used to, to control for the
	Quantification (16)	
		the sum of the, the size of the, the price of the, is the number of the magnitude of the, the extent to which, (is) one of the (most) the total number of, the present value of, the value of a/the the market value of, the absolute value of, a large number of as a measure of, more than of the, the risk free rate

Category	Sub-category	Bundles
		(continued)
	Description (26)	the volatility of the, the impact of the, the dependent variable is the nature of the, the effect of the, (to) the fact that (the) the results of/for the, the coefficient of the, the return on the the correlation between the, the relationship between the of the underlying asset, the standard deviation of, the difference between the, as a proxy for, in the form of (is) the ratio of (the), of the stock price, is defined as the that there is no, is a dummy variable, an increase in the is based on the, out of the money, (as) a function of (the) we find that the
Text-oriented	Transition signals (4)	on the other hand, in addition to the, on the one hand as well as the, at the same time
	Resultative signals (1)	as a result of/the
	Structuring/relating/referring signals (adapted) (8)	in line with the, in the next section, is/are consistent with in our study we, results are consistent with in this section we, in this article we, in this paper we
	Framing signals (9)	with respect to the, (on) the basis of (the), for each of the in the context of, in terms of the, significant at the level in this case the, (in) the case of (the) (in) the presence of (a)
Participant-oriented	Stance features (4)	(are) more likely to (be), it is important to, is likely to be we assume that the

4.2.1. Research-oriented bundles

The bundles in the research-oriented category are used in the MFRAC to speak about some activities, experiences, phenomena, etc., by indicating their time and place, by describing them, quantifying them and by speaking about some procedure related to these activities, experiences, phenomena, etc.

Of the four sub-categories, the bundles of the description sub-category form the largest sub-category, which may suggest that the description of miscellaneous phenomena, processes, events, etc., is a prevalent feature of writing in master's-level research articles. This is illustrated in the following extract, where the bundle is used in the description of a phenomenon:

> *the volatility of the* underlying asset also affects the critical asset prices because it influence of the time value of the option. Since exercising prematurely to avoid a credit loss means that the option holder is giving up the time value of the option, the effect of credit risk on the critical asset price should be higher for options with relatively high time value.

The quantification bundles are used, among other reasons, to evaluate or to refer to the size, value or amount of something, or to speak about something in relation to the size, value or amount of something, as illustrated in the following:

> Board size is *the total number of* board members sitting on the board. A board member is classified as an insider, if the person is an employee, or a former employee of the company. The number of outsiders (Outsiders) is calculated as Board size minus Insiders.

This specific bundle mostly occurs in the middle of a sentence where it is used to refer to all the members or elements of a group being considered while adding emphasis.

The third sub-category in terms of number of bundles is location, with bundles that indicate time and place in relation to some abstract, concrete or physical entities, including countries. This is exemplified in the following extract:

> Our evidence thus suggests that the underlying securities are generally very volatile *at the time of* issuance. We cannot identify the cause for the high volatility, whether firm specific or market wide.

The bundle in the example is mostly used to indicate that some event, phenomenon, process or activity occurs as another happens.

Contrary to the aforementioned sub-categories, procedure bundles were found to be almost non-existent, being limited to two bundles. They are used in relation to some procedure involved in the accomplishment or realization of something. This is exemplified in the following:

> For the first stage, the Heston semianalytic pricing formula […] *can be used to* achieve fast and accurate calibration for the term-structure Heston model as long as the characteristic function of the model is available (see Elices [2009)). The Levenberg-Marquardt nonlinear least squares optimization is then performed to find the optimal parameters.

The bundle in the above example is used to indicate that series of actions can be put to use to perform some action in addition to others.

4.2.2. Text-oriented bundles

Text-oriented bundles have 4 sub-categories, with framing and structuring/referring/ relating bundles being far more present than resultative and transition signals. Text-oriented bundles have as a common characteristic the organization and structuring of texts. Framing bundles specify the conditions of the arguments, as in the following:

> *With respect to the* value factor, this change represents a complete reversal of the low-risk portfolio's strong positive relationship since the 1980s up until the recent financial crisis.

In the above example, the bundle is used to refer to the specific element relative to which the information presented in the following proposition is valid or true. By framing the proposition, the writer also shifts the focus from one proposition/discourse element to the other.

Structuring/referring/relating bundles are used to refer or direct attention to specific discourse, author or work, or to organize texts. The original structuring sub-category was adapted to include bundles that also refer to an author or work, not only specific part of the text or the whole text. The following shows an instance of structuring/referring/relating bundle used in the wider textual context:

> Although we find that ROA is significantly and positively related to Size for SOEs and not significantly related for private firms, our result for ROA *is consistent with the* result for Tobin's q since both results imply that SOEs benefit relatively more from size than private firms.

In the specific example, above, the writer refers to previous work to point to similarity between their result and findings in the previous work. Referring to previous literature is a feature of academic prose, where writers relate or connect their research to previous literature on which they try to build while referring to similarities and differences between their findings and findings in previous literature.

Resultative signals, far fewer than the two text-oriented sub-categories referred to above, include bundles such as *as a result of, as a result the*, and their uses are illustrated in the following:

> *As a result of* this fraud, in October 1986, the ASC placed a moratorium on new blind pool stock offerings until the program could be reviewed. During the review, it was noted that small public companies share many of the challenges facing private companies that seek venture capital (VC) financing.

In this specific example, in the first sentence the bundle indicates that the first proposition is the reason why the actions in the second proposition were taken. This bundle refers to rather direct resultative and causative relationships between the propositions that it connects or shows have relationships to each other.

Transition signals are used to indicate addition as well as contrast relationships between elements of discourse, and they include bundles such as *on the other hand* or *we also find that.* In the specific example, below, the information following the bundle comes in addition to the one presented in the preceding proposition by not contradicting the preceding information:

> We demonstrate that government connections are associated with substantially less severe financial constraints […] *We also find that* those large non-state firms with weak government connections […] are especially financially constrained, due perhaps to the formidable hold that their state rivals have on financial resources […].

4.2.3. Participant-oriented bundles

Bundles in this category are far fewer than the ones in the two other broad categories, and only bundles in the stance features sub-category were identified in the MFRAC. These, according to Hyland (2008a), “convey the writer’s attitudes and evaluations” (p. 14) of information or knowledge being discussed as in the following:

> Their announcements might simply reflect a need to remind investors of the difficult economic times they face. Thus, their future company-specific performance *is likely to be* neutral. The more interesting result occurs if we find that firms that blame others exhibit poor company-specific performance prior to the announcement.

Here it can be argued that the speaker is using hedging in their evaluation. The use of cautious language has been reported to characterize expert writing (Chen and Baker, 2010).

4.3. The relationship between the structural and the functional characteristics of the MFRAC bundles

Phrasal bundles, the great majority of which are noun phrase- and prepositional phrase-based bundles, are overwhelmingly present in the research-oriented and text-oriented bundles. On the other hand, the very few participant-oriented bundles include anticipatory *it* + verb phrase/adjective phrase, and *that*-clause and *to*-clause fragments. These results are consistent with findings in previous literature such as Biber et al. (2004) or by Hyland (2008a, 2008b).

When one closely examines the structures in the functional categories, prepositional phrase fragments make up most of the bundles in the research-oriented and text-oriented bundles. Prepositional phrase fragments make up more than 80% of text-oriented bundles, that is, 19 of the 22 text-oriented bundles. The predominance of prepositional phrase fragments is probably due to the fact that many linking adverbials or transition phrases are prepositional phrases. These are used to organize the text, and guide the reader through the text and facilitate comprehension by showing how ideas connect or relate to each other. They are also used to make the text a cohesive and coherent whole by showing how and why ideas logically follow or precede each other (e.g. *as a result of/the, we find that the*). These relationships include addition ones (e.g. *in addition to the, as well as the*), contrastive ones (e.g. *on the one hand, on the other hand*) or cause and effect relationships (e.g. *as a result of/the*).

In research-oriented bundles, 24 of the 54 bundles are noun phrase fragments, and all of them are in the description and quantification sub-categories. In contrast, all location bundles are prepositional phrase fragments, even though they do not usually function as linking adverbials, but indicate time and place.

Participant-oriented bundles include an extraposed structure (*it is important to*), *to*-clause fragments (is *likely to be, (are) more likely to (be)*), and *that*-clause fragment (*we assume that the*). The predominance of adjective phrase and clause fragments in this category may reflect the fact that the expressions showing stance are usually verb, adjective and clause phrase fragments.

4.4. Pedagogical Implications

The findings of this study, like in earlier studies such as Hyland (2008a), Cortes (2004), Nattinger and DeCarrico (1992) are important for EAP teachers and students. First and foremost, they reinforce the idea that lexical bundles should receive more attention in EAP programs because they were consistently shown to serve important discourse functions in written academic texts or other registers.

At the same time the findings in the present study, as well as in previous literature on academic texts, run counter to the relatively widely held view that a core academic lexis is equally useful for students in different disciplines. This is because lexical bundles are used differently in different academic disciplines and fields in terms of frequency, distribution, uses and functions. The results of the analysis of four-word bundles in the MFRAC seem to show that a common or core list of lexical bundles is not useful for all students in a discipline like business either. In other words, the findings of the current study may suggest that the sub-fields of business rely on different sets of lexical bundles. Twenty-two different types of bundles of the MFRAC were not identified in the business studies sub-corpus in Hyland (2008a). In addition, while the bundles in the business studies corpus in Hyland (2008a) were found to include more text-oriented bundles than research-oriented and participant-oriented bundles, the bundles in the current study include more research-oriented than in the two other broad categories. This seems to suggest that it may be useful and advisable for EAP students in a master's-level finance program and their teacher to focus on different sets of bundles than those taught to business students.

In teaching text-oriented bundles, for example, the bundles can be taught as transition phrases or linking adverbials. The Academic English teacher might organize the instruction around the four functional sub-categories (transition signals, resultative signals, structuring/relating/referring signals, framing signals). In doing that, they might start by teaching transition and resultative signals as phrases to show contrastive (e.g. *on the one hand, on the other hand, at the same time*) and addition (*in addition to the, as well as the*), and cause effect relationships (*as a result of/the*). This is because transition signals are more frequent and more commonly used in less specialized text types than professional academic research articles. That is why it may be useful to start with those as they are more likely to be encountered by master's-level finance students, and the EAP teacher should start with these before proceeding to teaching more difficult strings such as structuring bundles and framing bundles. Of course, there should be a needs analysis in order to determine what form the instruction will take. Specifically, the EAP teacher may start by pulling concordance lines of the target bundles from the MFRAC. After they could ask students to read the extracts, try to figure out what functions the target bundles serve and where in the sentences they tend to occur. Learners would do that individually before sharing their insights with other learners in groups. After the teacher could elicit responses from students regarding the function and uses of the target bundles in a whole class discussion. Following this, the teacher might ask students to do a fill-in-the-blank activity in which students would complete a series of concordance lines with correct bundles (See Appendices D and E for sample tasks).

4.5. Limitations and Suggestions for Further Research

This study sampled texts from four common journals used in the finance and real estate program at Colorado State University, even though sixteen different journals were suggested to the researcher. Sampling texts from all the sixteen journals may have resulted in a different list of bundles. That is why it is important to study bundles in texts from other master's-level finance research articles to see how consistent the findings in this study will be focusing on articles other than from applied finance. Studying bundles in master's level research articles using texts from departments where the focus is different may give a fuller picture of the uses of lexical bundles in master's level texts.

Moreover, more research into bundles used in other sub-fields of business studies is needed to see if other sub-disciplines of business rely on specific lists, and if the findings in the current study will hold true for other sub-fields of business. It would also be useful to create a more general master’s-level list of lexical bundles that would include lexical bundles by sub-discipline of business.

Another limitation relates to the size of the corpus, which may be small by today’s standards. So, research involving larger corpora is needed to reinforce or undermine the findings reported in the present study. Additionally, what to exclude from the list of all bundles originally retrieved from the corpus is based on the researcher’s intuition. Another limitation with the identification of bundles in this study, as well as in others, is that the frequency and range criteria are arbitrary, and depending on the frequency and range used to identify bundles in the same corpus, a researcher can have widely different numbers of bundles to examine. Also, some important and useful expressions might be excluded because they are not frequent enough.

CHAPTER FIVE: CONCLUSION

5.1. The purpose of the study

In this study my purpose was to investigate the extent to which the forms, functions and uses of lexical bundles used in a sub-discipline or sub-field differ from those in the discipline or field. Specifically, using master's-level finance research-articles, I was interested in knowing whether students in business studies rely on a relatively undifferentiated list of lexical bundles irrespective of their sub-disciplines. In that, the current study to some extent extends earlier studies by examining the bundles used in a sub-discipline, not only the discipline.

Another of my purposes was to provide a focused and restricted list of lexical bundles for master's-level finance students in order to maximize the return for learning. Lexical bundles were shown to consistently serve important discourse functions in various text types, including academic discourse. They were also shown to vary across different academic disciplines. That is why it this study set out to identify the bundles frequently used in master's-level finance research, and to study their forms and functions. I chose research articles because they are widely used as reading assignments to students in American universities, and they are one of the many ways by which scholarly knowledge is disseminated. I chose master's-level texts in finance because a great proportion of the students admitted into the master's in finance program are international students, and they need to learn the bundles specific to their sub-disciplines in order to write in a discipline-appropriate way.

5.2. The summary of results

Eighty different types of lexical bundles were identified, representing 4281 tokens or instances. A number of bundles found to be among the most frequent in academic prose were equally found to be the most frequent in the MFRAC. These most frequent bundles include *the value of the/a, on the other hand, (in) the case of (the), (at) the end of (the)* and *as well as the.* However, a number of bundles were found to occur specifically in the MFRAC, compared to earlier studies. Of the 80 different types of bundles, 22 were not identified in earlier studies, which seems to reinforce the view that different disciplines rely on different sets of bundles. The bundles were found to vary in their structures, even though most of them are noun and prepositional phrase fragments. These results are consistent with the findings in earlier studies such as Biber et al. (1999); Cortes (2004) and Hyland (2008a, 2008b), which showed that noun phrases and prepositional phrases are predominant in academic prose. By including far more phrase fragments than clause fragments, the results of the analysis of the MFRAC bundles reinforce the notion that academic prose, especially professional research articles, is more phrasal than clausal in structure, and this is attributed to the mainly informational purpose of academic prose (Biber & Gray, 2013). In the noun and prepositional phrase fragments, the great majority involves various forms of post-modifications, and over half of the post-modifiers involve *of*-phrases.

Other phrase fragments -- far fewer than the noun phrase- and prepositional phrase-based bundles – include verb phrase-, adjective phrase-based bundles and anticipatory *it* + adjective phrase. Clausal bundles include four *to*-clause and three *that*-clause fragments; in the other expressions category there are only two different types of bundles.

With regard to functional categories, which are the ones used in Hyland (2008a), research oriented bundles form the majority of the MFRAC bundles. In this category, the bundles in the description sub-category are the most numerous, followed by the quantification bundles, which in turn, are slightly more numerous than location bundles. The procedure sub-category includes only two bundles while the topic sub-category has none.

The second largest group include the text-oriented bundles, the great majority of bundles are formed by framing and structuring/relating/referring signals while resultative and transition signals form less than twenty five percent of bundles in this category. The very few participant-oriented bundles include only bundles in the stance features sub-category.

As regards the relationship between structures and functions, prepositional phrase fragments make up the majority of research-oriented and text-oriented bundles, with prepositional phrase fragments forming more than 80% of bundles in the text-oriented. The structural characteristics of the very few participant-oriented bundles include *to*-clause and *that*-clause fragments.

REFERENCES

Adel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of english: A lexical bundles approach. *English for Specific Purposes, 31*(2), 81-92. Retrieved from http://search.proquest.com/docview/ 1010694401? Accountid =10223/

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

Beitul, B. (2010). *Analysis of four-word bundles in published research articles written by Turkish scholars* (master's thesis). Retrieved from http://scholarworks.gsu. edu/alesl_theses/2

Biber, D. (1999). *Longman grammar of spoken and written English*. Harlow, England: Longman.

Biber, D. (2006) *University Language*: *A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263-286. Retrieved from http://search.proquest.com /docview/85656722?accountid=10223/

Biber, D. & Conrad, S. (2001). Quantitative Corpus-Based Research: Much More Than Bean

Counting. *TESOL Quarterly, 35,* 331-336. Retrieved from http://www.jstor.org/stable /3587653/

Biber, Conrad and Cortes. (2004). *If you look at…*: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25,* 371-405. Retrieved from applij.oxfordjournals.org/ content/25/3/371.full.pdf+html/

Biber, D. and Gray, B. (2013). Nominalizing the verb phrase in academic science writing. In Aarts, A., Close, J., Leech, G. and Wallis, S. (Eds.), *The verb phrase in English* (pp. 99-132). Cambridge University Press.

Biber, D., Grieve, J., & Iberri-Shea, G. (2009). Noun phrase modification. In G. Rohdenburg, & J. Schlüter (Eds.), *One language, two grammars? Differences between British and American English.* (pp. 182-193). (Studies in English language). Cambridge University Press

Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30-49. Retrieved from http://search.proquest.com/ docview/753820628?accountid=10223/

Conrad, S., & Biber, D. (2004). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica. 20,* 56-71. Retrieved from http://search.proquest.com/ docview/85695901?accountid=10223/

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397-423. Retrieved from proquest.com/docview/85610836?accountid=10223/

Hyland, K. (2008a). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes. 27,* 4-21. Retrieved from http://www.sciencedirect.com/

Hyland, K. (2008b). Academic clusters: Text patterning in published and post-graduate writing. *International Journal of Applied Linguistics. 18,* 41-62. Retrieved from http://www.researchgate.net/publication/229879061/

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics, 32,* 150-169. Retrieved from http://search.proquest.com /docview/1541993002? accountid=10223/

Jalali, H., Rasekh, A. E., & Rizi, M. T. (2009). Anticipatory 'it' lexical bundles: A comparative study of student and published writing in applied linguistics. *Iranian Journal of Language Studies, 3*(2), 177-194. Retrieved from http://search.proquest.com/ docview/85697332?accountid=10223/

Matthews, D., Lieven, E., Theakston, A.L., Tomasello, M., (2005). The role of frequency in the acquisition of English word order. *Cognitive Development, 20,* 121–136. Retrieved from http://www.eva.mpg.de/psycho/pdf/Publications_2005_PDF/The_role_of_ frequency_in_the _05.pdf

McEnery, T., Xiao, R. (2006). *Corpus-based language studies: An advanced resource book.*

London: Routledge.

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purpose, 25,* 235-256. Retrieved from www.sciencedirect.com/science/article/ pii/S0889490605000360

Myles, F., Hooper, J., & Mitchell, R. (1998). Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning*, *48*, 323-364. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/0023-8333.00045/epdf

Myles, F. 2005. Interlanguage corpora and second language acquisition research. Second *Language Research, 21*, 373–391. Retrived from http://slr.sagepub.com/content/21/4/373.refs

Nattinger, J. R., DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford [England]: Oxford University Press.

Ortega, L. (2009). *Understanding second language acquisition*. London: Hodder Education.

Perez-Llantada. (2014). Formulaic language in l1 and l2 expert academic writing: Convergent and divergent usage. *Journal of English for academic purposes*, 14(Jun), 84-94. Retrieved from http://sciencedirect.com/science/article/pii/S1475158514000162/

Peters, AM. (1983). Units of language acquisition. Cambridge, UK: Cambridge University Press.

Reppen, R. (2010). *Using corpora in the language classroom.* New York: Cambridge University Press.

Rica-Peromingo, J. P. (2010). Corpus analysis and phraseology: Transfer of multi-word units. *Linguistics and the Human Sciences, 6*(1-3), 321-343. Retrieved from http://search. proquest.com/docview/1430171777?accountid=10223/

Robinson, M. S., Stoller, F. L. and Jones, J. K. (2008). Using the ACS journals search to validate assumptions about writing in chemistry and improve chemistry writing instruction. *Journal of Chemical Education, 85,* 650-654. Retrieved from www.JCE.DivCHED.org/

Schmitt, N. & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use (Language Learning & Language Teaching)* (pp. 1-19). Amsterdam: John Benjamins Publishing Company.

Simpson-Vlach & Ellis (2010). An academic formulas list: New methods in phraseology Research. *Applied Linguistics. 31*, 487–512. Retrieved from http://applij.oxfordjournals .org/content/31/4/487.full.pdf+html/

Tomasello, M., (2003). Constructing a Language: A Usage-based Theory of Language Acquisition. Harvard University Press, Cambridge, MA.

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

Wray, A. (2008) *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.

APPENDIX A:

The complete list of journals suggested by the 6 professors in the Finance and Real Estate Department at Colorado State University, with the journals sampled in bold.

1. The Journal of Applied Corporate Finance

2. **Financial Analysts Journal**

3. The Journal of Real Estate Research

4. Financial Services Review

5. The Journal of Money Credit and Banking

6. **The Journal of Derivatives**

7. The Journal of Insurance Issues and Practice

8. **The Journal of Portfolio Management**

9. The Journal of Alternative Investments

10. The Journal of Fixed Income

11. The Journal of Index Investing

12. The Journal of Finance

13. The Journal of Private Equity

14. The Journal of Structured Finance

15. The Journal of Trading

16. The Journal of Wealth Management

17. **The Journal of Corporate Finance**

APPENDIX B:

The complete list of the removed 20 bundles not occurring in at least 3 different journals.

Freq.	Range	Bundles
60	5	and non family firms
56	5	family and non family
51	24	panel a of table
46	24	panel b of table
41	24	the journal of portfolio
40	23	journal of portfolio management
37	24	et al find that
36	14	are less likely to
35	14	a dummy variable that
35	13	la porta et al
32	13	the investor x s
32	17	the results in table
31	16	et al show that
28	15	we also find that
27	12	of table shows that
27	5	value of the option
27	14	we found that the
26	17	are likely to be
25	11	dummy variable that equals

25	14	results are robust to

APPENDIX C:

The complete list of the merged overlaps.

Freq.	Range	Bundles
114	51	in the case of
26	15	the case of the
103	49	at the end of
79	39	the end of the
101	40	on the basis of
30	19	the basis of the
88	34	are more likely to
35	26	more likely to be
83	45	in the united states
26	15	the united states and
62	36	is consistent with the
39	28	are consistent with the
47	31	the fact that the
25	19	to the fact that
47	27	the s p index
38	20	of the s p
44	22	in the presence of
38	10	the presence of a
43	27	as a function of

26	20	a function of the
34	25	the results of the
25	21	the results for the
29	13	is the ratio of
25	19	the ratio of the
25	19	the ratio of the
30	24	one of the most
27	22	is one of the

APPENDIX D:

Task 1: Read and examine the concordance lines and try to figure out what functions the two underlined phrases serve, and where in the sentences they tend to occur. The two target expressions are in two sets of concordance lines. After you have examined the concordance lines, share what you have come up with the students to the left and to the right.

with a more enlightened regulatory environment. On the other hand, even if governments adopt

like an interest-sensitive cash flow. If, on the other hand, he goes to sleep

extreme cases, multi-stakeholder financial fights. On the other hand, the cure must also

idiosyncratic volatility and expected return. On the other hand, if global idiosyncratic volatility

the measurement of a fund's alpha. On the other hand, Back, Kapadia, and Ostdiek

instead of participating in share repurchase. On the other hand, institutional investors, who are

between efficiency and political objectives. On the other hand, Bai et al. (2006) suggest

controlled firms operating in its region. On the other hand, the central government may

relationship between Leverage and labor intensity. On the other hand, since most loans of

by the government or government agencies. On the other hand, the disciplinary role of

on the difference in Tobin's q. On the other hand, local government is concerned

development and remains trivial (Alan and Shen, 2012). On the other hand, private firms are still

are less likely to tunnel resources out. On the other hand, firms with tangible assets

performance. The negative coefficients on bank loans, on the other hand, are largely reduced or

non-tradable shares; individual shareholders, on the other hand, hold minority stakes. Therefore

tradable A-shares for a quicker resolution. On the other hand, foreign institutional investors

portfolio allocations are reported in Table 5. In addition to the industry-specific human capital investor perception is certainly of value, in addition to the measurement of risk in FAJ26.txt

similar to diversification across various stocks. In addition to the prevailing performance

the standard deviation of the underlying. In addition to the previous issues, maximum drawbacks

options, collect this volatility risk premium in addition to the equity risk premium earned

discussed. Regarding risky asset returns (and in addition to the S&P 500 return), these

with annual crystallization as the baseline. In addition to the increase in fee load

Annual Meeting in Lugano, Switzerland. In addition to the management fee, we also

out 70% above collateral and margin requirements. In addition to the use of leverage, the

model by adding size and value factors in addition to the market risk

influenced by their availability of internal funds. In addition to the sensitivity of investment to

suppliers are fundamentally different types of firms. In addition to the other factors, the large

firm's future investment opportunities. Finally, in addition to the industry dummy variables, we

All of our results are similar. In addition to the level model used in five quarters, inclusive of the

important determinant of the composition of board in addition to the commonly-used framework

ROA is 0.039 with a standard deviation of 0.16. In addition to the variables discussed above,

APPENDIX E:

Task 2: Complete the following extracts by filling in the blanks with the appropriate transitional phrases from the four phrases in parentheses (at the same time, on the one hand, on the other hand, with respect to the). Each of the four transitional phrases should be used twice.

1. _______________, cross-sectional regressions are clearly the methodology of choice in applied risk factor models, whereas time-series regressions against MRF, SMB, HML, and, sometimes, UMD (up minus down momentum portfolios) are usually associated with the measurement of a fund's alpha. _______________, Back, Kapadia, and Ostdiek (2013) argued that Fama-Macbeth cross-sectional regressions yield purer factor returns than do portfolio sorts and thus do a better job of measuring alpha.

2. ________________ 5% expected shortfall of standardized portfolio returns, Panel B of Table 3shows that all momentum portfolios exhibit expected shortfalls that are significantly lower than that of a standard normal distribution (which has a 5% expected shortfall of -2.06).

3. When the trade is uncollateralized (Figure 7), as the price of WTI increases, the swap becomes more valuable for us and hence the credit exposure increases. But _______________, the counterparty default probability decreases, and therefore, this trade has right-way risk.

4. There may be some confusion, however, ________________ standard error around excess returns from the moment strategy, ________________, and the IR as an estimated parameter itself, on the other. Therefore, Table 2 also includes t-statistics that are based on Jobson and Korkie 22.

5. Large firms might have higher market share and/or greater market power, which might positively impact firm performance. However, _______________, large firms might experience a greater degree of government bureaucracy or other organizational inefficiencies that are detrimental to firm performance (Sun and Tong, 2003).

6. If prices are too high, firms find issuing new shares attractive, but those shares will offer a poor return, eroding the equity risk premium. If prices are too low, _______________, firms find issuing new shares unattractive, hampering economic growth. Active investing ensures an efficient allocation of capital, which, crucially, is a positive-sum game.