Lexical Bundles within English for Academic Purposes Written

Teaching Materials: A Canadian Context

by

Julia Moynié

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

Master of Arts

in

Applied Linguistics & Discourse Studies

Carleton University
Ottawa, Ontario

Abstract

Research surrounding lexical bundles in English for Academic Purposes (EAP) is fundamental to the academic success of students pursuing postsecondary education. Recent studies suggest that awareness of the presence and function of lexical bundles in academic English is beneficial to students and teachers, and leads to a deeper understanding of the construction of discourse (Biber, 2004). Using corpus linguistics, this study presents the degree to which students of an EAP program at a mid-sized Canadian university encounter four-word lexical bundles in their written teaching materials. The study uses functional taxonomy to classify the lexical bundles of the Moynié corpus I created. The results indicate the need for further research to assess the frequency of lexical bundles amongst all the registers that students come across in their academic studies. This is to ensure the students achieve fluency in English, and can competently understand, recognize and utilize lexical bundles.

Keywords: lexical bundles, discourse functions, functional taxonomy, corpus linguistics, English for Academic Purposes (EAP)

Acknowledgements

I would like to express my sincerest gratitude to Dr. Marie-Odile Junker and Dr. Geoffrey Pinchbeck, whose support were instrumental in this process. Their wisdom, advice and guidance is what made this possible. I would like to thank Judy Senecal, whose assistance at the beginning of my journey guided my thesis direction, as well as the teachers who supported this research by sharing their materials.

I would also like to thank my internal examiner, Dr. Guillaume Gentil, and my external examiner, Dr. Melanie Gonzalez, for their time, their questions, and advice regarding my research.

To Erin Kelleher, thank you for keeping me grounded and laughing this year.

A special thank you to Alex Ross, my classmate, my friend, my lifeboat. We did it!

Dedication

A ma maman, Dadi Moynié, merci pour tout.

Table of Contents

List of Tables

List of Illustrations

List of Appendix

1 Chapter: Introduction

Educational programs of English for Academic Purposes (EAP) are growing increasingly and constantly evolving in Canada. Coutinho (2018) reports that Immigration Refugees and Citizenship Canada (IRCC) estimate that by the year 2022, more than 450,000 international students will be studying in Canada. Schools and universities are finding that more and more international students are registering for their various EAP programs with the aim to attain academic fluency in English so that they may pursue a postsecondary degree in English. This growing interest demonstrates the importance of EAP programs in the field of education. Applied linguists have had a dominant role in the creation of these EAP programs and continue to further research the content of the program materials. This research aims to improve the way English is taught to students, by making teachers and students more aware of the presence and functions of lexical bundles in academic discourse, so that they may perform academically in a competent manner. The rationale for this research came from a need to study the surrounding student community of my university, and seeing the potentially direct applicability of my study's findings. Being able to acquire and study data from the context of my own university is unique, and a privilege. Further research of lexical bundles may allow language acquisition to be more customizable to the desired register of the learners.

Described as "the most frequent recurring lexical sequences" (p. 183), Biber et al. (1999) paved the way for a multitude of studies in the field of lexical bundles. Wood and Appel (2014) state "an explicit awareness of the importance of lexical bundles in university textbook language might help EAP materials designers and teachers to craft

materials which can include a focus on relevant bundles" (p. 9). Lexical bundles are of interest to learners and teachers of English because to succeed in academic writing, students need to be able to identify and use them correctly in order to replicate the appropriate language in the academic genre (Sykes 2017). The focus on lexical bundles began with Biber et al. (1999) who performed corpora analysis of the *Longman Spoken and Written English* (LSWE) Corpus. Rooted in formulaic language, lexical bundles are pre-fabricated chunks of language that are found across a variety of registers in academic English. Some of the research in this field consists of identifying just how present these lexical bundles are within academic discourse, and aims to distinguish between them according to functions of purpose and use, all of which is of utter importance to both teachers and students of English.

In the current study, I focused solely on four-word lexical bundles in order to replicate the same parameters set by Biber (2004). Four-word lexical bundles are more often the focus since they are not as frequent as three-word lexical bundles, nor are they as rare as five-word lexical bundles. I took a careful look at the study conducted by Biber (2004) on the topic of lexical bundles in university teaching and textbooks. The research I performed in the current study aims to replicate the work of Biber (2004), but with a sole focus on the information and results that came out of the data for the textbook register. I examined data in the form of course packs collected from teachers of an EAP program at

a Canadian University. The purpose of my study is to answer the two following research questions:

1. How frequent are 4-grams, as identified by Biber (2004), in EAP written teaching materials?

2. What is the function of lexical bundles in written pedagogical materials used in an EAP program at a mid-sized Canadian university?

      EAP program materials, in the form of course packs, were used to create a university level corpus. This study uses corpus linguistic methods and software to compile an original corpus from the data collected. The use of the text corpus analysis software AntConc facilitates the analysis of the corpus, and extracts all lexical bundles from the texts. AntConc is able to identify the frequency of all lexical bundles found, and indicate in how many texts these bundles appear, which is also referred to as the *range*. A detailed comparison between the Moynié Lexical Bundle List (MLBL) and the Biber Textbook Lexical Bundle List (BTLBL) in chapter 4 outlines differences between lexical bundle frequency in the two corpora. These results indicate that the type of functions the lexical bundles serve is directly linked to the theme of the content being analyzed. They could be useful to academic discourse content developers so they are made aware of what to include, teachers will benefit from this information because it will allow them to teach important lexical bundles explicitly, and students will acquire a new set of vocabulary to use in their academic discourse. The rest of this paper is organized as follows: In the second chapter, I discuss previous literature on the topic and aim to assess and identify the research gaps. I describe the methodology used to conduct the present study in the third chapter. In the fourth chapter, I present the analyses of both corpora, outline the

results, and discuss their implications. I use the fifth chapter to conclude this thesis by presenting the pedagogical implications of the study, by identifying the limitations involved, and to suggest further research directions. My research aims to contribute to a broader understanding of teaching materials in EAP programs since they are developing and becoming quite in-demand in Canada. My research is unique in that the data collected is from course packs, versus textbooks, which seems to be what EAP programs are moving towards by implementing a content-based approach to teaching their students. The goal is to use this research to improve the quality of EAP teaching across Canada, and to determine if the written teaching materials used in this Canadian EAP program is comparable to that of English university textbooks in various fields of study. The purpose of using Biber's (2004) functional taxonomy, a framework developed to categorize the function of lexical bundles in language, is to analyze my findings and advance research on lexical bundles in order to provide teachers and students with a tool they can use to perform their own analysis.

2   Chapter: Literature Review

2.1   Introduction

In the discussion of vocabulary within the context of language teaching, Schmitt (2012) argues that "there is not necessarily a one-to-one correspondence between a meaning and a single word" (p. 2). Very often, in English at least, meanings are represented by multiple words. To handle these multiword units, the term *lexeme* (also *lexical unit* or *lexical item*) was coined. These three interchangeable terms are all defined as "an item that functions as a single meaning unit, regardless of the number of words it contains" (Schmitt, 2012, p. 2). Lexical bundles belong to the formulaic language spectrum, in which a multitude of terms exist to categorize and describe various chunks of pre-fabricated language. Though some of these terms may be considered interchangeable, lexical bundles are identified with a specific list of criteria and their study holds a particular purpose for linguists utilizing them. This literature review aims to acknowledge and bring forth the research of major linguists whose focus and passion for lexical bundles has allowed to make improvements in the teaching of English as a second language.

The study of lexical bundles emerged with Biber et al. (1999) while performing computational analysis of text corpora to compare the two dominant registers: academic prose and conversation, which are both represented in the *Longman Spoken and Written English* (LSWE) Corpus. This corpus exceeds 40 million words taken from four major registers: conversation, news, academic prose and fiction in British and American English, and aims to interpret the quantitative findings of grammatical functions of English.

The term lexical bundle is often associated or mistaken with other key concepts such as idioms and collocations. I will briefly define these categories below. Biber et al. (1999) describes lexical bundles as "the most frequent recurring lexical sequences" (p. 183) that are neither fixed expressions nor structural units such as *as well as* or *it is important to*. Lexical bundles are composed of three or more words and are categorized according to the registers in which they appear, they are much more frequent than idioms. Biber, a leading expert on lexical bundles, uses functional taxonomy to classify lexical bundles, and therefore demonstrate their unique uses within language.

Idioms are meaningful and opaque expressions. In other words, the meaning of their individual component words does not equate to the general meaning of the expression. A single verb can also often replace an idiom and capture its entire meaning. For example, the idiom *see eye to eye* can be replaced with the verb *agree*. Idioms are also quite rare within the context of natural speech or writing (Biber, 2004), which indicates that lexical bundles are more valuable to study and extract from registers since learners and users of language come across them more often.

Collocations are made up of two words that retain their individual meanings and co-occur more often than not, with particular collocates. For example, the collocation *strong coffee* describes a drink made from dark roast beans. The word *strong* in this context cannot be replaced with synonyms such as *tough* or *heavy* and retain the same meaning. In comparison, lexical bundles such as *in the next chapter* or *it is important to* cannot be categorized as either idioms or collocations since they have no opaque meaning, nor do they necessitate specific partner words.

I will be exploring the literature on lexical bundles to inform my research into their presence in the written teaching materials of a Canadian English for Academic Purposes (EAP) program. In this chapter, I provide a brief history of lexical bundles, present the main contributors and their discoveries and outline the research gaps and questions that still need to be addressed by the applied linguistics community. The following sections will inform these research gaps: (1) finding the most accurate definition and method of identification of fixed multi-word units, and (2) analysing their discourse functions.

2.2   Characteristics of Lexical Bundles

It is important to identify distinguishing factors of lexical bundles in order to analyse them independently from other chunks of language. Biber (2004) presents a multitude of ways scholars have attempted to describe these chunks of language without clearly setting definitions that are agreed upon. The criterion of frequency is prevalent in the research surrounding lexical bundles. Hyland and Tse (2007) state that "frequency patterns reveal clear disciplinary preferences as routine uses take on constancy of convention" (p. 246). This indicates the frequency of language forms such as lexical bundles inform the register, the context, the audience, and determines its relative importance. The frequency criterion is crucial in that it informs us how often we use or encounter these chunks of language. If this frequency criterion is found to be important, it would likely be a good candidate for teaching. The context in which these lexical bundles are found is also key to figuring out how to approach them from both a research and a pedagogical perspective. For the purpose of this research, due to time constraints and the nature of the work, only written texts have been analyzed, however, a comparison of registers, especially between spoken and written texts would allow a more well-rounded

image of the use of lexical bundles to be drawn. Similarly to Biber et al. (2004), this research was conducted using frequency as primary criterion in the investigation of the functions and presence of lexical bundles. More specifically, this research examined four-word lexical bundles within EAP written teaching materials used at a Canadian university.

The frequency-driven approach to research on chunks of language began with Salem (1987) who explored French government documents through corpus analysis. The first analysis of English sequences is attributed to Altenberg (1998, Altenberg and Eeg-Olofsson 1990) who based their research on the London-Lund Corpus. The work of Biber et al. (1999, see also Biber & Conrad 1999) in the *Longman Grammar of Spoken and Written English* was an extension of Butler's (1997) work on recurrent sequences found in a large-scale corpus of Spanish texts.

The study conducted by Biber (2004) focused on the analysis of use of lexical bundles within university classes in both classroom teaching and textbooks, the latter of which is of specific interest in the research outlined in my thesis. Biber (2004) denotes that textbooks and classroom teaching are each classified as separate registers, an important distinction since my thesis solely focuses on textbooks. Biber's (2004) was strongly influenced by the one conducted by DeCarrico and Nattinger (1988), which focused on lexical phrases. Instead of using perceptual salience as his basis for analysis, which was described by Tallat (2011) as "the information that captures the attention of the individual from a given situation or stimulus" (p. 7), Biber (2004) employed the frequency criteria to analyse the patterns of use of lexical bundles. Since academic prose and textbooks are close in terms of purpose they serve, they can be analyzed similarly for

their patterns of use. Textbooks are used in teaching when the goal is to present materials that are planned and have a clear focus and target audience in mind. There is more control of forms as compared to that of spoken registers. Biber et al (2004) utilized the T2K-SWAL Corpus (TOEFL 2000 Spoken and Written Academic Language Corpus, see Biber et al. 2002, 2004) as their source of data, which was created to encompass both written and spoken registers that university students in the United States encounter in their academic careers. The T2K-SWAL Corpus is composed of registers and texts from classroom teaching and textbooks from six dominant academic domains including: Education, Business, Humanities, Social Science, Natural Science and Engineering. These texts were taken from four academic institutions: Iowa State University, Northern Arizona University, Georgia State University, and California State University at Sacramento, all ranging from three levels: graduate, upper division undergraduate, and lower division undergraduate. Biber et al. (2004)'s study included 87 textbooks, which made up 760,600 words. The composition of this sub-corpus will serve as a baseline to compare the corpus within the present study. It is recognized that textbooks are most often written for students as their audience, and this is the same for the course pack corpus that I have also used in my study (see below). Since the articles and books included in the register labelled as academic prose corpus were created for other professionals, they will be excluded from my research, in order to compare lateral components.

Biber et al. (1999) defines lexical bundles in the simplest fashion as "the most frequent recurring lexical sequences in a register" (p. 376). Frequency data can thus be used to determine the relative rate at which chunks of language appear in any given

register, in this case written texts. Given that such sequences are often without an idiomatic meaning, and are prefabricated, it would make sense to add to Biber et al. (1999)'s definition of lexical bundles as follows: lexical bundles consist of the most frequent recurring lexical sequences absent of idiomatic meaning in any given register or genre. The lexical bundles that the study focuses on originate from academic course packs for learners of English as an academic language. Lexical bundles found in academic texts usually bridge two phrases together (Biber, 2004), meaning they are formed by two structural units that are composed of a clause and the first elements of the second structural unit. This structure helps identify their purpose in language and guides us on how they could be defined for the purposes of language pedagogy.

For the most part, lexical bundles have been studied within the context of academic teaching materials, such as textbooks, in various fields of study. This study is looking specifically at lexical bundles within written teaching materials presented to students whose first language is not English, and for students who are beginning their academic, university-level, career. Since lexical bundles have been found to be of importance and to have a noticeable presence in textbooks, should their presence not also be sizeable to students in the language learning process they are, or will be, studying in? EAP programs specialize in preparing adults whose first language is not English, to perform and succeed academically in English-medium university programs. For most students, these EAP classes are a mandatory part of their university program, and success in these classes qualifies them to continue their education in Canada and in English. The ELSA department is unique in that it uses course packs instead of textbooks to teach their students. These course packs include various articles from different authors, all

surrounding one theme. This type of teaching material is fairly new and uses a content-based instruction approach. Content-Based Instruction (CBI) has developed mainly in the context of second language teaching due to its student-centered methods, but has begun to branch out to other educational contexts as well, particularly to EAP. Support for CBI has increased because of the clear benefits and success it brings to student learning. CBI stems from Krashen's comprehensible input hypothesis, which claims that in order to acquire language, the focus should first be on understanding the meaning, and secondarily on inferring language structure. Krashen (1982) states that acquisition is done by "understanding language that contains structure beyond our current level of competence ($i + 1$)" (p. 21). According to Krashen, we acquire, we do not learn.

The textbooks used as samples by Biber (2004) do not contain a lot of lexical bundles, compared to the frequency and presence in which they are utilized within classroom teaching. Should the language in both contexts not match somewhat in order to present students with a coherent example of how to speak and write academically?

2.3   Functional Taxonomy of Lexical Bundles

Biber et al. (2004) present a functional taxonomy that indicates that lexical bundles have three identified functions: (1) stance expressions (SE), (2) discourse organizers (DO), and (3) referential expressions (RE). This scheme allows us to organize lexical bundles according to their purpose in language. This knowledge allows us to teach them more effectively.

The first type of lexical bundles, **stance expressions**, articulate the position one might hold regarding a particular situation. For example, the lexical bundle *I don't know if* expresses a weakening opinion towards something that has been said or written. The

11

second functional type of lexical bundles, **discourse organizers**, comment on links that exist between past and future discourse. For example, the lexical bundle *in this chapter we* refers to the knowledge that will be presented and most often relates it to the content of the previous chapter as a way to build on said knowledge. The third type of lexical bundles, **referential expressions**, refer directly to an entity, concrete or imaginary, in order to discuss it as a whole or bring attention to a part of it. For example, the lexical bundle *is one of the* is used to single out one part of a whole. All three functions of lexical bundles can be further divided into sub-categories. Table 3 outlines the various lexical bundles found in the data collected by Biber (2004), and groups it according to these functions, and subcategories, as well as across the various registers analyzed. All lexical bundles discussed within the present study were found within course packs utilized for academic purposes. The six course packs that make up the raw data of this study cover three different levels of instruction of a Canadian university EAP program. The course packs include instructional language on how to write in academic discourse, exercises that provide vocabulary and grammar practice, texts and chapters on various topics such as politics, history, or science, and examples, as well as guidelines, on how to format references and essays. These course packs replace textbooks in this particular EAP program. They contain a variety of registers so as to expose the students to different information, but also to gather most of the course materials in one place. A replication of Biber's (2004) Table 3 can be found in the Results section of this paper.

Stance bundles can be categorized as either epistemic stances (ES) or attitude/modality stances (A/MS). Epistemic stance bundles are characterized by Jalali (2017) as referring to "meanings of certainty, doubt, actuality, and definiteness, as well as

indicators of the source of perspective of knowledge (e.g. attributing an idea to particular reference)" (p. 31), while attitudinal/modality stance bundles are made up of "attitudes and evaluations, as well as personal feelings or emotions" (p. 31). Stance bundles can also be classified as personal (Per.), stemming directly from the speaker or writer (e.g. *I want you to*), or impersonal (Imp.), without a direct link to the speaker or writer (e.g. *it is possible to*). Similarly, attitudinal/modality stance bundles can be personal or impersonal, and can be broken down further to fit in one of four subcategories: *desire* (D), expressing a personal wish, *obligation/directive* (O/D), indicating an action that needs to be completed directed towards a second person pronoun, *intention/prediction* (I/P), indicating a personal future action, and *ability* (A), an impersonal expression to outline outcomes desired.

Discourse organizing bundles are used for two purposes: to *introduce* or *focus* (INT/F) on a particular topic, and the *elaboration* or *clarification* (E/C) of a topic. Simpson (2004) brought forth the importance of discourse organizing bundles due to their usefulness to "summarize, sequence, and focus information" (p. 40), a skill very much desired in academia. Topic introduction bundles, found mainly in classroom teaching in teacher-to-student interactions, alert the audience that a new topic will be presented (e.g. *I want to discuss*), and depending on their structure, will invite collaboration, consideration, participation, attention, or provide instruction on the new topic. These types of bundles can be both personal and impersonal.

Referential bundles, which Conrad (2005) defines as making "direct reference to physical or abstract entities, or to the textual context" (p. 67), carry out four different functions: *identification/focus* (I/F), *imprecision* (I), *specifying attributes* (SA), and

time/place/text reference (T/P/TR). The *identification/focus* bundles, mostly present in classroom teaching, are used to narrow down the focus of the noun phrase which follows the bundle. They may help in summarizing or emphasizing the key point being made. These bundles can also be used to summarize or introduce the discussion to follow. *Imprecision bundles* can be subcategorized into two distinct functions: to exclaim inexactitude or the presence of further references, such as *or something like that*. Bundles *specifying attributes* detail amounts and quantities following the head noun of a phrase. The size and form can also be detailed using these bundles, such as *the size of the*. Places, times, and locations are referred to in text using *time/place/text reference bundles*. These bundles help highlight the context in which the language is being studied or found. For example, if the bundle *in the United States* is recurrent in a text, one could assume that the text was written in the United States, and intended for a United States' audience. Text reference bundles are present mostly within text registers since they refer to tables, charts and figures that would rarely be used in speech.

The preceding information outlines the various functions of lexical bundles that exist. However, these bundles and their correlating functions are not found evenly across registers. Biber's (2004) data indicate that stance bundles are mostly found in classroom teaching and conversation, discourse organizing bundles are mostly found in classroom teaching and less in conversation, and referential bundles are prominent in classroom teaching, while rarer in textbooks and academic prose. All forms of lexical bundles are found to have a strong presence in classroom teaching due to the multi-functional purpose of the classroom which can encompass all registers. The argument could be made that this language use, and therefore this variety of lexical bundles, should be

reflected in the textbooks from which EAP students study and learn academic English. An increased correspondence between pedagogical and target language could facilitate learning, and would enable students to better imitate their teacher's language as the example and model they are learning from. The patterns identified by Biber (2004) in figure 6 of the "direct association between form and function of lexical bundles" (p. 398), could help guide teachers on how and when to teach lexical bundles to students. This form of explicit instruction could provide greater understanding of how to manipulate language. Biber (2004) notes that there exists a "complex interaction between structural form, discourse function, and the typical purposes and situational characteristics of registers" (p. 398).

2.4   Use of Lexical Bundles in Teaching

Coxhead & Byrd (2007) have proposed three main reasons as to why academic sequences, in this case lexical bundles, are important to writers and speakers of a language. The first reason is the fact that lexical bundles are repetitive and provides users, in our context ESL students, with sets of words that are prefabricated and that they can be used to construct language. The second reason is that the use of lexical bundles accounts for fluency in a language, and holds a certain accreditation and legitimacy in the discipline one writes or speaks in. The third reason is that the classification of these lexical bundles expose their grammatical structure and leads to their function in language. Finally, it has become apparent throughout my research and this study that lexical bundles facilitate linguistic inclusion in the community and context they are used in. All the reasons noted above help outline why lexical bundles should be an explicit part of English language teaching, particularly in EAP, in order to make sure students are

linguistically prepared for the language they will no doubt encounter in their future academic careers. Hyland (2008) insists "academic writing draws on a much larger stock of prefabricated phrases than either news or fiction" (p. 44). Considering students are consistently presented with academic writing, it seems rather crucial for them to have a true understanding of what they are reading, in order for them to master it and go on using it in their own speech or writing.

Biber and Barbieri (2007) pointed out the infrequency of lexical bundles in academic written registers, specifically in textbooks and academic prose, which is in direct contrast with the other academic registers who are also important to EAP students in their preparations for post-secondary education. Their explanation outlines the informational purpose of these registers. Written teaching materials are evolving into less traditional textbook types, the EAP program at the Canadian university from the current study is a great example. The students of the EAP program are enrolled in university, and are expected to eventually reach the ability to write academically and at a university-level. The change of written teaching materials they are presented with should reflect other materials they will encounter in non-EAP classes. Lexical bundles should be explicitly taught to them if it is expected to have students include them in their writing. Halliday's (1994) linguistic macro functions have categorized lexical bundles in a broader fashion, one that may be helpful to teachers without a specialization in applied linguistics and lexical bundles in particular. Halliday (1994) highlighted three main categories: *research-oriented* (ideational), helping with the structure of information, *text-oriented* (textual), looking at the text's organization, and *participant-oriented* (interpersonal), focusing on who is involved with the text.

Researchers, such as Meunier & Granger (2008), have performed studies that can now inform the pedagogy behind teaching lexical bundles. Hyland (2012) comments that "the absence of such clusters [lexical bundles] reveal the lack of fluency of a novice or newcomer to that community" (p. 165). In the Canadian context, newcomers can be subject to discrimination or isolation due to their linguistic gaps and lack of flow. My research takes a particular look at EAP classrooms, which are made up of students who are newcomers to Canada, and are depending on their success in these classes to move forward with their academic careers, meaning the stakes are high. Hyland (2012) supports my position by stating "it is possible, then, for bundles to be taught in EAP classrooms, although to date very little by way of practical applications has been published" (p. 165). My research thus aims to outline the presence of lexical bundles within EAP written teaching materials in the Canadian context of the EAP program from the current study.

3   Chapter: Methodology

The methodology for this study was guided by the following research questions:

1.   How frequent are 4-grams, as identified by Biber (2004), in EAP written teaching

materials?

2.   What is the function of lexical bundles in written pedagogical materials used in an

EAP program at a mid-sized Canadian university?

In order to answer these questions, I utilized a corpus-based approach to examine

the written pedagogical materials within an academic English context to identify the

frequency of lexical bundles. A corpus-based approach to research, according to Conrad

(2004) "has allowed us to understand patterns of variation more comprehensively" and

can also "describe variation in the use of a specific feature of language, rather than to

characterize a variety" (p. 69). English-language corpora are numerous, and vary in terms

of the registers, fields, and genres. Specifically looking at the presence of lexical bundles

in written teaching materials only, I chose to create my own corpus based on the course

packs used in the EAP program classes of a Canadian university. The decision to build

my own corpus was due to my interest in the process, the lack of accessibility to corpora,

and the fact that I had access to an EAP program whose method of teaching with course

packs was quite different than the traditional use of textbooks. I wanted to explore these

course packs further, and try to find out how close the linguistic style, structure and

content was to textbooks. By creating this corpus and conducting this study, I aim to

contribute to the conversation surrounding the presence of lexical bundles in academic

discourse. The aim is to identify the presence of four-word lexical bundles in the Moynié

corpus, analyze their discourse functions, and make pedagogical recommendations regarding which ones to teach explicitly to students in EAP programs.

3.1   Selecting a Corpus

A major research gap within the field of formulaic language was identified by Wood and Appel (2014) in their study relating to multiword constructions in EAP, business and engineering textbooks for first-year university students. These researchers outline the need for further research stating:

> An explicit awareness of the importance of lexical bundles in university textbook language might help EAP materials designers and teachers craft materials which can include a focus on relevant bundles... if EAP materials are to be considered authentic, they need to present language which typifies that which is commonly encountered by students in the academy, and materials developers need to attend to key features of that language. (p. 9)

Based on the research gap outlined above, I began to explore the variety of research around lexical bundles, where Biber stood out as one of the leading experts. As outlined in Chapter 2, my literature review, previous research focused mainly on academic discourse and textbooks at the university level. The study conducted by Biber (2004) has driven this present study due to its explicit focus on four-word lexical bundles, which narrows down the results, its detailed account of the functional taxonomy framework, including purpose and use, and finally its broad sampling of data across multiple fields of study. The current study aims to replicate the work done by Biber (2004) while analyzing a different and smaller data set. While searching through numerous corpora, I found many were not accessible, nor focused on the study of lexical

bundles. Other corpora included data from too many different registers, which did not suit the scope of the study I wanted to conduct. That is when I realized it would be best to create my own corpus.

3.2   Creating a Corpus & Data Collection

Considering my surrounding academic environment, I chose to focus on the EAP program at a mid-sized Canadian university as a way of having access to data that was from my own community, that was authentic, and with which I could relate to and observe who it would directly affect. Faculty members in this program are in charge of teaching the EAP Foundation Program. These ESL professionals utilize the most up to date and innovative teaching practices to educate their students. I contacted the department and set up meetings with the faculty members. I explained the basis of my research goals, and was generously provided with a wide sample of teaching materials the instructors had designed and used for their various classes. This EAP program offers three levels of classes: the course code 1300 is for beginner-level students, the course code 1500 is for intermediate-level students, and the course code 1900 is for advance-level students. The written teaching materials I was given came in various forms: course packs, books (fiction and non-fiction), as well as course binders, which included instructional handouts, quizzes, tests, assignment sheets, and exams. After careful study of these materials, I narrowed them down to the course packs given by one teacher, a senior instructor from the EAP program. I chose to focus on these materials alone because they were the most recent. She was able to give me access to course packs from all three levels offered by the department, as well as from different school terms and years. This variety allowed for a representative data set from which to build my corpus.

The creation of the Moynié corpus began with the complete digitization of all six course packs collected. I started by unbinding them, putting them through the scanner, then enhancing the scans using the software Adobe Acrobat Pro, and applying the function *recognize text*, which uses Optical Character Recognition (OCR) to capture the text digitally. Afterwards, I needed to export the searchable PDFs and transform them into Word documents, again using Adobe Acrobat Pro, in order to be able to perform error correction. Two types of error correction were performed. The first was a thorough spell check of each of the six documents in three different varieties of English: Canadian English, American English and British English. The reason for performing the spellcheck using those varieties of English was to make sure to catch all possible spelling variations of words that could be found in a lexical bundle, so that future analysis performed using software would recognize these variations as viable sequences. Table 1 depicts the most common spelling variations found in the texts.

Table 1 Spelling Variations in Course Packs

Varieties of English	*Words*	*Difference*
Canadian English (CAN)	honour, labour, colour, colourful, behaviour, neighbourhood, flavours, odours	*-our*
	kilometre	*-re*
	analyze, generalize	*-ze*
	licence (when used as noun)	*-ce*
American English (US)	honor, labor, color, colorful, behavior, neighborhood, flavors, odors	*-or*
	kilometer	*-er*
	analyze, generalize	*-ze*
	license	*-se*
British English (UK)	honour, labour, colour, colourful, behaviour, neighbourhood, flavours, odours	*-our*
	kilometre	*-re*
	analyse, generalise	*-se*
	licence (when used as noun)	*-ce*

Several copies of each document were created, one for every spelling and style, as well as the original copies for comparison. The second type of error correction that was performed involved a detailed look at all the OCR mistakes that had been made. The process was not perfect, and since the course packs are made up of various articles, instructional handouts, examples from different sources, and include multiple images and charts, at times the computer did not properly recognize the text, particularly if the font style and size are different. In that case, the text can appear distorted, for example: numbers can replace letters, and requires error correction to be performed. Illustration 1 is an example of this type of OCR error:

Illustration 1 Optical Character Recognition Errors

./ Very *(ow immferation* ₒ.₂₉₆ *forefen* Cabour
!Russia
1. *1vfore foss of yoyu(ation: (ow*fortifity, *fafff"B* fije *exyectancy (2005* 147 *m{fffon- 2050 107*
   *mi*ℓ *n)*
   1. *Poor infant/motlier lieaftli care*
   2. *'Afcolio(, ln'8, S'T'lJ;* '11.'11JS *risi"t3*
   3. *'Harato increase bittli rate in sytte of ac,vernment yo(tcy*
2. *96 ofifer liftJli,* % younε *small: 2005 20 1nt«wn 65+, 24.5 mt«wn 15-24; 2050 25 nd«wn 65+,* **10.8**
   mt«wn 15-24
3. *(oss* ofafo6a*(injfuence*

GUIDING QUESTIONS

Use these questions to make sure you're selecting the important information in your notes. You should be able to use your notes to answer these questions.

1. The first step is to clearly **understand what the question means.** Look up unfamiliar words and ask questions in class. Don't ask your friend-he / shemay get it wrong   too.

Source: EAP 1500 – winter 2013, page 11*.*

A careful and thorough review of the Word documents was done to ensure that the OCR errors were dealt with. Once all Word documents were cleaned, I moved on to the corpus analysis software stage. Error corrections were not performed on content such as references or credits since they do not pertain to the learning or teaching of the

students taking the EAP classes for which the course packs are used, and they would therefore not be used in subsequent analyses.

3.3   Corpus Analysis Software

To conduct the analysis of my corpus, I utilized the software *AntConc* created by Dr. Laurence Anthony. AntConc is a "freeware corpus analysis toolkit for concordancing and text analysis" (Anthony, 2018). Once the Word documents had been cleaned, they were converted to Unicode (utf-8) text (.txt) files and imported into AntConc. The number of sources the texts come from, is six, since I narrowed down my data to the six course packs previously mentioned. Once imported, I used the software to isolate all 4-word lexical bundles occurring within all six course packs and extracted the frequency in which they occurred.

As per Biber's (2004) study, the frequency cut-off for the lexical bundle to be included in the analysis was set at 40 times per million words. This decision was arbitrary, however the rationale for having a high frequency cut-off is due to wanting a more conservative approach to the data. Biber also strictly looked at four-word sequences only, as three-word sequences are too common, and five-word sequences are often attached to multiple semantic meanings or too many other words. The Moynié corpus contains a total of 335,202 words across a range of six texts. My frequency cut-off for the following analysis was therefore 40 per million words, in concordance with Biber's numbers. The analysis of the list of 4-word bundles, also called 4-grams, will be detailed in the next chapter. This list will be compared with Biber's (2004) list, however, only the lexical bundles found in the textbook register will be included in the analysis. AntConc was also utilized to locate where these lexical bundles occurred within the texts, in order

to categorize their functions according to the functional taxonomy of lexical bundles defined by Biber (2004) in the literature review of this study. I created the Moynié Lexical Bundle List (MLBL), which consists of the 94 most frequent lexical bundles in the Moynié corpus, and the Biber Textbook Lexical Bundle List (BTLBL), which consists of the 38 most frequent lexical bundles from Biber's textbook register. Using the VLOOKUP function in Excel, I was able to compare and match the MLBL with the BTLBL in order to extract the four-word lexical bundles that occur in both most frequent lists.

In my study, I aim to determine which four-word lexical bundles are most frequently recurring in the data. Without having a frequency cut off, the number of four-word lexical bundles in the Moynié corpus is 173,167. Since the results of this study are geared towards contributing to pedagogy, it is important to be selective when deciding which lexical bundles are relevant to students and their learning. The course packs are used in the EAP program as an alternative to textbooks. Their content includes informational directives on writing within academic discourse, which they are tested on in the courses. In order to compare the frequencies of both the BTLBL and the MLBL, I used the normalized frequency so as to compare corpora of the same size. This allows for a fairer comparison of both sets of data. I used the VLOOKUP function in Excel to find both the matching and non-matching values between the BTLBL and the MLBL.

My data resulted in seven corpora, one master Moynié corpus, and six individual sub-corpora for the six course packs gathered as data. To calculate the normalized frequency of the MLBL, I analyzed the frequency of each four-word lexical bundle, and found the equivalent ratio, keeping in mind that my corpus consists of 335,202 tokens.

This allowed for a fair comparison with Biber's numbers. The comparison was not exactly precise, however, since Biber did not publish the frequency of 4-grams in the text-book sub-corpora, but rather in a larger corpus that included that and other sub-corpora. Therefore, while the BTLBL consists of 760,600 tokens, and the normalized frequency of 40 per million can be calculated for his corpus, the normalized frequency for the textbook corpus cannot. Biber's corpus is made up of 2,009,400 words in its entirety, with 760,600 words for the textbook text, and 1,248,800 words for the classroom teaching text. This is an approximation that should not affect the frequency numbers of the individual frequency lists for the course packs. The size of the individual course packs is so small, if the frequency cut off is set at 40 occurrences per million tokens, a single occurrence of a four-word lexical bundle in any of my course packs is much higher than 40. Therefore, the normalized frequency allows me to claim that the study is a fair replication and adaptation of Biber's study, using comparable numbers. My thesis is not strictly quantitative, but also a qualitative one that examines the rough proportions of n-grams in each type of corpus. With the limitations in the data in this study, I cannot compare frequencies between these corpora using rigorous statistical methods, there is always randomness that can occur given the sample size. The creation of a textbook, or in this case a course pack, can include a particular vocabulary, a specific style or register, this is where I must acknowledge the potential for bias within the data. Therefore, I acknowledge this comparison is not quantitatively rigorous, but provided that these limitations are acknowledged, a qualitative comparison will be useful regardless.

4   Chapter: Results & Discussion

The following chapter outlines the results obtained from the analysis described in Chapter 3: Methodology. The results address the research questions of this study using corpus linguistics, which explore EAP written teaching materials according to: (1) the frequency of 4-gram lexical bundles in the data collected, and (2) their function. The results were obtained through a series of analyses using AntConc which identified lexical bundles and determined their frequency in the Moynié corpus. The list of lexical bundles was then organized and compared to Biber's list using MS-Excel. Below is a detailed account of the findings, as well as interpretations of their significance within the corpus. The focus of this chapter outlines the similarities and the differences between my corpus and Biber's (2004).

4.1   Findings from the Moynié Corpus

The Moynié corpus, created here, comprises 335,202 tokens, ranging over six different EAP course packs. The Moynié corpus contains 173,167 four-word strings with a minimum frequency of 1. Following Biber (2004), these raw frequencies were then normalized to the size of a 1 million-token corpus, and then a frequency cut off of 40 per million was applied. This was done for several reasons: (1) to limit the biases inherent in specific topics represented in this corpus and the individual vocabulary choices of the course pack authors, and (2) to generate a list of lexical bundles that would be a manageable size for teaching purposes. A list of 228 four-word lexical bundles were identified as the most frequently recurring in the Moynié corpus. Out of these 228 four-word lexical bundles, 94 remained, which will be referred to as the Moynié Lexical Bundle List (MLBL). The 134 four-word lexical bundles that were excluded from the

MLBL contained tokens that would have no use in the teaching of lexical bundles. Any four-word lexical bundle that contained a proper noun was disregarded, for example *the Gulf of Mexico*, as well as lexical bundles that were too context-specific to the themes of the course packs, such as *the tar sands and*. The final 94 four-word lexical bundles occurred a total of 6,386 times in the MLBL, meaning they make up 1.9% of the total words. Table 2 below depicts the raw data collected and categorized into the MLBL. The definitions of the abbreviations in Table 1 are as follows: *discourse organizer* (DO), *referential expression* (RE), *stance expression* (SE), *ability* (A), *attitudinal/modality stance* (A/MS), *desire* (D), *elaboration/clarification* (E/C), *epistemic stance* (ES), *imprecision* (I), *identification/focus* (I/F), *intention/prediction* (I/P), *introduction/focus* (INT/F), *obligation/directive* (O/D), *specification of attributes* (SA), and *time/place/text reference* (T/P/TR), *personal* (Per.) and *impersonal* (Imp.).

Table 2 Moynié Lexical Bundle List; Normalized Frequency, Range and Discourse Functions

Moynié Lexical Bundle List	Normalized Freq. (per million)	Range	Discourse Function
in your own words	185	5	**SE** – ES - Per.
can I do it	149	5	**DO** – E/C
is to blame for	128	5	**SE** – ES - Imp.
what we can do	128	6	**SE** - A/MS – A - Per.
different forms of the	125	5	**DO** – E/C
here is an example	125	6	**DO** – E/C
is an example of	125	6	**DO** – E/C
on the other hand	122	6	**DO** – E/C
everyone needs to know	119	6	**SE** - A/MS – O/D - Imp.
we can do about	119	6	**SE** - A/MS – A - Per.
what everyone needs to	119	6	**SE** - A/MS – O/D - Imp.
forms of the word	116	6	**DO** – E/C
original set of sentences	110	4	**RE** – I/F
you'll need to	110	6	**SE** - A/MS – O/D - Per.
the evidence of the	107	5	**DO** – E/C
all different forms of	104	2	**DO** – E/C
provide all different forms	92	2	**SE** - A/MS – O/D - Imp.
that we commonly use	92	2	**RE** – I/F
the word that we	89	2	**RE** – I/F
focus of the research	89	4	**DO** – INT/F
the end of the	89	6	**RE** – T/P/TR
a state or condition	89	5	**SE** – ES - Imp.

if you don't	87	6	**SE** - A/MS – D - Per.
select three unfamiliar words	84	2	**SE** - A/MS – O/D - Imp.
are more likely to	78	6	**SE** – ES - Imp.
combine these using the	78	5	**SE** - A/MS – O/D - Imp.
in one sentence make	72	2	**SE** - A/MS – O/D - Imp.
it is sensible and	72	2	**SE** – ES - Imp.
make sure it is	72	2	**SE** - A/MS – O/D - Imp.
using the same pattern	69	5	**SE** - A/MS – O/D - Imp.
at the end of	66	5	**RE** – T/P/TR
is likely to be	63	6	**SE** - A/MS – I/P - Imp.
on the one hand	63	6	**RE** – I/F
to read the whole	60	5	**SE** – A/MS – I/P - Imp.
you may have used	60	6	**SE** – ES - Per.
don't try to	60	6	**SE** - A/MS – O/D - Imp.
will be able to	60	4	**SE** - A/MS – A - Imp.
an example of a	60	5	**DO** – E/C
at the same time	60	4	**RE** – T/P/TR
to the research question	60	4	**RE** – I/F
use three of your	60	1	**SE** - A/MS – O/D - Per.
words in one sentence	60	1	**SE** – ES - Imp.
although I study hard	60	6	**SE** – A/MS – I/P - Per.
here's an example	57	6	**DO** – E/C
what you want to	57	6	**SE** – A/MS – D - Per.
an example of the	54	4	**DO** – E/C
if you need to	54	4	**SE** – A/MS – I/P - Per.
in the same way	54	6	**RE** – SA

is related to the	54	6	**RE** – SA
the information in your	54	5	**RE** – I/F
you don't need	54	6	**SE** - A/MS – O/D - Per.
you may have to	54	6	**SE** – A/MS – I/P - Per.
at the beginning of	54	6	**RE** – T/P/TR
based on the evidence	54	5	**DO** – E/C
because you don't	54	6	**DO** – E/C
but we must not	54	6	**DO** – E/C
explain the possible meanings	54	2	**SE** - A/MS – O/D - Imp.
on the evidence of	51	5	**DO** – E/C
the examples of the	51	5	**DO** – E/C
they can fit in	51	6	**SE** - A/MS – A - Imp.
we must not change	51	6	**SE** - A/MS – O/D - Per.
you may need to	51	5	**SE** – A/MS – I/P - Per.
you want to say	51	6	**SE** – A/MS – D - Per.
you will need to	51	5	**SE** - A/MS – O/D - Per.
can be used to	51	4	**SE** - A/MS – A - Imp.
can fit in a	51	6	**SE** - A/MS – A - Imp.
in the hands of	48	6	**SE** – ES - Imp.
is one of the	48	4	**RE** – I/F
is treated as a	48	6	**DO** – E/C
the content of the	45	6	**DO** – E/C
the future of the	45	3	**SE** – A/MS - I/P - Imp.
we must protect the	45	6	**SE** - A/MS – O/D - Per.
which is treated as	45	6	**DO** – E/C
words that tell you	45	5	**SE** – ES - Per.
may be able to	45	6	**SE** – A/MS – I/P - Imp.
of the research is	45	4	**DO** – E/C

you don't know	45	6	**SE** – ES - Per.
a side effect the	45	6	**DO** – E/C
are many forms of	45	5	**SE** – ES - Imp.
as a side effect	45	6	**DO** – E/C
for the same reason	45	5	**DO** – E/C
from the source that	42	3	**RE** – I/F
help you understand the	42	6	**SE** – A/MS – I/P - Per.
likely to be a	42	6	**SE** – A/MS – I/P - Imp.
make a list of	42	6	**SE** - A/MS – O/D - Imp.
make sure you've	42	6	**SE** - A/MS – O/D - Per.
one of the most	42	4	**RE** – I/F
the different forms of	42	5	**DO** – INT/F
the main focus of	42	4	**DO** – INT/F
the primary cause of	42	5	**DO** – INT/F
the same reason they	42	5	**DO** – E/C
there are many forms	42	5	**SE** – ES - Imp.
we might have to	42	4	**SE** – A/MS – I/P - Per.
you don't like	42	6	**SE** – ES - Per.

Most of the four-word lexical bundles in the MLBL occur across the majority of the course packs, the exact break down of the occurrences can be found in Table 2. This suggests the most frequent lexical bundles are present across all levels and all years of the university program, which could mean most students have approximately the same exposure to them. The most frequent lexical bundles in the MLBL were *in your own words*, occurring 185 times, *can I do it*, occurring 149 times, *is to blame for* and *what we can do*, both occurring 128 times. The first twenty-five lexical bundles in the MLBL occurred more than 80 times in the corpus, which is twice as many times as the frequency cut off for this corpus. The full MLBL of the most frequent lexical bundles for the Moynié corpus can be found in Table 2.

4.2   Discourse Functions in Moynié Lexical Bundle List

All three types of discourse functions outlined by Biber (2004) – stance expressions, discourse organizers, and referential expressions – whose definitions and use can be found in the literature review in Chapter 2, were present in the four-word lexical bundles that emerged from the MLBL. Of the 94 most frequent four-word lexical bundles, 51 were identified as stance expressions (SE), 28 as discourse organizers (DO), and 15 as referential expressions (RE). The classification of discourse functions of the 94 most frequent four-word lexical bundles from my corpus was fairly straightforward to apply. Using Biber's (2004) functional taxonomy classification of the BTLBL, I categorized the discourse functions of the lexical bundles in the MLBL. As a first step, I identified which lexical bundles in the MLBL were the same as in the BTLBL. These matching lexical bundles received the exact same discourse function since my study is a replication of Biber's (2004) study. As a second step, I identified which lexical bundles in

the MLBL were similar to the ones in the BTLBL, but different only by one word. I proceeded to examine the differences, which were often the distinction between the use of an impersonal versus personal pronoun. As a last step, with the remaining lexical bundles to categorize, I examined their structure and the function they served, and classified them by referring once again to Biber's (2004) functional taxonomy. More than half of the most frequent lexical bundles in the MLBL are stance expressions, indicating their importance and overwhelming presence in the data the students of the EAP program at a Canadian university come across. Of the 51 stance expressions, which are the only ones that can be differentiated by an *impersonal* or *personal* attribute, 28 were identified as *impersonal*, while the remaining 23 were identified as *personal*, demonstrating that there is no large difference in the use of *impersonal* and *personal* discourse in my sample of written academic teaching materials. Examples of the *impersonal* stance expressions are: *everyone needs to know*, *select three unfamiliar words*, while examples of the personal stance expressions are: *you may need to*, *you may have to*.

As indicated by Biber (2004), the discourse functions can be further broken down according to various categories of use and purpose. The 51 stance expressions in this compressed list include: six four-word lexical bundles described as *attitudinal/modality stances*, with the purpose of indicating *ability*, of which two are classified as *personal* and four as *impersonal*. Out of the 51 stance expressions, three four-word lexical bundles are described as *attitudinal/modality stances*, with the purpose of indicating *desire*, and are all classified as *personal*. Eleven four-word lexical bundles categorized as stance expressions, of which six are classified as *personal* and five as *impersonal,* are described as *attitudinal/modality stances*, and have the purpose of indicating *intention* or

prediction. The remaining eighteen four-word lexical bundles categorized as stance expressions are described as *attitudinal/modality stances*, with the purpose of indicating *obligation* or *directive*. Seven of these bundles are classified as *personal* and eleven as *impersonal*. The list also includes thirteen four-word lexical bundles described as *epistemic stances*, where five are classified as *personal* and eight as *impersonal*. The 28 discourse organizers in this compressed list include twenty-four four-word lexical bundles with the purpose of indicating *elaboration* or *clarification*. The remaining four discourse organizers in the MLBL have the purpose of indicating *introduction* or *focus*. The 15 referential expressions in this list include nine four-word lexical bundles with the purpose of indicating *identification* or *focus*. The next two four-word lexical bundles have the purpose of indicating a *specification of attributes*. The following four four-word lexical bundles have the purpose of indicating a *time, place* or *text* reference. The last categorization of referential expressions has the purpose of indicating *imprecision*, however, no such lexical bundle was found in the MLBL. The complete list of the 94 most frequent lexical bundles in the Moynié corpus can be found in Table 2.

4.3   Discourse Functions in the Biber Textbook Lexical Bundle List

All three types of discourse functions (stance expressions, discourse organizers, and referential expressions) were also present in the four-word lexical bundles from the Biber Textbook Lexical Bundle List (BTLBL). The complete list of the 38 lexical bundles in the BTLBL from the Biber corpus can be found in Table 3.

Table 3 Biber Textbook Lexical Bundle List; Approximate Frequency and

Discourse Functions

Biber Textbook Lexical Bundle List	Approx. Freq.	Discourse Function
are more likely to	**	**Stance Expression** – Epistemic Stance - Impersonal
the fact that the	**	
it is important to	**	**Stance Expression** – Attitudinal/Modality Stance – Obligation or Directive – Impersonal
to be able to	.	**Stance Expression** – Attitudinal/Modality Stance – Ability – Personal
can be used to	**	**Stance Expression** – Attitudinal/Modality Stance – Ability – Impersonal
it is possible to	*	
in this chapter we	**	**Discourse Organizer** – Introduction/Focus
has to do with	.	**Discourse Organizer** – Elaboration/Clarification
to do with the	.	
nothing to do with	.	
on the other hand	**	
as well as the	**	
is one of the	**	**Referential Expression** – Identification/Focus
was one of the	.	

the rest of the	**	
than or equal to the size of the	.	
	**	
in the form of	**	
the nature of the	**	
in the case of	**	**Referential Expression** – Specification of Attributes
in terms of the	**	
as a result of	**	
on the basis of	**	
in the absence of	*	
the way in which	*	**Referential Expression** – Specification of Attributes
the extent to which	*	
the United States and	**	
in the United States	***	
of the United States	**	
at the same time	**	
at the time of	*	
shown in figure N	**	**Referential Expression** – Time/Place/Text Reference
as shown in figure	**	
the end of the	**	
the beginning of the	**	
the top of the	**	
at the end of	***	
in the middle of	*	

Note: The legend for the approximate frequency is as follows.

. = 10-19 per million words
* = 20-39
** = 40-99
*** = over 100

Of the 38 most frequent four-word lexical bundles, 6 were identified as stance expressions, 6 as discourse organizers, and 26 as referential expressions. Therefore, more than 68% of the most frequent lexical bundles in the BTLBL are referential expressions. This is largely due to the fact that the Biber textbook corpus included texts from disciplines such as engineering and the sciences, which often refer to various figures, tables, charts, and concepts. Of the 6 stance expressions, 5 were identified as *impersonal*, while only 1 was identified as *personal*, demonstrating there is a clear preference for the use of impersonal discourse rather than *personal* in the academic textbooks involved in these fields.

The classification of lexical bundles in the BTLBL according to various categories of use and purpose is as follows. The 6 stance expressions that occur include: three four-word lexical bundles described as *attitudinal/modality stances*, with the purpose of indicating *ability*, of which one is classified as *personal* (e.g. *to be able to*) and two as *impersonal* (e.g. *it is possible to*). The BTLBL does not include any four-word lexical bundles described as *attitudinal/modality stances*, with the purpose of indicating *desire*. There are also no four-word lexical bundles described as *attitudinal/modality stances*, with the purpose of indicating *intention* or *prediction* in this corpus. Only one four-word lexical bundle described as *attitudinal/modality stances*, with the purpose of indicating *obligation* or *directive* is found in the corpus and classified as *impersonal* (e.g. *it is important to*). The list also includes two four-word lexical bundles described as

epistemic stances, both of which are classified as *impersonal* (e.g. *are more likely to* and *the fact that the*). The 6 discourse organizers in this corpus include five four-word lexical bundles with the purpose of indicating *elaboration* or *clarification* (e.g. *has to do with*). The sixth discourse organizer in the BTLBL has the purpose of indicating *introduction* or *focus* (e.g. *in this chapter we*). The 26 referential expressions in this corpus include two four-word lexical bundles with the purpose of indicating *identification* or *focus* (e.g. *was one of the*). The next twelve four-word lexical bundles have the purpose of indicating a *specification of attributes* (e.g. *the size of the*). The following twelve four-word lexical bundles have the purpose of indicating a *time, place* or *text* reference (e.g. *as shown in figure*). The last categorization of referential expressions has the purpose of indicating *imprecision*, however, no such lexical bundle was found in the BTLBL.

4.4 Comparison of Moynié Lexical Bundle List and Biber Textbook Lexical Bundle List

A comparison of content between the Moynié Lexical Bundle List (MLBL) and Biber Textbook Lexical Bundle List (BTLBL) is necessary to assess the similarities, the differences and their significance. The BTLBL contains 38 four-word lexical bundles, which all have a varying frequency of 10 to over 100 occurrences per million words. This particular textbook register includes 87 textbooks, which make up a total of 760,600 tokens. In comparison, the MLBL consists of 6 course packs, which make up a total of 335,202 tokens. Biber's minimum cut off of 40 per million includes his entire corpus, including other registers, such as academic prose, which were included in the Moynié Corpus. Biber indicates the frequency of his bundles using a star system which categorizes the bundle frequencies into ranges, rather than reporting the exact

frequencies. This means that a perfect comparison to Biber's study is impossible using only the published data. Therefore, this study aims to perform a limited replication and comparison.

Table 4 outlines the differences and similarities of both lists by comparing the presence of four-word lexical bundles according to the three functions discussed above.

Table 4 Lexical Bundles Across Functions

Type	Moynié Lexical Bundle List (# of lexical bundles)	Biber Textbook Lexical Bundle List (# of lexical bundles)
Stance Expressions	51	6
Discourse Organizers	28	6
Referential Expressions	15	26
Totals	94	38

As Table 4 indicates, there is a larger occurrence of stance expressions in the MLBL. On the other hand, the majority of the BTLBL contains lexical bundles that are classified as referential expressions. While there is an even occurrence of both discourse organizers and stance expressions in the BTLBL, the MLBL has almost twice as many lexical bundles classified as discourse organizers than referential expressions. Perhaps this is due to the fields and themes these lexical bundles occur in, or perhaps the instructional directions the creator of these texts prefer to use. For example, *the end of the* is a referential expression bundle that has 89 occurrences across all six course packs.

Table 5 compares the coverage, using percentages, of the four-word

lexical bundles present in both the MLBL and the BTLBL according to the three

functions discussed above.

Table 5 Percentages of Lexical Bundles Coverage

Type	Moynié Lexical Bundle List (% of lexical bundles)	Biber Textbook Lexical Bundle List (% of lexical bundles)
Stance Expressions	54%	16%
Discourse Organizers	30%	16%
Referential Expressions	16%	68%
Totals	100%	100%

As Table 5 demonstrates, 54% of the lexical bundles in the MLBL are stance

expressions. The discourse organizers in the MLBL make up 30%, whereas the

referential expressions only account for 16%. The results are quite different in the

BTLBL. The referential expressions make up 68% of the lexical bundles found in the

BTLBL, while the stance expressions and the discourse organizers each make up 16% of

them. The differences in proportions of the type of bundles that occur in the MLBL

versus the BTLBL stems from the nature of the texts. The textbooks Biber uses for his

corpus include disciplines such as science, which will naturally have more referential

expressions, since they include more figures and tables. Meanwhile, the course packs

used in the EAP program include texts carefully selected to teach English academic

discourse to students, with a focus on vocabulary variety. As previously discussed in

Chapter 2, the course packs in the EAP program at hand include specific instructional

language on how to write in academic discourse. These course packs are specifically

designed for ESL students, which is why some of their content could be absent from regular university textbooks, and would explain the differences in occurrences of certain lexical bundles.

Table 6 identifies the matching four-word lexical bundles in the MLBL and the BTLBL.

Table 6 Matching Lexical Bundles List Between MLBL and BTLBL

Matching Lexical Bundles	Normalized Freq. (Moynié Corpus)	Range	Discourse Function
on the other hand	122	6	**DO** – E/C
the end of the	89	6	**RE** – T/P/TR
are more likely to	78	6	**SE** – ES - Imp.
at the same time	60	4	**RE** – T/P/TR
can be used to	51	4	**SE** - A/MS – A - Imp.
is one of the	48	4	**RE** – I/F

Of the six matching lexical bundles, three are found in all six course packs of the Moynié Corpus, and the remaining three are found in four of the six course packs, all with a normalized frequency of greater than or equal to 48 occurrences. The fact that these bundles are common across academic texts and EAP course materials is interesting and signifies their importance in English academic discourse for students. The fact that these matching lexical bundles are recurrent throughout all the levels and years of the course pack suggests they may be of use to the students. It is my recommendation that these matching four-word lexical bundles should be taught explicitly in EAP classes, since they are common and useful for students in English academic discourse.

Specific explicit strategies for teachers to use in their classrooms could include, but are not limited to, student presentations and text analysis. The list of matching four-

word lexical bundles could be given to the students as the target language to use in oral presentations. These in-class presentations would allow the students to practice their oral production skills. The presentations could be filmed or recorded, and the teachers could have other students review the recordings and outline the presence and function of the lexical bundles. A series of texts from various fields of study could be assigned to students, in which they would be required to identify the occurrence of lexical bundles, as well as their function. Wood (2015) outlines several activities that EAP instructors could utilize when teaching lexical bundles. Instructing students to listen to the radio or watch the news while keeping track of the formulaic language used, in this case lexical bundles, would help expose them to a more authentic occurrence of the target language (Wood, 2015, p. 148). Wood (2015) also refers to *student dictations* (p. 151), where students partner up, dictate a portion of a text to one another, while using their working memory to identify and remember formulaic sequences. This activity would be useful and engaging for the students in EAP programs, particularly when going over the informational texts on the various themes of the content.

This chapter presents the findings from the Moynié corpus, the discourse functions in the Moynié Lexical Bundle List, as well as in the Biber Textbook Lexical Bundle List, and compares both the MLBL and the BTLBL to extract differences and similarities in both sets of data. Out of the Moynié corpus, a list of the most frequent 94 four-word lexical bundles was extracted and classified according to Biber's (2004) functional taxonomy. Biber's (2004) list of the most frequent 38 four-word lexical bundles in the textbook register portion of his data is then analyzed and compared with the MLBL. Based on the results of my study, the most frequent four-word lexical bundles

account for 1.9% of the total number of words in the Moynié corpus. The purpose of those same lexical bundles is broken down into the three major classifications: 54% to express stance expressions, 30% to express discourse organizers, and 16% to express referential expressions. This breakdown details what students in the EAP program are exposed to. The comparison of these numbers to the Biber corpus helps determine if the two sets of data are on par with one another regarding the inclusion of lexical bundles, which they are.

5   Chapter: Conclusion

In this study, I aimed to investigate the presence and the function of four-word lexical bundles in the English for Academic Purposes (EAP) program's written teaching materials at a Canadian University. I examined six course packs from a variety of levels and years to assess the content. Based on Biber's (2004) functional taxonomy of lexical bundles, a framework of the various discourse functions of language, I classified the most frequent four-word lexical bundles in the Moynié corpus. The use of classifying lexical bundles according to functional taxonomy aims to facilitate how they might be taught by teachers in the EAP classroom and how they might be learned by students.

5.1   Pedagogical Implications of this Study

The act of explicitly teaching lexical bundles in EAP programs, as well as the importance they hold in becoming fluent, should be discussed amongst applied linguists and teachers. Hyland (2012) suggests lexical bundles are "an important component of fluent linguistic production" (p. 150). Lexical bundles are frequently used, particularly in academic discourse, and make up an important component of both comprehension and expression of the English language. The presence of lexical bundles in the context of

EAP programs, such as the that of the program examined in this thesis, indicates the need to teach them explicitly. A list of 94 four-word lexical bundles were indeed identified as the most frequently occurring in the Moynié corpus. I classified the Moynié Lexical Bundle List (MLBL) using the concept of functional taxonomy to outline their presence in academic written materials. The functional taxonomy framework poses certain challenges and difficulties. Biber's (2004) functional taxonomy has the intention to encompass "the major discourse functions served by lexical bundles" and to "describe the extent to which each register uses lexical bundles for each of these functions" (p. 383). The fact that this taxonomy captures the majority of the discourse functions implies that some discourse functions could not be accounted for. Another issue with the framework stems from the bias of the data used, and how it influenced the organization of the discourse functions and their purpose. The sample of the 87 textbooks used by Biber in his study (2004) has a topic sample bias, being that the textbooks are used in American universities, that despite the variety of fields they come from, they are still selective, and they come from a limited number (four) of academic sites. Biber (2004) does admit "this sampling does not achieve complete demographic representativeness" (p. 375). The quality of the materials I had access to gives better results for direct pedagogical application since the data was taken directly from the EAP program. However, Biber's textbook texts are important to compare my data to, in order to assess the similarity in English academic discourse, which the students of the EAP program will also be exposed to. This functional taxonomy framework has been used in past studies such as Hymes (1974) and Halliday (1978). As such, I decided to utilize this framework to make my results comparable. By being aware of the occurrence of lexical bundles in English

academic discourse, as well as their purpose and use in the English language, teachers will be more likely to understand and find creative ways to introduce lexical bundles to their students and into their teaching materials. In order to reach teachers of EAP, I would like to share the results of my study with the teachers of the EAP program, particularly the list of matching four-word lexical bundles between the MLBL and the BTLBL.

In this study, I have identified the most frequent four-word lexical bundles in the Moynié corpus I created, which could indicate which lexical bundles would be the most useful for students to learn. The MLBL outlines the most frequent four-word lexical bundles in the Moynié corpus that could be explicitly taught to students in EAP programs. The discourse functions of these 94 four-word lexical bundles are there to guide teachers on the context and purpose in which to use them. The four-word lexical bundles that were removed from the MLBL, such as *the tar sands and*, can be ignored since they have no practical use for students. Explicit instruction should focus on the academic language needs of the students especially in the context of the EAP programs. In Chapter 4, where I present the results of my study, a four-word lexical bundle such as *on the other hand* is very frequent, occurring 122 times per million words. This particular lexical bundle is classified as a *discourse organizer*, with the purpose of *elaboration/clarification*, and is an example of what could be most useful for students to learn in terms of lexical bundle. The question of whether or not students should be taught

lexical bundles explicitly can also depend on the teaching philosophy of the teacher or the institution.

5.2   Limitations of the Current Study

The analyses presented in this thesis have several limitations, including the limited access to corpora, the process of compiling the corpus, the challenge of coding the results, and the size of the corpus. Being that the field of research centered around lexical bundles is relatively new, access to corpora can be difficult and limited in quantity. Since there is a limited access to corpora, I created the Moynié corpus and used it for the current study. When compiling the Moynié corpus, Optical Character Recognition errors in the digital versions of the texts incited more thorough cleaning of the data before it could be used. The results are described as approximate because of the size of the corpus. The Moynié corpus is significantly smaller than the Biber corpus, creating obstacles with the comparative analysis of the data. The strict focus on the written versus spoken register narrows the scope and perhaps does not allow for a wide enough sample. In the present study, I did not conduct a cross-comparison of lexical bundles within different registers, which could help identify more patterns of occurrence for the key lexical bundles to teach. Although the Moynié corpus might not be generalizable to all EAP classrooms, it will provide a very direct benefit to the EAP classrooms in which the source texts are used as teaching materials.

5.3   Directions for Future Research

Further research on the presence of lexical bundles within EAP teaching materials, as well as their discourse functions, should be continued on a larger scale. A collaboration that would compile all EAP course packs, and EAP corpus data as a whole,

from all the EAP programs in Canada would improve the validity of the kind of study conducted in this thesis, as well as the generalizability of the results. At the moment, the Moynié corpus is restricted in size. This study is not a rigorous quantitative study, but rather is of a qualitative nature. The use of the functional taxonomy analysis as outlined by Biber (2004) can be used by EAP teachers to categorize and decide which lexical bundles to teach to their students. Though Biber's (2004) functional taxonomy provided an excellent base for the classification of lexical bundles, a revision of the classifications, further examples, and a broader scope of data should be explored. Further research should include a focus on the spoken register of these EAP classes. A comparison of the presence of lexical bundles in the course packs versus the spoken register would allow for a more definitive conclusion of use and purpose of lexical bundles. Lexical bundles show a clear presence in academic language. This implies their importance and their need to be an explicit part of instruction, particularly in EAP programs that intend to prepare students to write academically. In this study, I do not explore if lexical bundles of any length are explicitly taught in the EAP program's classes by the teachers at the mid-sized Canadian university, and this would indeed be an interesting research focus. This study also specifically looks at the course packs alone, and none of the other academic materials the students encounter, such as handouts, worksheets, books, etc. The focus of this study is the written teaching materials for EAP within the context of the EAP program. Further research on lexical bundles include, but are not limited to: the analysis of EAP corpora across Canada, spoken as well as written teaching materials, and longer or shorter n-grams.

Due to the complexity of the course packs, and the various genres they contain (instructional versus informational content), conducting further research and including a more detailed analysis would allow for better guidance on which lexical bundles to use and when. Another layer of analysis to conduct in future research would be on the specific context in which the lexical bundles appear by separating the various genres within each of the course packs. With this future research, perhaps different lists of lexical bundles could be extracted, and would refer directly to the different genres.

References

Altenberg, B. & Eeg-Olofsson. (1990). Phraseology in spoken English. In J. Aarts and W. Meijs (eds): *Theory and Practice in Corpus Linguistics*. Amsterdam: Rodopi, pp. 1-26.

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

Anthony, L. (2018). AntConc (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

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

Biber, D. & Conrad, S. (1999) Lexical bundles in conversation and academic prose. In H. Hasselgard & S. Oksefjell (Eds.). *Out of corpora: Studies in honour of Stig Johansson*. Amsterdam: Rodopi, pp. 181-9.

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

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., … Urzua, A. (2004). Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic (ETS Research Memorandum No. RM-04-03, TOEFL-MS-25). Princeton, NJ: Educational Testing Service. Retrieved from: https://www.ets.org/research/policy_research_reports/publications/report/2004/ibyq

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

Butler, C. S. (1997). Repeated word combinations in spoken and written text: Some implications for Functional Grammar. In C. S. Butler, J. H. Connolly, R. A. Gatward, & R. M. Vismans (Eds.). *A Fund of Ideas: Recent developments in Functional Grammar*. Amsterdam: IFOTT, University of Amsterdam, pp. 60-77.

Conrad, S. (2004). Corpus linguistics, language variation, and language teaching. *How to use corpora in language teaching*, 67-85.

Conrad, S. M. & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica.*

Coutinho, D. (2018, March 26). Canada registers a new record of international students' enrolment. *Immigration News*. Retrieved from http://www.immigrationnews.ca/2018/03/26/record-students-enrolment/

Coxhead, A. & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing, 16*, 129-147.

DeCarrico, J. & Nattinger, J. R. (1988). Lexical phrases for the comprehension of academic lectures. *English for Specific Purposes, 7*(2), 91-102.

Halliday, M. A. (1978). K (1978). Language as social semiotic: The social interpretation of language and meaning. *London: Arnold*.

Halliday, M. A. K. (1994). Functions of language. (2nd Ed.). London, UK: Edward Arnold.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"?. *TESOL quarterly*, *41*(2), 235-253.

Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. International Journal of Applied Linguistics, 18, 41-62.

Hyland, K. (2012). Bundles in Academic Discourse. Annual Review of Applied Linguistics, 32, 150-169.

Hymes, D. (1974). Foundations in Sociolinguistics: An Ethnographic Perspective. *Philadelphia: U of Pennsylvania P*.

Krashen, S. (1982). Principles and practice in second language acquisition. Oxford: Pergamon.

Jalali, H. (2017). Reflection of stance through *it* bundles in applied linguistics. Ampersand, 4, 30-39.

Meunier, F. & Granger, S. (2008). Phraseology in foreign language learning and teaching: Where to from here? *Phraseology in Foreign Language Learning and Teaching,* 247-252.

Salem, A. (1987). *Pratique des Segments Répétés*. Paris : Institut National de la Langue Française.

Schmitt, N. (2012). *Vocabulary in language teaching*. New York: Cambridge University Press.

Simpson, R. (2004). Stylistic features of academic speech: The role of formulaic expressions. In T. Upton & U. Connor (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 37-64). Amsterdam, the Netherlands: John Benjamins.

Sykes, D. (2017). *An Investigation of Spoken Lexical Bundles in Interactive Academic Contexts* (Doctoral dissertation, Carleton University Ottawa).

Tallat, G. (2011). *Understanding the social and perceptual salience* (Doctoral dissertation, University of Birmingham).

Wood, D. C. & Appel, R. (2014). Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes, 15*, 1-13.

Wood, D. (2015). Chapter 9: Formulaic language and language teaching - research and practice. In *Fundamentals of Formulaic Language: An Introduction*. Bloomsbury Publishing.

Appendix

Appendix A

Table 3: Functional classification of common lexical bundles across registers

	Classroom teaching	Textbooks	Conversation	Academic prose
I. STANCE EXPRESSIONS				
A. Epistemic stance				
Personal:				
I don't know if	**		**	
I don't know what	**		***	
I don't know how	**		**	
I don't know I	**			
and I don't know	**		.	
I think it was	**		**	
and I think that	**			
you know what I	**		**	
I don't think so	*		**	
I thought it was	*		**	
well I don't know			**	
I don't know whether			**	
I don't know why			**	
oh I don't know			**	
Impersonal:				
are more likely to		**		.
the fact that the	.	**		**

	Classroom teaching	Textbooks	Conversation	Academic prose
B. Attitudinal/modality stance				
B1) desire				
Personal:				
if you want to	***		**	
I don't want to	**		***	
do you want to	*		***	
you want to go			**	
do you want a			**	
what do you want			**	
B2) obligation/directive				
Personal:				
I want you to	***			
you don't have to	**		**	
you don't want to	**		**	
you have to be	**		.	
you have to do	**		*	
you look at the	**			
you might want to	**			
you need to know	**			
and you have to	**		.	
going to have to	**		**	
you want me to			**	
do you want me			**	
Impersonal:				
it is important to		**		*
it is necessary to				**
B3) Intention/prediction				
Personal:				
I'm not going to	**		**	
we're going to do	**			
we're going to have	**		*	
and we're going to	**			
I was going to	*		***	
what we're going to	**			
are we going to	*		**	
are you going to	*		***	
Impersonal:				
it's going to be	**		**	
is going to be	***			
are going to be	**			
going to be a	**		**	
going to be the	**			

Table 3: cont.

	Classroom teaching	Textbooks	Conversation	Academic prose
not going to be	**		*	
going to have a	*		**	
B4) ability				
Personal:				
to be able to	***	.	.	*
to come up with	**			
Impersonal:				
can be used to		**		*
it is possible to		*		**

II. DISCOURSE ORGANIZERS
A. Topic introduction/focus

	Classroom teaching	Textbooks	Conversation	Academic prose
what do you think	**		**	
if you look at	**			
take a look at	**			
if you have a	**			
if we look at	**			
going to talk about	**			
to look at the	**			
to go ahead and	**			
I want to do	**		*	
what I want to	**		.	
want to do is	**			
want to talk about	**			
you know if you	**			
a little bit about	**			
I would like to	*		**	
in this chapter we		**		
I/I'll tell you what			**	
have a look at			**	
let's have a look			**	
do you know what			**	

B. Topic elaboration/clarification

	Classroom teaching	Textbooks	Conversation	Academic prose
has to do with	**	.		
to do with the	**	.	.	.
I mean you know	**		.	
you know I mean	**		.	
nothing to do with	*	.	**	
on the other hand	**	**		***
as well as the	.	**		*
know what I mean			**	

55

	Classroom teaching	Textbooks	Conversation	Academic prose
was going to say			**	
what do you mean			**	

III. REFERENTIAL EXPRESSIONS

A. Identification/focus

	Classroom teaching	Textbooks	Conversation	Academic prose
that's one of the	**			
and this is a	**			
and this is the	**			
is one of the	**	**		*
was one of the	**	.		
one of the things	***			
and one of the	**			
one of the most	*		**	**
those of you who	**			
of the things that	***			

B. Imprecision

	Classroom teaching	Textbooks	Conversation	Academic prose
or something like that	**		**	
and stuff like that	**			
and things like that	**		*	

C. Specification of attributes

C1) Quantity specification

	Classroom teaching	Textbooks	Conversation	Academic prose
there's a lot of	**		*	
have a lot of	**			
and a lot of	**			
a lot of people	**			
a lot of the	**			
how many of you	**			
in a lot of	**			
the rest of the	**	**	*	*
a little bit of	**			
a little bit more	**			
a lot of times	**			
than or equal to	**	.		
greater than or equal	**			
per cent of the				**

C2) Tangible framing attributes

	Classroom teaching	Textbooks	Conversation	Academic prose
the size of the	.	**		*
in the form of	.	**		**

C3) Intangible framing attributes

	Classroom teaching	Textbooks	Conversation	Academic prose
the nature of the	.	**		**
in the case of	**	**		***
in terms of the	**	**		*

Table 3: cont.

	Classroom teaching	Textbooks	Conversation	Academic prose
as a result of	*	**		**
on the basis of	.	**		**
in the absence of		*		**
the way in which	*	*		**
the extent to which		*		**
in the presence of				**
D. Time/place/text reference				
D1) Place reference				
the United States and	*	**		
in the United States	**	***		*
of the United States	*	**		
D2) Time reference				
at the same time	**	**	*	**
at the time of	.	*		**
D3) Text deixis				
shown in figure N		**		*
as shown in figure		**		*
D4) Multi-functional reference				
the end of the	**	**	**	**
the beginning of the	*	**		*
the top of the	*	**	.	.
at the end of	**	***	**	**
in the middle of	**	*	*	

	Classroom teaching	Textbooks	Conversation	Academic prose
IV. SPECIAL CONVERSATIONAL FUNCTIONS				
A. Politeness				
thank you very much			**	
B. Simple inquiry				
what are you doing			**	
C. Reporting				
I said to him/her			**	

Key to symbols:
. = 10–19 per million words
* 20–39 per million words
** = 40–99 per million words
*** = over 100 per million words

Source: Biber (2004), page 3