The
University
Of
Sheffield.

Profiling lexical bundles in an EAP pre-sessional course: A corpus-based study on textbooks and instructors' materials

By:

Reem Abdulkareem I Fattani

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Arts
Department of English Language and Linguistics

October 2018

Profiling lexical bundles in an EAP pre-sessional course: A corpus-based study on textbooks and instructors' materials

Reem Abdulkareem I Fattani

A thesis submitted in partial fulfilment of the requirements for
the Degree of Doctor of Philosophy

The University of Sheffield
Faculty of Arts
Department of English Language and Linguistics

October 2018

ABSTRACT

The present study adopts a corpus analysis method and focuses on lexical bundles, a type of formulaic sequence, examining English for Academic Purposes (EAP) textbooks and addressing the unexplored area of EAP instructors' materials. This study adopts a frequency-driven approach to identify the most frequently used four-word bundles in textbooks and instructors' materials aimed at teaching academic writing in an EAP pre-sessional course at one of the UK's leading universities. To my knowledge, this study is among the first attempts to analyse lexical bundles by comparing the frequencies, functions and structures of the most frequently found four-word bundles in EAP materials to an empirically derived list called the Academic Formulas List (AFL) (Simpson-Vlach and Ellis, 2010). It employs the written AFL sub-list as an instrumental tool in the comparison to show whether EAP materials are using the most frequently occurring lexical bundles, which are common in academic writing. It also reveals findings related to in-context information, including analysing the pedagogical treatment of the identified four-word bundles.

The comparison of the four-word bundles between the three lists indicates that only four lexical bundles are shared (e.g. *on the other hand* and *it is important to*). The structural comparison reveals that the VP-based form is the most common structure in the bundles across the three lists. Furthermore, the functional comparison shows that referential expressions are common in the instructors' materials and in the written AFL sub-list, whereas stance expressions are more common in textbooks. In addition, the EAP textbooks provided learners with the highest-frequency lexical bundles, which mostly appear in instructional parts (e.g. *focus on your subject*). This means that lexical bundles have the features of being language classroom-based and instructional but not specific to academic writing. Furthermore, instructors' materials provide learners with frequently occurring lexical bundles that are usually commonly found in academic prose (e.g. *at the end of*).

This study reflects on the reality of EAP academic writing materials and highlights the nature of the bundles that learners encounter during EAP courses. Overall, the treatment and the teaching of lexical bundles in the EAP materials provided is not appropriately focused on by instructors and in textbooks. Thus, this thesis concludes with implications for instructors, EAP course syllabuses and textbook designers, and outlines areas of future research related to the inclusion of lexical bundle lists in EAP writing classes.

DEDICATION

I wish to praise *Allah* (God Almighty) for providing me with the perseverance, endurance, determination and strength needed. For that, I attribute the accomplishment of this thesis to Allah, who continues to grace me with His mercy and generous blessings, granting me the will to continue, surrounding me with positive and loving people in my life and during the completion of my PhD in the UK.

It is with great pleasure that I dedicate this PhD thesis to my late mother Dr. Ebtisam Muhammed Saleh Fatani (*God bless her soul*) and my father Abdulkareem Ibraheem Fatani. I am beyond grateful to my mum, for establishing a solid grounding and strong spiritual base for me to follow. It was during my teenage years that my mother set the stage for me as she was a PhD researcher working in academia in the Kingdom of Saudi Arabia. She is my muse, inspiration and role model, who ignited my love for reading and sought every opportunity to instill in me a love of learning. She embraced me with remarkable love, kindness, care and continuous prayers. She is unforgettable. I am also beyond grateful to my father for having faith in me and for his endless support. His continuous tenderness, modesty, wisdom and compassion helped me as a researcher. I thank him for raising me to be confident, independent, authentic and strong. I love both my parents endlessly and I hope I make them proud parents.

ACKNOWLEDGMENTS

My extended heartfelt gratitude goes to all the wonderful people who have contributed to the completion of this thesis and helped in the achievement of this academic milestone. I am sincerely thankful to my supervisor Dr. Valerie Hobbs at the University of Sheffield for supervising my work. I am most grateful for her invaluable guidance and solid advice throughout the development of this thesis. Her uplifting, comforting and passionate attitude has given me the confidence and encouragement to work enthusiastically. I am indebted to her insightful comments and constructive feedback on my written drafts. All her efforts in helping me to cope with times of difficulty are most appreciated. I would also like to thank my second supervisor Dr. Richard Steadman-Jones for his academic support and helpful comments in the early days of this thesis. Special thanks go to Dr. Nigel Harwood and Dr. Seth Mehl at the University of Sheffield for their expert advice during the early stages of this thesis.

I am thankful to the Saudi Government and King and Abdul Aziz University for granting me a scholarship and for all the assistance they provided me to undertake higher education and carry out my research in the UK. I would like to thank the pre-sessional course director and all the instructors who participated and submitted documents which were used in the present study. Without their data and appreciated involvement, this research would not have been possible.

I would like to express my deepest gratitude to my beloved husband Amro for his sincere love, distinguished support, wonderful patience and understanding. He is my soulmate. I am grateful to my son, Abdul Aziz, for his true love, for believing in me, and for his IT support. I owe a great debt to my extraordinary daughters, Waad, Nadia and Ola, for providing me with unconditional love, unfailing support, great understanding, and continuous encouragement throughout my years of study. Words are not enough to express the warmth and compassion that my daughters surrounded me with. I am privileged and honoured to have them as my daughters. Without their courage, constant reassurance, determination, and putting up with the enduring absence of my company back in Saudi Arabia, I would not have been able to complete my PhD. I love you greatly.

I am greatly indebted to my devoted siblings Salah Aldean and Samir for their genuine love, empowering support, for believing in me, and for their continuous encouragement. I am enormously thankful to my aunts Seham, Dr. Eman and

Elham, who have provided me with genuine love, surrounded me with prayers, believed in me, and offered uplifting support throughout the years of my study. I am immensely thankful to Seham for her tireless encouragement, constant phone calls from Saudi Arabia, and for pushing me to pursue my dreams. Special thanks and gratitude go to three lovely women Nidaa, Dalia Milibari, Enass, and my handsome cousin Fadhul who surrounded me with unforgettable love and warm and beautiful support. I am grateful to all of my relatives, cousins, friends and neighbours in Sheffield, the UK, and Saudi Arabia, who encouraged me in this meaningful journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

DEFINITION OF TERMS

TERMS

Academic discourse: this refers to using language that is representative of the different styles that are commonly associated with academia. Academic discourse is found when reading textbooks and journals, writing research articles and dissertations, and attending lectures and conferences (Hyland, 2009). The present study focuses on the use of lexical bundles in one area of growing interest recently linked to academic discourse, which is concerned with the development of EAP for pedagogical purposes, targeting the teaching of academic writing (e.g. EAP materials).

Academic genres: defined, for the purpose of this study, as the different kinds of texts used in academic discourse, particularly related to written discourse such as the textbook and teachers' handouts genres.

Academic register: the conventionalized language, such as vocabulary and grammar, of academia. It is characterized as being informative in purpose and geared towards a specialist audience (Biber *et al.*, 1999). The present study is interested in investigating lexical bundles of the written register in the field of EAP.

AntConc: a freeware concordance program for storing, handling and generating lexical data, which was created and developed by Laurence Anthony (Anthony, 2018b).

Collocations: the combination of words formed when two or more words are often used together in a way that sounds correct: the phrase "a hard frost" is a collocation (Cambridge Advanced Learners Dictionary, third edition: 268).

Concordance: a list of target words or sequences extracted from a specified text or set of texts, often presented in a certain way in order to show the context in which the word is used, and usually produced by most lexical handling software programs.

Corpus-based approach: a method in which the corpus is used "to expound, test, exemplify theories and descriptions that were formulated before large corpora became available to inform language study" (Tognini-Bonelli, 2001: 65). The present study is adopting this approach to investigate lexical bundles.

Corpus-driven approach: a method "where linguists use a corpus not only to provide examples but to go beyond to support linguistic argument or to confirm theoretical accounts" (Tognini-Bonelli, 2001: 84).

Corpus-informed materials: defined, for the purpose of this study, as a term used in the literature to refer to material derived from corpora for pedagogical purposes (e.g. in the form of a reference tool – a dictionary; Ädel, 2010b).

EAP: English for Academic Purposes

Formulaic: "words and word strings which appear to be processed without recourse to their lowest level of composition are termed *formulaic*" (Wray, 2002: 4).

Formulaic language/sequences and phraseology: these are two general terms that refer to a wide range of recurrent multi-word combinations/constructions, including collocations, idioms, chunks, formulas, ready-made sequences, discontinuous frames, fixed expressions, and prefabricated routines. The present study will use the general term formulaic sequences as a cover term.

Frequency-driven approach: a method where the frequency count is considered a determining factor in investigating formulaic sequences.

Idioms: these "are relatively invariable expressions with meanings that cannot be predicted from the meanings of the parts. That is, idioms are expressions which have to be learned as a whole, even if we know the meanings of the individual words composing them (e.g., piece of cake: an expression which is used when we refer to a task or activity that is easy or simple to do)" (Biber *et al*., 1999: 988).

Instructional language: defined for the purpose of this research as the language used by instructors and textbook writers to present headings, exercises, tasks, group work, and comprehension questions after reading passages.

Lexical bundles: "the combinations of words that in fact reoccur most commonly in a given register… a recurring sequence of three or more words" (Biber *et al*., 1999: 990–992).

Lexicography: the act or job of making dictionaries

Lexis: the vocabulary of a language

Native-like language: the language spoken by native speakers of the target language. The present research focuses on the English language.

N-grams/clusters: these are terms used in a corpus tool such as WordSmith and AntConc to refer to formulaic sequences such as lexical bundles.

Non-academic register: includes the conventionalized language of non-academic language such as conversation, news and fiction. For example, conversation is characterized as being spoken, interactive, and as having communicative purposes to serve the needs of interlocutors who share the same physical and temporal context (Biber *et al*., 1999: 16).

NNS: non-native speakers

NS: native speakers

Optical Character Recognition: refers to the electronic identification of handwritten or printed text characters by means of optical scanning devices and specialized computer software.

Pragmatics: the study of how language is used in communication to accomplish a specific purpose (Levy, 2008).

Prefabs: "a combination of at least two words favoured by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization" (Erman and Warren, 2000: 30).

Raise learners' attention/awareness: many research studies in Second Language Acquisition (SLA) have been conducted on this term (e.g. Schmidt, 1993, 1995; Robinson *et al.* 2014); however, the present study uses this term to refer to explicit instructions on lexical bundles through the use of activities, tasks and exercises on teaching lexical bundles in writing classes.

Readiris Pro 15: a software program that converts paper, PDFs and image documents into editable format and searchable documents, enabling the user to edit, store and retrieve the content of these texts easily and more quickly.

Teachable units or teachable bundles: a term used by Wood and Appel (2014), but, in this study, the term refers to academic lexical bundles that were highlighted to EAP learners. This phrase includes any sets of tasks and exercises which focus learners' attention on sequences, allowing learners to practise using these lexical bundles.

CHAPTER I

INTRODUCTION

Teachers and ELT publishers in the field of English for Academic Purposes (EAP) are involved in helping learners acquire the language skills needed to succeed in an English-language academic environment. In this process, a major concern relates to identifying which lexical items or multi-word combinations are most useful for students to learn for use in the academic register, such as in academic writing. At the word level, there have been significant efforts to identify frequently occurring words for pedagogical purposes, such as Coxhead (2000), who used frequency counts and range criteria to develop the Academic Word List (AWL). The AWL is a widely known list of academic vocabulary that was produced to assist teachers in choosing academic words. The list is useful for EAP students to use when constructing their written work. Learners encounter academic vocabulary or academic multi-word combinations in textbooks, journal papers and in university settings (e.g. lectures, seminars and conferences), and use this type of language in essay writing.

More recently, developments in corpus-based research have allowed large-scale lexical studies to reveal different linguistic patterns through investigating longer sequences (see Biber *et al.*, 1999; Byrd and Coxhead, 2010; Cortes, 2006; Hyland, 2008b; Martinez and Schmitt, 2012; Simpson-Vlach and Ellis 2010; Wood and Appel, 2014) (see section 2.4.6). Recent research conducted by Simpson-Vlach and Ellis (2010) on the formulaic sequences used in academic writing and research on multi-word constructions (MWC) by Liu (2012) and Wood and Appel (2014) has helped us further our knowledge on these important combinations, which have key functions in the construction of academic discourse and provide a useful resource to use in English-language learning in EAP settings.

Thus, at a multi-word level, academic multi-word combinations found in academic writing have different structures and employ a mix of diverse pragmatic functions compared to those often found in everyday conversation, or in newspapers or fiction (Biber *et al.*, 1999). Although multi-word combinations found in academic text have certain features that can sometimes be shared with other registers, they have certain qualities that are specific to academic text (Hyland, 2008a). According to Hyland (2008b: 5), these multi-word combinations help “shape text meaning and contribute to our sense of distinctiveness in a register”.

Lexical bundles are one type of multi-word combination that are often labelled under the umbrella term "formulaic sequence", and are identified using frequency counts and range measures (see sections under 2.2.2). In addition, multi-word combinations such as lexical bundles are considered an important defining feature of academic discourse (Hyland, 2012). To illustrate, lexical bundles such as *it should be noted that* help mark a text as belonging to an academic register, while bundles like *in pursuance of* are likely to identify a text as being a legal text (Hyland, 2008b: 5). Similarly, Biber *et al*. (2004) found that the bundle *on the other hand* is most common in academic writing, and functions as a discourse organizer. Moreover, in conversation, the bundle *I don't want to* is commonly used, which has the function of modality stance, reflecting the desire of the speaker.

In recent years, lexical bundles have started to receive considerable attention by linguists, teachers and researchers due to the significant role they play in language learning, particularly in academic writing. For example, studies such as Tang (2012) and Kazemi *et al*. (2014) revealed that teaching "lexical chunks" or lexical bundles raises students' awareness of the concept of sequences and helps them develop their English writing abilities. There is a general consensus among corpus linguistics researchers that most (spoken or written) texts make considerable use of multi-word sequences, formulas or bundles (Biber *et al*., 2004; Ellis *et al*., 2008; Granger and Meunier, 2008b; Hyland, 2012; Pawley and Syder, 1983; Sinclair, 1991, 2004; Wray, 2002).

Moreover, lexical theory and corpus linguistic studies have pointed to the pedagogical value of formulaic sequences such as lexical bundles in academia (Biber *et al*., 1999; Cortes, 2004; Hyland, 2008a; Hyland, 2012). In accordance with the notion that it is useful to teach the most frequently occurring bundles (Wood and Appel, 2014), corpus studies can help us by accurately providing information on which frequently occurring academic sequences should be taught to students to develop their writing proficiency.

In response, on the one hand, teachers, ELT publishers and course materials writers have been keen to incorporate these bundles into EAP materials in order to address the phenomenon of formulaicity. On the other hand, although course material designers are interested in incorporating these sequences into the syllabus, the process of selecting sequences has been somewhat subjective and has taken place without reference to data from corpora. Despite the growing attention on the importance of formulaic sequences of different types of collocations and lexical bundles (Biber *et al.*, 1999; McIntosh *et al.*, 2009), there is, however, the problem of the vast number of sequences and word combinations for learners to learn and the issue of how these sequences are presented in mainstream EAP textbooks (Byrd and Coxhead, 2010; Jones and Haywood, 2004; Wood and Appel 2014). The vast number of sequences is considered a major problem confronting course material designers. The example of the Oxford Collocations Dictionary for Students of English (McIntosh *et al.*, 2009), a corpus-based dictionary, highlights the problem facing course material designers. This dictionary lists 250,000 word combinations and 9,000 noun, verb and adjective collocations, which are the most frequently used word combinations in British and American English. As stated previously, such a vast number of combinations makes the selection process for course material design much more problematic. Therefore, although different formulaic sequences appear in EAP coursebooks, it is useful to conduct further research on this topic, such as the study conducted by Koprowski (2005), because it is worthwhile knowing whether these sequences are selected on the basis of them being the most frequently used ones.

Koprowski (2005) in his study found that nearly a quarter of the multi-word lexical items profiled in ELT textbooks may have limited value for learners. Likewise, Jones and Haywood (2004) argue that the academic writing textbooks that they analysed do not provide very useful formulaic sequences for students to acquire. This is because textbooks may use sequences that are rare in academic prose, and students are not learning the most frequently used ones. Jones and Haywood conclude that if coursebooks do not provide learners with the appropriate sequences in academic writing, then it is up to the instructors to present these bundles to their learners. Harwood (2010) states that teachers aim to cater for their learners' needs by using other materials because textbooks alone do not meet the needs of their classes. He recommends that it is not only important to examine unpublished teacher-produced materials but it is also essential to focus on published textbooks because teachers are to a certain extent required to use them in their teaching. The present study is among the first attempts to examine EAP instructors' materials as well as textbooks to uncover the lexical bundles students are exposed to during their EAP writing classes.

Several studies have revealed a lack of lexical knowledge and use among non-native speakers (Ädel and Erman, 2012; Karabacak and Qin, 2013; Wei and Lei, 2011) and native speakers (Cortes, 2004, 2006) in the academic writing of university-level students (see section 2.4.6.1). These studies have confirmed that formulaic sequences are seen as a problematic issue for language learners and that there is a gap between published writing and learner writing in academic disciplines. Cortes (2004: 421) and Chen and Baker (2010) suggest that future studies and course material designers should concentrate on providing useful ways to bridge this gap, “not only in the use of lexical bundles, but in the use of a wide variety of linguistic features”.

Other studies have compared ELT textbooks and corpora of students’ writing, in which the researchers found gaps between the language presented in textbooks and that of natural language – termed “real world” by Wood and Appel (2014: 3) (e.g. see Biber and Reppen, 2002; Harwood, 2014). The issue of how EAP materials, such as textbooks, deal with lexical bundles is attracting many researchers’ attention, as previously mentioned (e.g. Jones and Haywood, 2004; Koprowski, 2005; Wood, 2010; Wood and Appel 2014). Wood (2010) conducted a small-scale corpus research study on EAP textbooks, which found that textbook activities paid limited attention to formulaic sequences or lexical bundles.

Having briefly reviewed some of the research studies regarding the concerns associated with the teaching and treatment of lexical bundles, and since phrase/collocation learning is an essential component of EAP teaching (Lewis, 1993), it is clear that it would be useful to examine lexical bundles by reviewing the materials provided to learners in EAP academic writing classes. This will lead to furthering our understanding of the treatment of lexical bundles in EAP materials in general and in textbooks and instructors’ materials in particular. The present corpus-based study will not only provide quantitative data but also will be supported by qualitative results, adding valuable insights via an in-depth analysis of the in-context information of bundles. In addition, for pedagogical decision-making, material designers, instructors and curriculum developers, it would be useful to report on which bundles were presented to learners in terms of the notion of “teachable units” (see Chapter 4, section 4.5.6), and how they were presented. This involves analysing concordance lines to determine the type and kind of tasks and/or exercises in which these bundles occurred, extending the body of research on the development of formulaic learning and its methodological principles.

In addition, the present study extends the research into the use of lexical bundles, focusing on academic writing and targeting EAP materials. Besides my key interest in exploring academic writing, the choice to investigate lexical bundles in academic writing derives from the fact that most university disciplines and academic language courses currently include some form of academic writing as an important requirement in their curricula. With research levels on the rise across a wide range of fields and disciplines, it seems that there is a strong need among both students (e.g. university-level and EAP) and institutes and universities to develop native-like academic compositions (see Flowerdew, 2015; Hyland and Hamp-Lyons, 2002; Paltridge, 2004; Tribble, 2015). Furthermore, since many research studies such as Biber (2009) and Miao (2014) have revealed that the spoken and written registers display different sets of lexical bundles, it was logical and useful for the present study to focus on bundles commonly related to academic writing, as can be seen in Chapter 2.

To my knowledge, studies have investigated the use of lexical bundles by only focusing on EAP textbooks (Jones and Haywood, 2004; Wood, 2010; Wood and Appel, 2014). However, the present study extends the research on EAP materials, by analysing the use of lexical bundles in unpublished materials (e.g. instructors' handouts) that are presented to students rather than on focusing on published textbooks alone. The present study adopts a design that incorporates corpora (compiling EAP texts) and assumes a frequency-driven approach (see section 2.2.2.4) to generate the most frequently used four-word bundles extracted from materials in an EAP pre-sessional course.

In addition, to my knowledge, this study is among the first attempts to analyse lexical bundles by comparing them to a corpus-based derived list called the Academic Formulas List (AFL). The aim of this approach is to acknowledge the usefulness of corpus-based research, showing how the usefulness of the AFL is not limited to teaching purposes. Instead, it employs the AFL as an instrumental tool to use in the comparison to show whether EAP materials are using the most frequently occurring lexical bundles and presenting them to learners (see Chapter 4, section 4.5.4.1 for reasons for choosing the AFL). This will be achieved by identifying the most frequently occurring four-word bundles generated from EAP materials and comparing those found to the AFL. Such a comparison will result in profiling the most frequently used four-word lexical bundles found across two genres within the academic register: textbooks and instructors' handouts.

The results provided by this research will serve as a reflection of the reality of EAP materials targeted at academic writing and will highlight the nature of the bundles that learners encounter during the EAP course. EAP material must be presented in a way that is useful to students. Given recent trends towards making textbooks and supplementary materials engaging for students, I predict that they may not be employing the most frequently occurring lexical bundles from the AFL. In addition, it is important to highlight the types, structures, and functions of bundles that these EAP materials are employing. I will also present in-context information on where the bundles occur in the text and highlight pedagogical information to see if EAP materials explicitly point out lexical bundles to EAP learners.

This thesis is structured into 11 chapters. Following this first introductory chapter, Chapter 2 presents a review of the literature that informs this research. This includes a brief overview of formulaic sequences, and the role of corpora in analysing sequences and bundles, with a particular emphasis on lexical bundles. Chapter 3 provides information on the background and context of the EAP pre-sessional course selected for the present research. The chapter details the study's context and decision-making process, covering the reasons behind the selection of the pre-sessional course. It serves to provide a description of the EAP pre-sessional course and the nature of the academic writing materials used.

Chapter 4 takes into account the three main methodological phases involved in creating the corpora of lexical bundles from EAP materials for analysis in this study. After the data was processed, Phase 1 involved the process of compiling the corpora of EAP materials. Phase 2 involved the process of generating, refining and organizing the bundles extracted from the two corpora using a computer software program, with the application of well-defined identification criteria. This resulted in the production of EAP bundles ready for analysis. Phase 3 involved an exclusion specifications process that was applied to filter out unrelated EAP lexical bundles from the analysis. The chapter also presents the procedure used to undertake the types and frequency, structural, and functional classification of the EAP bundles. These classifications set the stage for the analysis and for a comparison between the bundles found in EAP materials (instructors' materials and textbooks) and those found in the modified list, which in this thesis is named the written AFL sub-list. In addition, the chapter includes the procedure used to explore the bundles in context and as teachable units to establish whether they are presented to EAP learners as teachable bundles. In addition, this chapter outlines the reasons underlying the selection of the AFL, as well as the modifications made to the selected bundles from the AFL in the course of producing the written AFL sub-list and its corresponding functional and structural categories.

Chapter 5 presents the quantitative results derived from the corpus regarding bundle use in instructors' materials, in terms of frequency and structural and functional analysis. Chapter 6 presents the quantitative results derived from the corpus regarding bundle use in textbooks in terms of frequency and structural and functional analysis. Chapter 7 seeks to discuss the comparison of the most frequent four-word bundles found in the EAP materials (instructors' materials and textbooks bundles) with the written AFL sub-list. In addition, the chapter will provide an in-context analysis of the bundles identified, and present a complete analysis of the types of bundles in relation to them being teachable units. Chapter 8 discusses the structural aspects of the EAP bundles found in the instructors' materials and textbooks, describing their form compared to the written AFL sub-list and that found in other research studies. Chapter 9 compares the functions of the EAP bundles in the present study to those in the written AFL sub-list and other studies, establishing the types of bundles identified and discussing the treatment of overall functions of lexical bundles in EAP materials.

Chapter 10 presents a general discussion, addressing and answering each of the five research questions (see Appendix A). Chapter 11 concludes with a discussion of the limitations of the study, stating the implications for teaching lexical bundles, and providing recommendations for further research.

CHAPTER II

REVIEW OF LITERATURE

2.1 The scope of formulaic sequences

The study of formulaic patterns in language has attracted linguists' attention since the beginning of the nineteenth century. Early studies on formulaic patterns referring to "lexical co-occurrences" dates back to researchers such as Jespersen (1924) and Firth (1951), who studied the concept of collocations (Cortes, 2004). In 1950, Firth (1951, 1957) used and popularized the term "collocations" to describe how words combine with each other or other words; he has also been recognized for his famous slogan: "you shall know a word by the company it keeps" (cited in Ellis, 2008: 1).

Over the years, a substantial amount of research has been conducted on the study of formulaic patterns under different terms, with different research perspectives and across different fields. *Formulaic language/formulaic sequences* (Schmitt and Carter, 2004; Wray 2002) and *phraseology* (Cowie, 1998; Granger and Meunier, 2008b) are two umbrella terms often used in the literature to refer to several types of multi-word combinations such as idioms, proverbs, collocations, routines, and set phrases. There is a wide range of terms to describe the phenomenon of formulaic language or phraseology (e.g. ready-made utterances, prefabricated routines, chunks, collocations, formulas, multi-word combinations) (Wray, 2002: 9).

The scope and diversity of the terminology made it problematic to decide on an overall definition for the notion of formulaicity. The issue, according to Wray (2002), is that different scholars and researchers may use some shared terms to express different perspectives across different fields, which do not mean entirely the same thing. For example, in addition to the terms mentioned above, Wray and Perkins (2000) and Wray (2002: 9), found over 50 other shared terms used by different researchers to describe the different types of formulaic sequences. Schmitt and Carter (2004: 3) observe that the diversity lies in the length and purpose of word sequences; for example, formulaic sequences could be short as in *Oh no!* or long as in *You can lead a horse to water*, and *but you can't make him drink.*

Regarding the purpose of these formulaic sequences, they can be used for different social requirements such as fixed expressions (*Happy Birthday* and *ladies and* *gentlemen*), idioms (*kick the bucket*), or lexical bundles (*on the other hand*) (which will be explained in detail in this chapter). They can also express a function in the language ([*I am] just looking, [thanks]* to decline an offer politely when offered assistance from a salesperson), or can be used in collocations (such as in words that collocate with the word *exam*, e.g. *[take/pass] the exam* but not *[do] the exam*).

Biber *et al*. (2004: 372), in agreement with Wray (2002), argues in relation to multi-word units that "there is little agreement on their defining characteristic, the methodologies to identify them, or even what to call them; and, as a result, there is little agreement across studies on the specific set of multi-word units worthy of description". Biber *et al*. (2004) also state that empirical studies approach "multi-word units" or formulaic sequences differently in terms of: (1) research goals (identifying full-range vs. small-range multi-word units); (2) the identification criteria used for retrieving sequences (perceptual salience and frequency criteria); (3) formal characteristics of the multi-word units selected (continuous sequences, discontinuous frames, or two-word collocations); (4) the amount of text samples selected (choosing from small, large or very large corpora); and (5) the choice of including or excluding register comparisons (many research studies ignore register totally, while others only analyse the spoken register).

Similar to Wray (2002), Chen and Baker (2010) contribute to the argument that different words or expressions may be used to refer to the same concept. For example, according to Chen and Baker (2010: 30), expressions or terms such as *clusters* (Hyland, 2008a; and also used in the corpus software WordSmith), *recurrent word combinations* (Altenberg, 1998; De Cock, 1998), *n-grams* (Stubbs, 2007), and *lexical bundles* (Biber and Barbieri, 2007; Cortes, 2006) are often found and used in a corpus-driven approach. These expressions refer to frequently retrieved fixed multi-word units or formulaic sequences that have pragmatic discourse functions that are often recognized by native speakers within certain contexts (Chen and Baker, 2010). Due to the wide range of these terms and the problematic issues discussed, I have decided to use the blanket term *formulaic sequence* that is in wide circulation among linguists (e.g. Schmitt and Carter, 2004), and which is defined by Wray (2002: 9) as:

> a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.

Schmitt and Carter (2004: 4) agree with Wray (2002) and find the term formulaic sequence to be “all-encompassing, covering a wide range of phraseology”. In addition, in this definition the general descriptive term of formulaic sequence includes the different meanings of word combinations and their different internal patterns. This means that Wray uses the term formulaic sequence “fairly loosely” as an overall term to include “any kind of linguistic unit that has been considered formulaic in any research field” (Wray, 2002: 9). As mentioned in the introduction chapter, the present study is interested in one type of formulaic sequence called *lexical bundles*, so this ‘inclusive’ formulaic sequence definition provides a suitable starting point to introduce and establish a clear understanding of the meaning and use of lexical bundles.

2.1.1 Formulaic sequences and generative language

Having discussed the matter of terminology, it is important to distinguish between two theoretical views related to language processing and production. Simply put, Sinclair’s (1991) view is that language as a whole is formed on the basis of two main structuring principles: an *open choice principle* and an *idiom principle*. This means that when language users construct their spoken or written discourse, their language use results from two different options. The *open choice principle* focuses on the speaker’s ability to generate new linguistic units by creating novel patterns and word combination structures, and also the ability to understand sentences at the time of production that they have never encountered before (Chomsky’s modern linguistic theory; Chomsky, 1964). On the other hand, the *idiom principle* emphasizes the speaker’s reliance on ready-made formulaic sequences that the speaker has heard, used or produced before. Similarly, Wray (2002: 14) offers a dual-processing system solution: *analytical processing*, which accounts for interaction between words and morphemes with grammatical rules to create and decode new or semi-new language, and *holistic processing*, which entails relying on “prefabricated strings” that are saved in memory. She views the choice of strategy as depending on the demands of the material and the communicative setting.

Pawley and Syder (1983: 196) made a comparison between the idiomatic expression *I want to marry you* and non-idiomatic, but also grammatical, sentences (*I wish to be wedded to you*, *I desire you to become married to me*). They argue that although the non-idiomatic sentences are grammatically correct, it is clear that they are not the preferred way in which to offer a marriage proposal, and do not sound native-like language. This may indicate that during language processing, certain default phrases are likely to be preferred over other new and grammatically suitable phrases (Wray, 1999). This argument seems to support the idiom principle.

Analytical processing is preferred, for example, when the language user paraphrases someone else's utterance instead of relying on their exact words. In such a case they create their own new structures, which are grammatically suitable and accepted by language users. In this light, the two sentences previously mentioned (*I wish to be wedded to you*; *I desire you to become married to me*), for example, may be perceived as acceptable replacements for *I want to marry you* to achieve a humorous effect or in formal surroundings. Irrespective of which principle is found to be more dominant, it is in line with the second principle, proposed by Sinclair (1991), that the notion of formulaicity has gained increased interest among linguists and researchers.

2.1.2 Formulaic sequences: storage and retrieval

This section will shed some light on some of the characteristics of formulaic sequences that are being discussed and investigated: storage, processing and retrieval. From a psycholinguistic, cognitive science and language study perspective, researchers try to offer descriptions and evidence in an attempt to describe how formulaic sequences are perceived, stored, remembered and processed. In addition, they are interested in addressing the question of how formulaic sequences are stored and remembered in the brain. Many studies on language processing suggest that language is stored in the form of fixed phrases or long memorized chunks, and they are often retrieved from memory as pre-assembled chunks (Bolinger, 1976). Another example can be derived from an influential article on memory written by Miller (1956), in which he claims that short-term memory may hold a reasonable number of "bits" or "chunks" of information rather than individual items. Chunking is a psychological process, which focuses on combining individual units of items to form groups. He argues that we use chunking every day without even noticing it. For example, we usually tend to remember and write down phone numbers in chunks.

This finding also supports Wray's (2002) review of the work of linguists such as Saussure (1916/1966), Becker (1975), and Bolinger (1976). These linguists were found to use expressions such as "short-cuts", "ready-made frameworks", and "prefabs", respectively, to put forward the concept that the formulaic sequence "offers processing benefits" to language users and creates a "short cut to production and comprehension" (Wray, 1999: 213). Thus, language users store chunks of words as individual phrases with holistic meanings in order to facilitate communication and to save processing time.

Regarding language learning, linguists today are seeking insights and employing approaches from different fields such as social psychology and cognitive psychology to examine language in general and formulaic sequences in particular. Entrenchment is considered a significant result of this interdisciplinary approach. The concept of entrenchment is based on the idea of a continuing process of restructuring and adaptation of communicative knowledge. There is an association between the role of frequency (see section 2.2.2.1 of this chapter), entrenchment and cognitive studies. As cited in Schmid (2017: 9), according to the well-known cognitive linguist Ron Langacker (1987: 59):

> Every use of a [linguistic] structure has a positive impact on its degree of entrenchment, whereas extended periods of disuse have a negative impact. With repeated use, a novel structure becomes progressively entrenched, to the point of becoming a unit; moreover, units are variably entrenched depending on the frequency of their occurrence.

One key assumption associated with entrenchment is that the repeated usage of a specified linguistic structure or item triggers its processing as a holistic unit. More recently, psycholinguistic research has investigated the relationship between language processing and formulaic sequences. Ellis *et al.* (2008: 377) state that "language processing is sensitive to formulaicity and collocation". According to Ellis (1996), Nattinger and DeCarrico (1992), Wray (1999, 2002), Schmitt and Carter (2004), and Jiang and Nekrasova (2007), a formulaic sequence is perceived as a common unit that is stored in the mind as a single significant unit with its own associated holistic meaning. This assertion is supported by scholars and many studies on language processing and memory. Pawley and Syder (1983) agree with Wray (1999) and affirm that one benefit of the formulaic sequence is that it saves processing time, because the single memorized strings are processed faster and more easily than individual words that are generated from scratch.

A study conducted by Jiang and Nekrasova (2007) used material derived from corpus data on "word combinations" in two online grammatical judgement experiments. In the second experiment, the researchers wanted to find out the participants' rapid reaction times to formulaic sequences and whether or not this is due to their familiarity with the sequence. The results showed "prevailing evidence in support of the holistic nature of formula representation and processing in second language speakers" (Jiang and Nekrasova, 2007: 433). They also revealed that both native speakers and non-native speakers responded to formulaic sequences in English much faster and with fewer mistakes than to non-formulaic sequences.

Millar (2011) conducted a study in which he examined how native speakers process learner collocations which digress from the patterns of the target language. His results showed that the errors produced in learners' collocations added a greater and sustained cognitive burden to native speakers' processing. More importantly, his finding supports the widely emphasized claim that "formulaic sequences provide processing advantages" (Millar, 2011: 129).

Underwood *et al*. (2004) in their research showed that the words included in formulaic sequences could be read more quickly than the same words which are not part of a sequence. In a similar study, Conklin and Schmitt (2008) investigated the notion that formulaic sequences save processing time by comparing reading times for formulaic sequences against matched non-formulaic phrases for native and non-native speakers. The results showed that both groups read the formulaic sequences more quickly than the non-formulaic phrases. This result supports the claim that formulaic sequences have a processing advantage over newly created language. The findings also indicate that it is possible for both native speakers and learners to enjoy the same type of processing advantage.

There is also some evidence concerning how formulaic sequences are retrieved. Wray (2002) concludes that the holistic system has the advantage of reducing processing effort and retrieving prefabricated strings, which is more efficient than creating novel sentences. Erman (2007) examined the pause frequency and pause duration of prefabricated language (*prefabs*) in spontaneous speech, in which pausing is seen as a measuring tool of cognitive effort in lexical retrieval. The results indicated that prefabs are stored and retrieved as whole units from long-term memory.

Having reviewed the literature on the scope of formulaic sequences, covering the terminology, addressing the formulaic and generative aspect, and considering issues associated with storage and retrieval, I will now turn to present my exploration of the literature on lexical bundles.

2.2 Lexical bundles

The term *lexical bundle* was first described and explored in detail in the Longman Grammar of Spoken and Written English (LGSWE) (Biber *et al.*, 1999: 13), in which the authors compared the most common recurrent expressions in conversation and academic prose based on their inclusive corpus-based study of English grammar. According to Biber and Barbieri (2007: 263), lexical bundles are strings of words which co-occur repeatedly and "they are important building blocks of discourse in spoken and written registers".

In addition, Cortes (2006) defines lexical bundles as combinations of three or more words which are identified in a corpus of natural language. A similar viewpoint is held by Biber *et al.* (1999) and Hyland (2008a, 2008b, 2012); the latter states that "bundles are statistically the most frequent recurring sequences of words in any collection of texts. They are extended collocations that appear more repeatedly than expected by chance" (Hyland, 2012: 150). They can be identified empirically and retrieved automatically by using a computer program – corpus analysis software – with specified frequency and distribution criteria, which will be discussed in section 2.2.2 of this chapter.

Due to their occurrences and distributional criteria, these bundles have not been found to be idiomatic nor complete grammatical units (Biber *et al.*, 2004; Hyland, 2012). For example, lexical bundles are not idiomatic in terms of meaning, such as *kick the bucket (detonating death)* and do not form a complete grammatical phrase such as *I am going home now*. However, they are, as mentioned previously, extended collocations that help form meanings in special contexts and provide a sense of unity and coherence in a text (McCulley, 1985). Their internal structure will be fully explained in section 2.2.3.1 of this chapter. Some examples of lexical bundles in the academic written register are *as a result of* and *on the other hand* and in the spoken academic register *I don't know if* and *we are going to do*. Although frequency is a distinctive criterion for the identification of lexical bundles, it is not the only feature of determination, as will be explained in section 2.2.2.4. Taking into consideration the general definitions of lexical bundles presented in this section, it is important to establish a working definition of lexical bundles for use in the present study. This will be particularly useful when employing exclusion criteria in the methodology chapter (see Chapter 4, section 4.4). Thus, in the present study, lexical bundles are defined as continuous fixed sequences extracted by a defined frequency threshold and range criteria and which are used as fixed phrases.

2.2.1 Idiomaticity and fixedness

A distinguishing characteristic of lexical bundles is their lack of idiomaticity. Most lexical bundles are not considered idiomatic but perceptually salient because their meaning is transparent and can be comprehended from the individual words that constitute the bundle (Biber, 2009; Cortes, 2004). This contrasts with idioms, where their meaning may not be derived from their internal constituents. A typical example is the idiom *kick the bucket*, where its components do not refer to the act of dying. In contrast, *as a result of* and *in the presence of*, among many others, are seen as fully compositional.

Fixedness is another distinctive characteristic of lexical bundles, and refers to their "fixed collocational patterns" based on frequency of occurrence (Hyland, 2012: 152). Most lexical bundles are characterized to a large degree by their fixed word order (e.g. *as a result of* and *in terms of the*). Additionally, Cortes (2004) points out that the fixedness feature of lexical bundles is determined by the criterion of frequency, which differentiates lexical bundles from other word combinations. This means that not all forms of word combinations qualify to be defined as lexical bundles; only the fixed sequences that meet a specified cut-off frequency. For example, Cortes (2004) in her study only selected the combinations that met the cut-off frequency of 20 occurrences per million words, such as *these results suggest that* but not its singular form *this result suggests that*, which did not come up as a bundle when the cut-off frequency was established and the computer program generated the frequent bundles. Also, Salazar (2011) selected the bundle *are expressed as*, which has a high frequency rate, but not its singular form *is expressed as*, because it does not qualify as a bundle according to the frequency approach.

In contrast, DeCarrico and Nattinger (1988) and Nattinger and DeCarrico (1992: 8) view "lexical phrases" or lexical bundles as being "subject to differing degrees of syntactic modification". Nattinger and DeCarrico state that some sequences such as *for the most part* are fixed sequences that do not allow any modification to their pattern (and which are the focus of the present study). Other combinations such as *it is only in X that Y* are also formulaic sequences that have optional slots in addition to their fixed components, which allow a certain degree of modification to their construction. The Y slot is only filled with full sentences, but the X slot may be filled by a noun or gerund phrase (e.g. *it is only in running uphill that I have trouble breathing*). This type of formulaic sequence is not considered a lexical bundle because frequency and fixedness are together two defining qualities of lexical bundles (Cortes, 2004).

2.2.2 Identification of bundles

2.2.2.1 Frequency of bundles

Wray (2002) sees frequency as a salient and determining factor in the identification of formulaic sequences. In corpus linguistics, researchers use automated corpus tools to determine the number of lexical bundles to be included in a study, which is done on the basis of frequency counts. This reveals the number of times these phrases – lexical bundles – occur within texts. Although this method is widely used by many researchers, threshold frequencies are nevertheless somewhat arbitrary and depend on the scope of each study (Hyland, 2008b) (see Chapter 4, section 4.3.2.3).

Research on lexical bundles has employed different lexical bundle cut-off frequencies that range from 10 to 40 occurrences per million words, depending on the size of the corpora. For example, for larger written corpora, the normalized frequency cut-off ranges from 20 to 40 occurrences per million words, while researchers handling relatively smaller, usually spoken, corpora use a lower cut-off frequency, which often ranges from 2 to 10 occurrences per million words. Studies such as Ädel and Erman (2012), Biber *et al*. (2004), Chen and Baker (2010), Cortes (2004, 2006), Csomay (2012), Hernandez (2013), Hyland (2008b), and Jablonkai (2010) used frequency bands ranging from 10 to 40 occurrences per million words. For example, in their study, Biber *et al*. (2004) set the frequency cut-off to 40 times per million words in their sub-corpus of around 2 million words. In studies by Altenberg (1998) and De Cock (1998), the frequency cut-off ranged from 2 to 10 occurrences per million words. For the present research, the frequency cut-off will be fully discussed in Chapter 4, section 4.3.2.3.

Researchers handling smaller spoken corpora usually use lower frequency bands, which could lead to problems in identifying lexical bundles, particularly when comparing a small-sized corpus with larger corpora. As a result, further research and calculations may be required before reliable comparisons can be initiated

(Hyland, 2012). For example, Chen and Baker (2010) in their methodology section discuss the appropriate cut-off frequency that they managed to set when comparing three corpora of various sizes. After repeated experiments the cut-off frequency was assigned to suit the three corpora in their study (see Chapter 4 – methodology). This normalization process is regarded as a critical factor when dealing with frequency counts across corpora of different sizes. The normalization process "involves extrapolating raw frequencies from the different-sized corpora which are being compared so that they can be expressed by a common factor such as a thousand or a million words" (Evison, 2010: 126). In the methodology chapter (see section 4.3.2.3), the calculation of the raw and frequency counts needed for this study are fully reported.

2.2.2.2 Distribution of bundles

According to Chen and Baker (2010) and Hyland (2012), a second standardized identification criterion that is considered another clear-cut approach for identifying and classifying lexical bundles is the range or breadth of use. This specifies the number of files or texts in the corpus in which the bundles occur. Previous corpus-driven research on lexical bundles has adopted various measures for specifying the range, including percentage occurrence (Hyland, 2008b) and a minimum number of different files or texts (Chen and Baker, 2010; Csomay, 2012; Wei and Lei, 2011). For example, Chen and Baker (2010) set the frequency and distribution threshold for identifying lexical bundles to around 25 occurrences per million words on average, appearing across at least three texts.

Similarly, Cortes (2004) and Csomay (2012) both took a conservative approach in identifying lexical bundles by setting a minimum frequency of 20 occurrences per million words and the appearance of the lexical bundle in at least five or more different texts. The frequency and dispersion thresholds employed for extracting lexical bundles vary from one study to another, depending on corpora sizes (Biber *et al.*, 1999; Chen and Baker, 2010; Salazar, 2011). This criterion is essential for avoiding the quirks of individual users (Chen and Baker, 2010; Cortes, 2006; Csomay, 2012; Hyland, 2008b, 2012). (see Chapter 4, section 4.3.2.2 for information on the distribution of bundles selected in the present research.)

2.2.2.3 Length of bundles

The third identification criterion is to determine the length of the strings, which requires specifying the three-, four-, five-, or more word sequences to be included in any investigation. For example, some studies focused on only examining four-word bundles (Biber *et al*., 2004; Chen and Baker, 2010; Cortes, 2004, 2006; Hyland, 2008b). Others such as Biber *et al*. (1999) included four-, five-, and six-word sequences and Simpson-Vlach and Ellis's (2010) Academic Formulas List (AFL) included three-, four-, and five-word sequences in the data set to provide a more comprehensive retrieved list.

In addition, lexical bundle studies have reported that longer bundles have a lower frequency range (Hyland, 2008a; Simpson-Vlach and Ellis, 2010). For example, Biber *et al*. (1999) found that three-word bundles appear ten times more frequently than four-word bundles and that four-word bundles occur ten times more frequently than five-word bundles. Drawing from previous research, most corpus-based studies focus on four-word lexical bundles because "they are over 10 times more frequent than five-word strings" (Hyland, 2012: 151). In addition, their pattern provides a wider selection of structures and functions for analysis than five-word strings (Cortes, 2004; Hyland, 2012).

Three-word bundles are particularly common while five- and six-word bundles are rare and can include shorter strings (Biber *et al*., 1999; Cortes, 2004; Hyland, 2012). To illustrate, the five-word lexical bundle *it has been suggested that* includes the four-word sequence *it has been suggested.* Additionally, Chen and Baker (2010), at the level of data analysis, manually excluded overlapping word sequences, where two four-word bundles are actually derived from a five-word string (e.g. *it has been suggested* and *has been suggested* are part of the bundle *it has been suggested that*). The authors argue that these overlapping bundles could inflate the results of the analysis if they were all included in the study. The present research only focuses on four-word bundles (see Chapter 4 section 4.3.2.1 for the reasons behind only focusing on four-word bundles).

2.2.2.4 An additional method of bundle identification

Mutual Information (MI) is a statistical measure frequently used in the field of theoretical computer science known as information theory. Recently, researchers in corpus linguistics who are interested in investigating formulaic sequences have been turning to additional methods to compute the association between words in a bundle. Simpson-Vlach and Ellis (2010: 493) derived a pedagogically useful list of formulaic sequences called the Academic Formulas List (AFL) by employing MI to calculate "the degree to which the words in a phrase occur together more frequently than would be expected by chance". A higher MI score indicates that there is a strong association between the pair of words, while a lower MI score shows that the co-occurrence is more possibly due to chance (Oakes, 1998). Other studies have also used MI as a reliable metric for extracting lists from a corpus (e.g. Tsai, 2014).

Having an understanding of the MI score, the role of frequency over MI is of key importance to the present research. While many current studies employ MI as a statistical measure in their studies to produce reliable empirical derived lists for teaching purposes (Simpson-Vlach and Ellis, 2010), many studies question its reliability when trying to account for long sequences. Although this study views the methodology used in Simpson-Vlach and Ellis (2010) as very useful, it is important to address certain issues relating to why the present research is not employing the MI measures and is only relying on the frequency-driven approach.

On the one hand, researchers in corpus linguistics who are interested in investigating formulaic sequences are turning to sophisticated statistical methods to compute the association between words in a bundle. MI measures the strength between any given pair of words by comparing the frequency of the whole pair (together) to the overall frequencies of each of the individual words in the pair (McEnery and Wilson, 2001). On the other hand, since the application of the MI statistic is designed to calculate the strength of two-word collocations, it does not take into account word order (Biber, 2009). Therefore, the results may be unreliable when determining the frequency of longer sequences (Hyland, 2012). Furthermore, the MI statistic tends to favour low-frequency words, and simply reflects the likelihood that a pair of words will occur together, regardless of word order (Biber, 2009; Hyland, 2012). Since the present study examines lexical bundles that have unique fixed structures such as *on the other hand* and is investigating bundles that occur the most frequently in corpora, MI is not seen as a required identification method in this line of research.

In addition, in the frequency-driven approach, researchers look at high-frequency counts as one important reflection of multi-word combinations, in terms of bundle identification. This is because speakers and writers commonly use these combinations as “unanalyzed chunks”, which have noticeable discourse functions in texts. When analysing language data using their intuition, linguists usually ignore lexical bundles and do not recognize them because bundles are not structurally complete (Biber *et al*., 2004: 376). For example, *I don’t know if* and *it is possible to* are common lexical bundles, but they are unlikely to be identified as complete lexical chunks based on intuition only. Thus, as I will demonstrate in the following sections, it turns out that lexical bundles identified simply on frequency counts do have strong functional correlates, with speakers and writers regularly using them as “basic building blocks of discourse” (Biber *et al*., 2004: 371; Biber and Barbieri, 2007).

2.2.3 Structural and functional classifications of bundles

2.2.3.1 Structure of bundles

Due to the established fact that lexical bundles are identified empirically, not intuitively, as multi-word combinations that occur frequently in a register, they are, in most cases, “not complete structural units, but rather fragmented phrases or clauses with new fragments embedded” (Cortes, 2004: 400). Biber *et al*. (2004: 399) state that lexical bundles are unlike the grammatical structures found in traditional linguistic theories, but rather, most lexical bundles have “well-defined structural correlates”, which make it possible to group them into several basic structural types.

Based on these typical grammatical correlates, a structural classification of lexical bundles was created by Biber *et al*. (1999) in the LGSWE corpus that has been widely used in many studies on word sequences (Ädel and Erman, 2012; Biber *et al*., 2004; Chen and Baker, 2010; Cortes, 2004, 2006; Csomay, 2012; Hernandez, 2013; Salazar, 2011, 2014; Wei and Lei, 2011). For example, the bundles *the end of the* and *the beginning of the* have a noun phrase with an of-phrase fragment structure. Table 2.1 shows the structural categories of bundles found in academic prose. The present study will use the following structural framework in the classification of bundles (see Chapter 4, section 4.5.3 for the structural taxonomy used in the present study).

Table 2.1. Structural classification of lexical bundles in academic prose (Biber *et al*., 1999: 1015–1024)

Structure	Examples
Noun phrase with *of*-phrase fragment	*the end of the, the beginning of the, the base of the*
Noun phrase with other post-modifier fragment	*the way in which, such a way that, the fact that the*
Prepositional phrase with embedded *of*-phrase fragment	*by the end of, as a result of, as a matter of*
Other prepositional phrase fragment	*at the same time, in contrast to the, in the sense that*
Anticipatory *it* + verb phrase/adjective phrase	*it is impossible to, it is important to, it is not surprising*
Passive verb + prepositional phrase/adjective phrase	*are shown in the table, is based on the, be taken into account*
Copula *be* + noun phrase/adjective phrase	*is one of the, was no significant difference*
(Verb/phrase +) *that*-clause fragment	*should be noted that, does not mean, has been suggested*
(Verb/adjective +) *to*-clause fragment	*are likely to be, will be able to, should be able to*
Adverbial clause fragment	*as shown in the figure, as we have seen, if there is a*
Pronoun/noun phrase + *be* (+…)	*this is not the, there was no significant*
Other expressions	*as well as the, may or may not, than that of the*

2.2.3.2 Functions of bundles

In addition to the structural classification of lexical bundles, some attempts have also been made to classify them functionally. Cortes (2004: 400) states that the functions in previous taxonomies “refer to meanings and purposes of the language, functions that try to provide texture or organize the discourse according to contexts or situations”. For example, lexical bundles such as *on the other hand* are used to express a textual function, which is concerned with the meaning of the sentence, establishing contrastive relations between elements.

Similarly, Nattinger and DeCarrico (1992) view lexical phrases or lexical bundles as parts of language that often have defined roles in shaping the overall discourse and help in signalling the direction of discourse, whether they are spoken or written texts. Nattinger and DeCarrico (1992: 59) also add that “lexical phrases represent various categories of meaning and pragmatic characteristics of discourse and conversational structure that exist in many different types of situations”.

To illustrate, Biber *et al*. (2004) designed a preliminary taxonomy, which indicates the meanings and purposes of lexical bundles within written and spoken texts and classifies them according to three main functions: (1) stance expressions, which are used to convey the writer's or speaker's attitude (e.g. *I don't know why*, *are more likely to*); (2) discourse organizers, which are employed to structure texts by reflecting relationships between prior and coming discourse (e.g. *on the other hand*, *as a result of*); (3) referential expressions, which are used to structure the writer's experience and establish their way of looking at things (e.g. *at the beginning of*, *at the same time*). Table 2.2 shows Biber *et al*.'s (2004) taxonomy. This initial taxonomy for analysing the functions of bundles has been widely adopted, utilized, and even modified, in many studies, such as Ädel and Erman (2012), Chen and Baker (2010), Cortes (2004, 2006), Csomay (2012), Hernandez (2013), and Wei and Lei (2011).

Table 2.2. Functional classification of lexical bundles in academic prose (Biber *et al*., 2004: 384–388)

Stance expressions Express attitudes or assessments of certainty that frame some other proposition	Discourse organizers Reflect relationships between prior and coming discourse	Referential bundles Make direct reference to physical or abstract entities, or to the textual context itself	Conversational functions
(A) Epistemic stance *the fact that the* (B) Attitudinal/modality stance (B1) Desire *I don't want to* (B2) Obligation/directive *it is important to* (B3) Intention/prediction *it is going to be* (B4) Ability *to be able to*, *can be used to*, *it is possible to*	(A) Topic introduction/focus *if you look at* (B) Topic elaboration/clarification *on the other hand*, *as well as the*	(A) Identification/focus *is one of the* (B) Imprecision *or something like that* (C) Specification of attributes (C1) Quantity specification *the rest of the* (C2) Tangible framing attributes *the size of the* (C3) Intangible framing attributes *the nature of the* (D) Time/place/text reference (D1) Place reference *in the United*	(A) Politeness *thank you* (B) Simple inquiry *what are you doing?* (C) Reporting *I said to him/her*

	States (D2) Time reference *at the same time* (D3) Text deixis *shown in Figure three* (D4) Multifunctional reference *at the end of*

Hyland (2008a) extended and modified this framework when he investigated the frequency, structures and functions of lexical bundles in his study of a large corpus. He made further modifications to Biber *et al.*'s (2004) functional classification by creating sub-categories that better describe the lexical bundle functions that he identified in his large corpus. Similarly, in terms of modification, Simpson-Vlach and Ellis (2010) credit the work of Biber *et al.* (2004) and developed a categorization scheme which is an adaptation of the functional taxonomy outlined in Biber *et al's.* (2004) work. Their purpose for developing the functional taxonomy is to achieve pedagogical goals and for it to be used in EAP curricula. Table 2.3 presents Simpson-Vlach and Ellis's (2010) functional classification and its sub-categories, which is significant for the present research.

Table 2.3. Functional classification of lexical bundles in academic prose – AFL (Simpson-Vlach and Ellis, 2010: 498–502)

(A) Referential expressions	(B) Stance expressions	(C) Discourse organizing functions
(1) Specification of attributes (a) Intangible framing attributes *in the course of*, *in accordance with the* (b) Tangible framing attributes *as part of the*, *an increase in the* (c) Quantity specification *there are a number of*, *a wide range of* (2) Identification and focus *as can be seen*, *that there is no*	(1) Hedges *is likely to be*, *it is likely that* (2) Epistemic stance *assumed to be*, *be considered as* (3) Obligation and directive *it should be noted*, *take into account the* (4) Expressions of ability and possibility *it is possible that*, *can be found in* (5) Evaluation *it is important to*, *it is clear that* (6) Intention/volition,	(1) Metadiscourse and textual reference *in the next section, in this paper we* (2) Topic introduction and focus *for example if, what are the* (3) Topic elaboration (a) non-causal *are as follows, in more detail* (b) Topic elaboration: cause and effect *due to the fact*, *for the purposes of* (4) Discourse markers *even though the*, *in*

(3) Contrast and comparison *be related to the*, *on the other hand*	prediction *to do so, we do not*	*conjunction with*
(4) Deictics and locatives *at the time of*, *b and c*		
(5) Vagueness markers *and so on*		

In this study, after identifying the most frequent four-word bundles extracted from the English for Academic Purposes (EAP) corpora, I will use Simpson-Vlach and Ellis's (2010) framework as a guide to functionally analyse the identified bundles found in the EAP materials (see Chapter 4, section 4.5.2 for the functional taxonomy employed in the present study).

The issue of multifunctionality, which refers to a lexical bundle having more than one function, is an important issue that needs to be considered when attempting to functionally classify formulaic sequences. In Biber *et al.*'s (2004) study, influenced by other researchers such as Halliday (1978), lexical bundles that served similar functions were grouped together according to their meanings and the uses of each bundle in its discourse contexts. So, when the lexical bundles were grouped together, Biber *et al.* (2004: 383) needed to decide "the discourse functions associated with each of the groups". However, according to Biber *et al.* (2004), in some instances, an individual bundle or sequence could have multiple functions even in a single occurrence. For example, the bundle *take a look at* can carry two functions: directive and topic introducer. In other instances, a single bundle may serve different functions based on the context in which it appears. For example, the bundles *at the beginning of* the and *at the end of* can be time-, place-, or text-deictic references due to their textual environment.

As reported by Simpson-Vlach and Ellis (2010: 497), "the creation of a functional taxonomy for formulaic sequences is an inherently problematic endeavour", due to a proliferation of types and subtypes. Simpson-Vlach and Ellis (2010: 502) also note that "this proliferation of categories does indeed make it difficult to distil the data into a compact functional model applicable across corpora and domains of use". The main problem when assigning functional classifications to bundles is that there are no fixed criteria provided in the literature on how to decide which sub-category a single bundle should belong to (Ädel and Erman, 2012). They argue that there are sub-categories that are clear and well-formed by previous studies (e.g. *topic introduction*, *quantifying*), whereas other sub-categories are considered "vague" (e.g. *identification / focusing*, *framing*). Ädel and Erman (2012) argue that this vagueness in the sub-categorization process has been accompanied with several inconsistences, as can be seen in previous work. For example, they report that the sub-category *focusing* was labelled under the category *Discourse organising* in Chen and Baker's (2010) study; however, Biber *et al.* (2004) and Simpson-Vlach and Ellis's (2010) studies both labelled the sub-category *focusing* under the *referential* category.

Thus, the proposed solution by many researchers, such as Ädel and Erman (2012), Biber *et al.* (2004: 384), Chen and Baker (2016), and Simpson-Vlach and Ellis (2010), is that when in doubt of the function of a lexical bundle, it is recommended to examine concordance lines of potentially multifunctional bundles and categorize them based on "their most common use". Ädel and Erman (2012: 90), as a solution to the multifunctionality of their lexical bundles, note that: "[d]espite considerable difficulty, we were able to improve the initial 66% agreement rate through extensive discussion, checks of bundle classifications in the literature, and the contextual analysis of concordance lines and finalise the classification with almost 100% agreement". They add, "it is not clear, however, to what extent our understanding of some of the categories matches that of other researchers".

This gives a clear picture of the level of difficulty regarding the multifunctionality of lexical bundles. Therefore, the present study adopts a clear and inductive approach to finalize the functional classification. Being aware of such a problematic issue, in the present study the assignment of functions is based on two main aspects: concordance checks and the categorization of Simpson-Vlach and Ellis's (2010) taxonomy. This is because EAP learners are more likely to encounter lexical bundles found in their EAP materials, mostly depending on the textual environment in which the bundles occurred. Therefore, basing the decision regarding bundle functional assignment on "the most common use" is not always a straightforward solution. This is because, for the present research, in some cases even a single lexical bundle with its functional data may be presented in the material of a particular EAP teacher and encountered by a group of EAP learners. However, the same lexical bundle (with a different function, depending on the textual environment the bundle appeared in) might be encountered by a different group of EAP learners, having been presented in the materials of a different teacher. Any challenging or problematic issues that may arise from the functional classifications as well as the issue of multifunctionality will be fully discussed in Chapter 11 in the limitations section (11.1.1).

2.3 Lexical bundles and metadiscourse

Regarding lexical bundles, it has been established in the literature that the lexicon of a language does not only contain a vast list of words, but it also contains a much larger range of items: sequences, phrases, idioms, and even complete expressions. Nowadays, it is widely accepted, as previously mentioned in the present study, that language is largely formulaic, pervading most language use and constituting a great proportion of any discourse (Biber *et al*., 2004; Erman and Warren, 2000; Schmitt and Carter, 2004). "This store of formulaic sequences is dynamic and constantly changing to meet the needs of the language user" (Wray, 2002: 101). The pervasive spread of the notion of formulaicity led researchers and linguists to identify and study lexical bundles (Biber *et al*., 1999; Biber *et al*., 2004; Cortes, 2004, 2006; Hyland, 2008a, 2008b). Therefore, together, metadiscourse and lexical bundles became two interesting phenomena in language production and language learning that are significantly linked to the lexical view of language which is embedded in Hoey's (2005) theory of words and language called "Lexical Priming". Metadiscourse and lexical bundles play a vital part in language discourse under the theory of lexical priming. Hoey's (2005: 12) theory, *Lexical Priming: A new theory of words and language*, is based, as stated in the title, on a new view of language, describing the existence of words, the collocation and combination of words, and drawing on evidence from language corpora.

Metadiscourse, a widely used term, is one aspect of understanding language use that usually surfaces when surveying and analysing lexical bundles. Both metadiscourse and lexical bundles are significantly linked to Hoey's (2005) theory of words and language called *Lexical Priming*. One important point to bear in mind is that metadiscourse and the underpinning theoretical background – *Lexical Priming* – are clearly regarded as two major concepts in their own right; therefore, it is impossible and inapplicable to discuss all their aspects. Instead, more importantly, a practical application for this study would be to consider relevant and related elements of these two important concepts to lexical bundles, focusing on the academic written register, and involving two key issues: definition and function.

2.3.1 The issue of definition

In a general sense, the dynamic view of language embodies the idea of how people come to use lexicon and grammar for communication. In its core essence, language is perceived as a tool for communication and social interaction (Chomsky, 1964). This means that language users, as they write, articulate and construct different interactions, utilize lexicon and grammar to present their ideas, attitudes, feelings and experiences, adapting to different social and academic contexts.

In this regard, according to Hyland (2005), metadiscourse can be viewed as an important element of this dynamic view of language because language users employ explicit metadiscourse devices (Hyland, 2017), expressing and constructing interactions when negotiating with others to make decisions on the kind of impact they wish to have on their readers. Under such a concept, metadiscourse is viewed as extra language or linguistic resources, which language users employ to express and frame the propositional content with regard to the standards and values of a particular academic discourse community. Hyland and Tse (2004: 156) give an explanation of metadiscourse with regard to the writer/reader connection:

> Metadiscourse is self-reflective linguistic material referring to the evolving text and to the writer and imagined reader of that text. It is based on a view of writing as social engagement and in academic contexts reveals the ways that writers project themselves into their discourse to signal their attitude towards both the propositional content and the audience of the text.

For example, Hyland (2010), in a corpus of 240 doctoral and master's dissertations, studied the work of advanced second-language (L2) writers' use of metadiscourse. The L2 writers were students from five different Hong Kong universities and were selected from six different academic disciplines: electronic engineering, computer science, business studies, biology, applied linguistics, and public administration. The study suggests that advanced L2 writers and members from different disciplines use metadiscourse language to represent themselves in quite distinctive ways. It also revealed that metadiscourse is used on the basis of the writers' evaluation of their readers. This means that advanced L2 writers' decisions regarding elaboration and reader involvement are fulfilled by providing sufficient signals to make sure that their readers understand and accept the propositional content. For example, in metadiscourse, words such as *might*/*perhaps*/*possible* are called hedges, which involve the reader in the argument by withholding the writer's full commitment to the proposition. The following is an example of metadiscourse usage provided by Hyland's (2010: 133) study: "It is *possible* that instruction in one would lead to increased ability in the other".

Metadiscourse, then, is a concept that seems to offer a means of organizing academic texts. Crismore *et al.* (1993) and Vande Kopple (1985: 83) state that "on the level of metadiscourse, we do not add propositional materials but help our readers organise, classify, interpret, evaluate, and react to such materials". In addition, Hyland (2005: 14) views metadiscourse as a key factor in discourse; thus, without it, the readers could not "contextualize a text" and writers may not be able to communicate their ideas effectively. However, Hyland and Tse (2004: 174) argue that metadiscourse should not be viewed and treated as an "independent stylistic device" from which language users can obtain a stock of varied forms. The underlying rhetorical dynamics of metadiscourse is an important factor, which relates metadiscourse to the context in which it appears. Metadiscourse is, then, the connection between the standards and codes of a specific discipline or community and the writer's ability to generate adequate signs to allow the reader to understand and accept the propositional content (Hyland and Tse, 2004).

A study conducted by Khedri *et al.* (2013) validated Hyland and Tse's (2004) claim that the significance of metadiscursive features as textual devices is based on their relationship with the contexts in which they occur and with the conventions of a particular professional discourse community. In their study, the authors investigated abstracts of 60 research articles (RA) from two different disciplines: applied linguistics and economics. The findings revealed a major difference between the two disciplines in terms of how RA abstract writers organize their argument, employing *transition markers* to guide readers to their intended argument. For example, they found that both applied linguist and economist abstract writers planned *transition markers* in the same way for communicative or discursive reasons but they used them differently. *Transition markers* found in RA abstracts written by applied linguists were employed to ease cognitive relations between information breaks, using *and*, *also* and *as well as*. Economists in their abstracts favoured *transition markers*, using comparative devices such as *however* and *but* to indicate logical relations. The following are examples from Khedri *et al.*'s (2013: 326) study, presenting both applied linguists' and economists' use of *transition markers*:

> The interpersonal categories were *also* broken down into subtypes depending on the linguistic items used, *and* analyzed for distribution in the conclusion sections. (From applied linguists)

> The use of e-commerce does nothing to boost entry into export markets, *but* the intensity of its use is associated with increased export intensity. (From economists)

Hoey (2005) suggests that language users acquire vocabulary and language through several features of lexical priming. According to Hoey, every word is primed for use in discourse, depending on the context (specific situations), which includes linguistic and contextual information, and the social interaction which language users are repeatedly involved in. This means that every time language users come across a single word or a phrase, they store it along with all the information and associations of the kind of context in which it appeared (e.g. semantic, pragmatic, grammatical, textual associations etc.). The notion of priming includes a number of features. The first feature of lexical priming is what language users are aware of or are "primed" for (using prior experience or knowledge of words). A second feature is to expect words to be in the parts of other words (collocation). A third feature is to expect words to appear in certain grammatical situations (grammatical colligation). The fourth feature is to expect words to be in particular position in the text and discourse (textual colligation) etc.

These features connect us to the concept of lexical bundles and metadiscourse, which offers a way of understanding and analysing language use. According to Hoey (2005: 129), "every lexical item (or combination of lexical items) is capable of being primed (positively or negatively) to occur at the beginning or end of an independently recognised 'chunk' of text". Hoey claims that when we meet language, for example, in writing, we are aware of the 'contexts and co-texts' which we come across. Hyland (2008b: 6) agrees with Hoey that lexis is "systematically structured through repeated patterns of use". Hyland believes that our knowledge of a word is based on our encounters with it. Thus, when we try to express what we want to say, the choices of words and phrases are formulated by the way we repeatedly encounter them in a similar context. Flowerdew (2012: 177) supports Hoey's theory of lexical priming in which the theory "maps out a theoretical relationship between lexis and textlinguistics, showing how semantic associations, collocation and colligation operate at a discoursal level". Moreover, according to Flowerdew (2012), at a textlinguistic level, metadiscourse and lexical bundles are considered two types of linguistic devices which carry discourse functions. The functional factor of metadiscourse in relation to lexical bundles is another important issue that will be discussed in the next section.

Furthermore, frequency counts (see section 2.2.2.1) and the concept of entrenchment (see section 2.1.2) are also associated with lexical priming in terms of acceptability. Entrenchment suggests that due to repeated contact and usage, forms and constructions are set and produced, and that these entrenchment items form the basis of users' linguistic structure (Bermel and Knittl, 2012). Acceptability judgments are found by many researchers to be a useful and rich source of linguistic data. Understanding how acceptability judgements relate to frequency counts is significant in providing evidence for lexical priming related to lexis and linguistic structures. For instance, there is a correlation between acceptability and frequency counts (Quirk *et al.,* 1985). According to Quirk *et al.* (1985: 33), generally, native speakers may judge a [phrase] or a sentence to be unacceptable because they consider these forms or lexical items to be logically "absurd, cannot find a plausible context for its use, or they sound clumsy or impolite". On this basis, the phenomenon of entrenchment suggests that language users' linguistic knowledge is continuously revived and reorganized, which is encouraged by social interactions. Frequency counts play a central role in the occurrence and entrenchment of linguistic units. For example, higher-frequency words are acknowledged more quickly than lower-frequency words, as shown by Cattell (1886). In addition, Bybee and Eddington (2006) found that high-frequency sequences will be judged as more acceptable than those of low-frequency sentences. Similarly, Bybee and Hopper (2001) claim that language users' experience and frequent exposure to a certain form makes these forms easier to access and use.

Hoey (2005) also adds that the items and sequences that language users usually encounter obtain the 'cumulative effects' of these encounters, which become part of the language users' knowledge or background of the word or combination that regularly co-occurs with certain other words or with certain structures and functions. For example, the word *word* collocates with the word *say*, forming the phrase *say a word*. The phrase *say a word* in turn collocates with the word *against*, producing the phrase *say a word against*. According to Hoey (2005: 11), in this way, "many lexical bundles are created and produced". Sinclair (2004) and O'Donnell *et al.* (2013), similarly, state that language is rich in collocational and colligation constraints and semantic patterns, and that the phrase is the basic building block of language representation where form and meaning connect with high dependability. It is within such a theoretical background that both lexical bundles and metadiscourse are considered an important part of language production such as academic writing.

2.3.2 The issue of function

The issue of function is vital here and is considered as a key determiner in approaching metadiscourse and lexical bundles. One common thread between metadiscourse and lexical bundles lies in the fact that researchers and linguists seek to survey various texts to understand, describe, explore and analyse language-in-use. Methodologically, metadiscourse and lexical bundles may be studied as an interlinked concept or as two independent notions, each involving its own classifications, sub-categories and functional analysis. As established in the definition section above, lexical bundles (e.g. Jablonkai, 2010; Salazar, 2011, 2014; Wei and Lei, 2011) and metadiscourse (e.g. Alyousef, 2015; Aull and Lancaster, 2014; Khedri *et al.*, 2013; Kuhi and Behnam, 2011) are two types of linguistic devices with discourse functions that are commonly used in corpus studies (Flowerdew, 2012). A more distinctive feature is the fact that metadiscourse is a functional category that examines different linguistic items, including single words, sequences, exclamation marks, etc., and "allows writers to specify the inferences that they wish their readers to make" (Barton, 1995: 219).

For example, metadiscourse tends to incorporate a wide range of linguistic items, ranging from verbs such as *explain*, *claim*, quotes such as *I admit that the term 'error'*, single words such as *I*, *first*, to short phrases such as *I believe*, short clauses such as *could potentially*, and long sequences such as *on the other hand* (Crismore *et al.*, 1993; Hyland and Tse, 2004). Within metadiscourse, a set of lexical bundles function in an almost identical manner to individual words, matching a particular meaning or function to a form, although that form involves multiple orthographic or phonological words (Martinez and Schmitt, 2012). For example, the sentence "*Increasingly, extreme weather events indicate that climate change is upon us"* has the meaning or concept of "a situation becoming noticeably prevalent" (Martinez and Schmitt, 2012: 299). Martinez and Schmitt explain that this concept or meaning can not only be realized by the single word *increasingly*, but could be equally well realized and replaced by the formulaic sequence – lexical bundle: *more and more*.

As emphasized in this chapter, formulaic sequences are considered by many linguists and researchers (e.g. Biber *et al.*, 1999; Biber *et al.*, 2004; Cortes; 2004, 2006; Hyland, 2008a, 2008b; Martinez and Schmitt, 2012; Wray, 1999) as linguistic items that involve the relationship or association of two or more words that represent formulaicity such as a collocation, lexical bundles, idioms, etc. Thus, it is safe to conclude that not all linguistic items of metadiscourse are considered formulaic sequences. For example, metadiscoursal items presented in single words are not perceived as formulaic sequences because formulaicity involves more than one word working together.

However, a key matter needs to be addressed, that is, all types of formulaic sequences such as the collocation *red carpet* and lexical bundle *on the other hand* may be considered to be linguistic devices that can be analysed under the concept of metadiscourse. There are a range of ways in which language users achieve some sort of interactional goal, using lexical bundles with functions of linking ideas, structuring and organizing the text, and conveying the writer's attitudes and evaluations. In this respect, these linguistic items of formulaic sequences – lexical bundles – are interlinked with metadiscourse.

From a functional approach, metadiscourse includes linguistic devices that carry different discourse functions:

> Metadiscourse is an intuitively attractive concept as it seems to offer a motivated way of collecting under one heading the range of devices writers use to explicitly organize their texts, engage readers, and signal their attitudes to both their material and their audience (Hyland and Tse, 2004: 156).

Thus, the present study adopts the perspective of metadiscourse, in accordance with Hyland and Tse (2004) and Hyland (2005, 2010), which is perceived to have its own associated frameworks or taxonomies. Metadiscourse embodies two traditional positions in terms of functions. The first position adopts a narrow description, labelled the *reflexive model*, which is represented by the research of Ädel (2006, 2010a), Ädel and Mauranen (2010), and Mauranen (1993). The second position offers a broad description, which is called the *interactive model*, which is represented by the research of Hyland and Tse (2004) and Hyland (2010). According to Hyland (2010), in the *reflexive model* of metadiscourse, the concept is restricted to focusing on certain elements, referring to the text itself and examining features of discourse which help to arrange the text as a text. In this model, adopted by Ädel (2010a), writers focus on elements of rhetorical organization by using only those text features which refer to the text itself (e.g. the expression: *this will be discussed in the next chapter*).

In contrast, following the research of Vande Kopple (1985) and Crismore *et al.* (1993) and others, Hyland (2010) describes the *interactive model*, as can be seen in Table 2.4. This model includes a range of devices which writers use to shape their texts, appeal to their readers, and signal their positions in their materials and to their readers. Hyland (2010) argues that this model is more encompassing and contains a coherent set of interpersonal choices. These frameworks are considered to be different from those classified for lexical bundles. Table 2.4 shows the functional categorization of metadiscourse according to Hyland (2010).

Table 2.4. A model of metadiscourse in academic texts (Hyland, 2010: 128–129)

Category	Functions	Examples
Interactive	Help to guide reader through the text	Resources
Transitions	express semantic relation between main clauses	*in addition / but / thus / and*
Frame markers	refer to discourse acts, sequences, or text stages	*finally / to conclude / my purpose is*
Endophoric markers	refer to information in other parts of the text	*noted above / see Fig / in section 2*
Evidentials	refer to source of information from other texts	*according to X / (Y, 1990) / Z states*
Code glosses	help readers grasp meanings of ideational material	*namely /e.g. / such as / in other words*
Interactional	Involve the reader in the argument	Resources
Hedges	withhold writer's full commitment to proposition	*might / perhaps / possible / about*
Boosters	emphasize force or writer's certainty in proposition	*in fact / definitely / it is clear that*
Attitude markers	express writer's attitude to proposition	*unfortunately / I agree / surprisingly*
Engagement markers	explicitly refer to or build	*consider / note that / you*

	relationship with reader	*can see that*
Self mentions	explicit reference to author(s)	*I / we / my / our*

For analytical reasons, Hyland (2010) distinguishes interpersonal choices into two elements: (1) *interactive resources* and (2) *interactional resources.* By using such metadiscourse devices, the writer using (1) the *interactive resources* is expected to guide the reader through the text. For example, the prepositional phrase *in addition* is a metadiscourse device that is categorized as a logical connective, which has the function of linking ideas in the text and referring to a writer's interpretation of his or her unfolding text. The writer using (2) the *interactional resources* involves the reader in the text or argument. For example, the endophoric markers found in phrases such as *in this paper* and *I explore* signal the connections of information found in the different parts of texts. These metadiscourse devices are very common in the academic genre (Ädel, 2010a).

Within the academic register, metadiscourse across different genres has been investigated. For example, Hyland (1999) investigated extracts from 21 introductory university textbooks across three academic disciplines: microbiology, marketing, and applied linguistics, and compared them with a similar corpus of research articles (RA), in terms of metadiscourse features. He wanted to ascertain how textbook authors introduce themselves and construct and signal their arguments to their readers.

The most important finding, which is useful for the present study, is that in the RAs writers used metadiscursive language to address their readers as experts, in order to show solidarity by drawing on shared knowledge. Textbook authors, in contrast, were found to employ metadiscursive language differently from RA writers. Textbooks used sophisticated metadiscursive devices to inform, clarify and order their material, and to explain propositional connections to their readers. In this way, according to Hyland, learners seeking information from research sources such as textbooks are faced with limited rhetorical guidance and are not encountering the written forms found in most research. Furthermore, he states that the different patterns of metadiscourse in the two corpora were used differently. For example, an interpersonal stance was commonly found in textbooks; the authors used this to highlight a professional stance towards both information and the readers. RAs used both textual and interpersonal metadiscourse, which authors employed to exclude non-experts in the field and to allow the authors to control the information they created.

The functions of lexical bundles, in contrast to metadiscourse, are viewed or examined in terms of lexico-grammatical features, with different frameworks being used, as previously mentioned (see section 2.2.3.2 of this chapter for functional classifications). In the present study, in the analysis (see Chapters 5 and 6) and discussion parts (see Chapter 9), aspects of lexical bundles in terms of their metadiscursive and textual reference features are examined but within lexical bundle functional classifications. Defining and presenting functional aspects of metadiscourse in this way will shed light on its distinctive features and draw the line between lexical bundles and metadiscourse. At the same time, understanding metadiscourse in terms of the writer/reader relationship, and making relevant connections to its functional model, when needed, may be useful for the present study to enrich the interpretations and draw conclusions.

2.4 The role of corpora in analysing sequences and bundles

2.4.1 A brief overview of corpus linguistics

In recent years, computer technology and sophisticated software tools have paved the way for linguists and researchers to compile large collections of texts of natural language, thereby forming and developing corpus linguistics as a methodology for language analysis. Corpus linguistics investigates linguistic phenomena through the use of large electronic collections of machine-readable text called corpora (Granger, 2002). Accordingly, a corpus is a large collection of naturally occurring texts that are electronically stored and processed by special computer software. It provides access to naturalistic information and descriptions of language (Sinclair, 2005).

The body of texts in a corpus is an accumulation of spoken and written material from different genres such as newspapers, interviews and published research articles. The British National Corpus (BNC), COBUILD/Birmingham Corpus, and the British Academic Written English Corpus (BAWE) are some examples of well-known corpora, in which authentic materials are available for researchers, teachers, linguists, and even students. In addition, ConcGram, WordSmith, and AntConc are corpus linguistic software packages that are especially designed to help linguists and researchers analyse their lexical data by providing different text-handling functions such as word lists, concordancing, and keyword and cluster analysis (Oakes, 1998).

2.4.2 The issue of the size and representativeness of corpora

The strength of a corpus comes from its ability to help the researcher answer a wide range of questions related to various aspects of language use. In this regard, the representativeness of a corpus is known to be one of the most crucial issues in corpus design. This is because the goal of many corpus studies is focused on identifying quantitative linguistic patterns in corpus samples and on generalizing those results to apply to a larger linguistic population. In corpora creation, representativeness refers to "the extent to which a sample includes the full range of variability in a population" (Biber, 1993: 243). However, representativeness of a corpus can be achieved by balancing a number of issues, including the kinds of texts, the number of texts, the selection of particular texts, the selection of text samples from within texts, the length of text samples, and implementing appropriate sampling.

Representativeness is closely connected to corpus size (Reppen, 2010) and it is seen as one of the commonly defining features of a corpus. The key for deciding the appropriate size of a corpus is highly dependent on the purpose and questions of each research study (see Sunderland, 2010), the kind of texts being included and analysed, and the number of words considered sufficient to obtain valuable results and achieve representativeness (Biber, 1993; Biber *et al.*, 1998; Reppen, 2010).

According to Reppen (2010), representativeness (the issue of whether there are enough words to accurately represent the type of language being studied) is a key factor in resolving the issue of corpus size but not the only factor. In the past, there have been some efforts to support large over small corpora. Sinclair (1991: 18) stresses that "corpus should be as large as possible, and should keep on growing". He adds "In order to study the behaviour of words in texts, we need to have available quite a large number of occurrences". Large corpora are needed for lexicographical goals, while small specialized corpora are valued for their pedagogical purposes (Nelson, 2010). For example, a large corpus is extremely practical if the main aim of the research is to create general-purpose dictionaries, which entails looking at language as a whole and requires the compilation of billions of words to obtain an adequate description of the system of naturally occurring language. The Collins COBUILD English Language Dictionary, which was developed by John Sinclair, is a perfect example of a corpus-based dictionary.

While Sinclair (2004) stressed for years the notion that bigger is best, like many linguists he also gave credit to smaller specialized corpora (Baker 2006; Koester, 2010). In recent years, smaller and specialized corpora have been gaining much attention, and have a unique and distinctive role in corpus linguistics. The most important considerations when designing a small and specialized corpus is that it must be representative and balanced. A *balanced or sample corpus*, according to McEnery and Hardie (2012), refers to building a corpus that represents a specific type of language over a given duration. In a *balanced or sample corpus*, the researcher seeks to include samples of data that are "balanced and *representative* within a particular *sampling frame* which defines the type of language, the *population* that we would like to characterise" (McEnery and Hardie, 2012: 8; emphasis in original text). This means that it is important ensure that the sample being examined is representative of the language under consideration.

Simply put, samples should be gathered from a range of fairly typical situations. So, for example, if we want to look at the written language used by first-year undergraduate learners across science-oriented classes such as biology, chemistry, and physics during a course of one term, the *sampling frame* is apparent. We would only take written data into our corpus, which represents the essays of first-year undergraduate learners across these classes – biology, chemistry, and physics during the first term. However, if we only took written data gathered in biology classes, we would not get an accurate set of data for the science-oriented population. Instead, much of the data would be of context-specific vocabulary targeted at the biology discipline such as *cell*, *tissue* and *bacteria*, leading to a misrepresentation of the science-oriented writing community. In this regard, "corpus size is determined by capturing enough of the language for accurate representation" (Reppen, 2010: 32).

One disadvantage from working with large corpora is that the vast amount of data from high-frequency items may be unmanageable to work with, resulting in the analysts working with a sub-sample (Koester, 2010). Sinclair (2004: 189) argues that working with small corpora "is simply a limitation". This means that the results from small corpora will be limited. However, in a smaller corpus, all of the frequent items and not random samples of highly frequent items can be examined (Koester, 2010). A good example of a small corpus comes from Li and Schmitt's (2009) study, where they built a corpus from the work of one student. They conducted a longitudinal case study to report on a Chinese MA student's acquisition and use of lexical bundles within one full academic year. The results showed that the learner depended greatly on a limited range of bundles, which at times were considered non-native-like.

A smaller corpus is valued for its pedagogical objectives. Unlike a larger corpus, a smaller corpus is suitable for studying frequent lexical items which tend to generate reliable results. For example, grammatical items, prepositions, and formulaic sequences such as lexical bundles are commonly investigated in a small corpus. Moreover, smaller corpora can be more specialized and genre-related, looking into different fields of study including pedagogy, language learning, EAP and ESP language (Römer, 2010; Koester, 2010) (e.g. Cortes, 2004, 2006; Simpson-Vlach and Ellis, 2010) and more focused and specific considerations such as classroom management (O'Keeffe *et al.*, 2007).

More importantly, small and specialized corpora can be planned to provide information on contextual features, including "information about the setting, the participants and the purpose of communication" (Koester, 2010: 67). This entails, according to Koester (2010), a connection between the corpus and the contexts in which text in the corpus appeared. Therefore, in a small corpus, the analyst with a certain degree will be familiar with the context or will have access to data regarding the contexts. As a result, corpus analysis can add qualitative findings to support, balance and complement the quantitative findings revealed by the corpus (Angouri, 2010; Koester, 2010; O'Keeffe *et al.*, 2007). Examples of studies that provided qualitative findings to balance and complement their quantitative findings are: Biber *et al.* (2004), Chen and Baker (2010), Cortes (2004), and Hyland (2008a, 2008b).

In the field of corpus linguistics, it is a commonly known feature that written corpora are considerably larger than spoken ones, which are normally small. Furthermore, O'Keeffe *et al*. (2007: 4) state that "for corpora of spoken language, anything over a million words is considered to be large; for written corpora, anything below five million is quite small". According to Flowerdew (2012: 19), "large scale, general-purpose corpora are generally in the range of 100 million to

500 million words, whereas more specialised, genre-related corpora can be from around 50,000 to 250,000 words".

Based on the literature, the corpora in the present study may be considered relatively small. The word count and other related issues are fully discussed in the methodology chapter (see Chapter 4). However, the arguments regarding the usefulness and adequacy of size and representativeness which are addressed here show how the data in this research can produce sufficient results, both quantitatively and qualitatively. Because there are no existing corpora that adequately represent my interest in exploring lexical bundles on a particular EAP course, focusing on the academic writing register, I needed to build two corpora. Regarding the degree of representativeness, samples from EAP materials (textbooks and instructors' materials) were compiled to build two corpora to investigate the frequent lexical bundles that EAP learners are more frequently exposed to during their writing classes in an EAP pre-sessional programme (see Chapter 4, section 4.2.2.1 for a discussion of representativeness in relation to the present study).

2.4.3 Influence of corpus studies on the intersection between lexis and grammar

Corpus studies have resulted in significant shifts in the theory and practice of teaching language, especially vocabulary and grammar. Researchers have been able to analyse patterns of language use by using corpora, leading to discoveries concerning lexico-grammar, which were once considered two separate areas.

In addition, corpus research has made it possible for linguists not only to explore choices of what to say but also how to say it. For example, Lewis (1997) highlights lexico-semantic knowledge, where the goal of the language user is to achieve successful communication. He also realizes the importance of words and their patterns, and how they may offer grammatical knowledge to language learners. This means that a basic principle behind the lexical approach (a language teaching approach adopted by Lewis, where the core of its teaching starts from the lexis) is that "language consists of grammaticalised lexis, not lexicalized grammar" (Lewis, 1993: 34).

Simply put, Lewis (2000) believes that lexical items have their own grammar, so his syllabus pays less attention to individual words and their grammatical structure. Instead, it highlights the teaching of collocations by paying more attention to their patterns and their surrounding context, which in turn provides more information on grammatical structure than is found in any traditional grammar syllabus. In this respect, it is a more "grammatical" approach than the traditional structural syllabus (2000: 150). Willis (1990) similarly believes that the grammar or structure of the language should be addressed from the point of view of the lexis, which will help learners focus on some complex features of language, such as the complex noun phrase, through exploring different grammatical patterns and/or functional realizations. Harwood (2002) also reports that one of the main concerns of the lexical approach is its determination to end the misleading grammar/vocabulary separation, which has remained as a great influence in the construction of ELT materials.

In a similar way but with a different perspective, corpus-based researchers approach language by accessing *language-in-use*, where the focus is on the patterns which are actually used by native speakers or non-native speakers for communication purposes. *Language-in-use* theory focuses on "what is actually done when a language is used for a particular type of communication" (Coxhead and Byrd, 2007: 130). To illustrate, Coxhead and Byrd (2007) explain that the verb *required* usually occurs in sequences which can be followed either by an *infinitive* or by a *that-clause.* In academic writing, for example, the word *required* is mostly followed by an infinitive, such as in the sequence *is required to make*. This means that the combination of the word *required* and its following word (in the infinitive form) is considered a lexico-grammatical pattern, which refers to a "frequently occurring combination of words and grammar, where a particular word requires particular grammar" (Coxhead and Byrd, 2007: 130). This led to many attempts to explain and illustrate the relationship and the interdependency of lexis and grammar.

Furthermore, in line with the discussion in this chapter on entrenchment and frequency, it is evident that entrenchment is considered a process that strengthens language users' linguistic knowledge. Both frequency and entrenchment are not only associated with cognitive studies (see section 2.1.2) and lexical priming (see section 2.3.1), as previously discussed, but also have their effects on linguistic structures or grammaticalization. In this regard, entrenchment provides a continuous, natural and useful explanation of how the constructional system is formed by observing language use. According to Hilpert and Diessel (2017: 57), “constructions that are used frequently become, overtime, more entrenched than constructions that are used only rarely”. They explain that activation occurs any time a connection is established between a linguistic form and a meaning, which is referred to as pair understanding. This activation affects how the form and meaning combination will be processed in the future. With every form/meaning usage, language users update their representation and knowledge of the construction. The degree of entrenchment depends on the frequency of activation for each construction and with each connection (Hilpert and Diessel, 2017). For example, corpus linguistic studies such as Stefanowitsch and Gries (2003) found irregularities in distribution with regard to the lexical elements the constructions appear in, some elements appearing much more frequently while others are almost absent. These irregularities “motivate the view that entrenched lexico-grammatical interrelations form an important part of linguistic knowledge” (Hilpert and Diessel, 2017: 58).

In addition, Biber *et al.* (1999), in a large-scale corpus-based research study, created the Longman Grammar of Spoken and Written English (LGSWE) from the Longman Spoken and Written English Corpus (LSWE), containing approximately 40 million running words of written and spoken texts. It is considered a corpus-based grammar, and provides a comprehensive basis for the analysis of grammatical patterns. This project shows that when native speakers use language, they combine their knowledge of words (lexis) with their knowledge of grammatical patterns, in line with Hoey’s (2005) theory. These two aspects of language interact in “lexico-grammatical patterns” (Biber *et al.*, 1999: 13). In the present study, this interrelationship between lexis and grammar plays a key role in analysing and providing a qualitative interpretation of lexical bundle structures.

2.4.4 Contributions of corpus studies to the analysis of sequences and bundles

The introduction of corpus linguistics research has brought about insightful findings and substantial support to the analysis of formulaic sequences, together with statistically fruitful data. Corpus linguistics research has revealed that natural language is composed of a considerable number of formulaic sequences (Ellis, 2008; Ellis *et al.*, 2008; Wray, 2002). According to Schmitt and Carter (2004) and

Nattinger and DeCarrico (1992), formulaic sequences form a large proportion of any discourse. Furthermore, corpus linguistics analysis of large collections of text provides statistical analysis and empirical evidence on the presence and use of formulaic sequences.

For example, Erman and Warren (2000) found that formulaic sequences of various types constituted 58.6% of the spoken English discourse and 52.3% of the written discourse that they analysed in their corpora. Foster (2001) estimated that 32.3% of the unplanned native speakers' speech that they analysed consisted of formulaic sequences. Similarly, Altenberg (1998) estimated that around 80% of the language that is produced by a native speaker could be formulaic. Interestingly, Erman and Warren (2000) also estimated that approximately half of the text produced by native speakers is constructed in accordance with the idiom principle, which entails that instead of native speakers consistently creating new phrases of individual words, they often rely on formulaic sequences.

2.4.5 Methodological approaches to investigating sequences and bundles

Generally, there are two distinctive approaches to investigating formulaic sequences or lexical bundles in large text corpora: the corpus-driven approach and corpus-based approach (see Tognini-Bonelli, 2001: 84–87). Corpus-driven studies of formulaic sequences and lexical bundles exploit the potential of a corpus to depict linguistic categories and to explore elements that have not been previously covered. Tognini-Bonelli (2001: 84, 87) states that in a corpus-driven analysis, the "descriptions aim to be comprehensive with respect to corpus evidence" so that even the 'linguistic categories' are extracted "systematically from the recurrent patterns and the frequency distributions that emerge from language in context".

A study of lexical bundles (e.g. Biber, 2009) is considered to have a corpus-driven methodology when the corpus functions as an empirical base from which researchers derive their data and identify uncovered linguistic constructs, having only limited prior assumptions and expectations about the data being extracted (see Tognini-Bonelli, 2001). This means that in corpus-driven studies any conclusions are made solely on the basis of corpus observations and sequences and are discovered through corpus analysis.

Biber's (2009: 301) corpus-driven study is based on identifying the most common formulaic sequences in conversation and academic writing. The study aimed at focusing on the different pattern types of the two registers. He examined empirically the patterns represented by formulaic sequences across the two registers by only investigating frequently occurring sequences of simple word forms. He also identified the same basic contrast between speech and writing. In both speech and written academic discourse, formulaic sequences are considered very important. However, these sequences are observed in very different forms linguistically. In conversation, for example, fixed sequences are represented in clause fragments such as *I don't know if*, while academic writing consists of formulaic sequences that contain noun phrases and prepositional phrase fragments (*the extent to which*, *on the other hand*) (Biber, 2009: 301).

In contrast, the second approach is the corpus-based one, which the present study adopts as a method for investigating bundles. The corpus-based approach uses a selected corpus as an inventory of language data. Biber (2009: 276) states that "corpus-based research assumes the validity of linguistic structures derived from linguistic theory". The main goal of corpus-based research is to analyse the appropriate material that is extracted from the corpus to either: (1) allow linguistic structures to be quantified, or (2) to find evidence for existing theories or to retrieve explanatory examples. In other words, the researcher pre-selects sequences/bundles and then analyses the underlying corpus to investigate how these sequences are used to confirm linguistic pre-set explanations and assumptions (Biber, 2009: 276). Thus, the corpus acts as an information supplier, by providing additional supporting information.

> In this case, however, corpus evidence is brought in as an extra bonus rather than as a determining factor with respect to the analysis, which is still carried out according to pre-existing categories; although it is used to refine such categories, it is never really in a position to challenge them as there is no claim made that they arise directly from the data (Tognini-Bonelli, 2001: 66).

In addition, corpus-based studies employ automated, frequency-driven means and other criteria for extracting formulaic sequences (see section 2.2.2), which allow researchers to explore the corpora and provide quantitative analysis to offer empirical proof related to the sequences under examination (Biber and Conrad, 2001). There are a number of leading studies that have employed a quantitative, corpus-based approach as a method of investigating lexical bundles (see Biber *et* *al.*, 1999; Biber *et al.*, 2004; Cortes, 2004; Csomay, 2012; Hyland, 2008b). Biber *et al.* (2004) adopted a frequency-driven method to investigate lexical bundles in classroom teaching, textbooks and conversation. Biber and his colleagues found that there were 43 lexical bundles in conversation, 84 in classroom teaching, and 19 in academic prose. The results show that overall the two spoken registers employ a larger number of different bundles than the written registers.

2.4.6 Corpus-based research studying sequences and bundles

Corpus-based research does not only rely on reporting frequency counts of linguistic data, but also helps researchers to provide qualitative interpretations of quantitative data. Biber *et al.* (1998) view corpus-based research, besides reporting numerical findings, as helping to reveal patterns of language use through the analysis of these results, using concordance lines. In addition, corpus-based research has opened up new opportunities in a wide range of research and presented important insights across different fields (e.g. EAP, translation, academic discourse, lexicography, language variation studies, psycholinguistics, etc.). However, due to limited space and for the sake of focusing on relevant issues, I shall limit my discussion to a brief examination of the impact of corpus-based research on the investigation of the formulaic sequences that are particularly relevant to the present study: those occurring in academic discourse (Biber *et al.*, 2004; Hyland 2008a, 2008b) and EAP materials (Chen, 2010; Koprowski, 2005; Wood, 2010; Wood and Appel, 2014).

2.4.6.1 Sequences and bundles in academic discourse

Formulaic sequences such as collocations and lexical bundles have proven to be "pervasive in academic language use (see section 2.4.4) and a main element of fluent linguistic production, distinguishing novice and skilled use in both spoken and written contexts" (Hyland, 2012: 153). Corpus-based research has investigated sequences in corpora of written and spoken language, affirming that they appear more frequently in academic than in non-academic registers.

Hyland (2008a: 44) examined the four-part British National Corpus Baby edition, representing, for example, academic writing, imaginative writing, newspaper text, and spontaneous speech. The findings show that "both academic writing and conversation draw on a much larger stock of prefabricated phrases than either news or fiction" (Hyland, 2008a: 44). For example, there are around 800 different four-word clusters in the conversation corpus and over 450 in the written corpus, appearing more than 10 times in one million words.

Likewise, Biber (2009) in his frequency-driven approach (mentioned previously) aimed at uncovering the most common multi-word combinations in both conversation and academic writing. He integrated a corpus-based approach to investigate the varied ways in which the patterns (sequences) are different in the two registers. As established in previous studies such as Biber *et al*. (2004), Biber (2009), unlike Hyland's (2008a) conclusion, found that conversation employs a larger stock of recurrent multi-word sequences than is the case in academic prose. From his corpus analysis and based on the frequency rate, Biber confirmed that 140 multi-word sequences in conversation occurred more than 10 times per million words, compared to 94 multi-word sequences in academic writing.

In addition, a series of corpus-based studies were carried out to investigate how lexical bundles behave across different academic registers in academic contexts (Biber, 2006, 2009; Biber and Barbieri, 2007; Biber *et al*., 1999; Biber *et al*., 2004; Conrad, 1996; Csomay, 2012, Hyland, 2002, 2008a, 2008b), confirming that lexical bundles have distinctive distribution patterns in the academic register. For example, Biber *et al*. (2004), as previously mentioned, investigated the use of lexical bundles in the context of university classroom teaching and textbooks, and compared these bundles from classroom teaching and textbooks to those found in conversation and academic prose. The authors examined texts from university classroom teaching and textbooks found in the TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL Corpus). The T2K-SWAL Corpus offers a range of spoken and written registers that university students encounter during university life in the US (Biber, 2006).

The two registers in the corpus include a wide range of spoken and written activities learners encounter such as classroom teaching, office hours, study groups, on-campus service encounters, textbooks, course packs, and other written materials – university catalogues, brochures – as reported by Biber *et al*. (2004). However, Biber *et al*. (2004) only analysed two of these registers: classroom teaching and textbooks. The texts in the T2K-SWAL Corpus are taken from six major academic disciplines (business, education, engineering, humanities, natural science, and social science) and from three levels of education (lower-division undergraduate, upper-division undergraduate, and graduate). The conversation and academic prose texts were compiled from the Longman Spoken and Written English Corpus (see Biber *et al.*, 1999). The study showed that in classroom teaching (as a spoken register), the use of lexical bundles was surprisingly different from that of the other two registers in relation to frequency and functions. Classroom teaching uses more lexical bundles that convey stance and discourse organizers than conversation does, but uses more bundles with referential expressions than academic prose.

The present study shares with these previous studies a similar perspective of adopting a corpus-based approach to examine lexical bundles, which provides descriptive facts that entail clarification or explanation. However, the present study has a different goal from the studies mentioned, which is to focus only on the written register instead of reporting on both spoken and written registers. While Biber *et al.*'s (2004) goal was to investigate the use of multi-word sequences in the context of university classroom teaching and textbooks compared to conversation and academic prose, the present study will examine bundles derived from a corpus of two different academic genres: textbooks and instructors' handouts (materials). It will compare the bundles found in the different academic genres to the AFL (Simpson-Vlach and Ellis, 2010).

Many studies have been conducted in regard to teaching lexical bundles to learners from different levels, disciplines, and as part of their writing classes (Cortes, 2006; Kazemi *et al.*, 2014; Jones and Haywood, 2004; Schmitt *et al.,* 2004). For example, Cortes (2006) taught a set of lexical bundles to a group of university students in a writing history class. Before and after instruction (pre- and post-instruction) analyses were employed on students of different levels in order to examine their use of the taught bundles. The study suggested that there were no marked differences between the pre- and post- instruction with regard to the production of lexical bundles, but learners' awareness and interest increased.

In terms of identifying lexical bundles across academic disciplines and investigating university students and professional native speaker writers, Cortes (2004) compared the use of lexical bundles by published authors in history and biology and the use of lexical bundles by university students at three different levels of the same discipline. The results showed that students rarely used lexical bundles in their writing. In addition, the bundles employed by university students did not correspond to the lexical bundles used by professional authors. To illustrate, students studying history, who were from different academic levels, occasionally used different lexical bundles to convey particular functions that differ from those found in published history papers. For example, according to Cortes (2004: 413), some students used the bundle *at the same time*. This bundle was employed “for addition instead of simultaneity”. In the following example of an undergraduate lower-level student, Cortes observed that the student used the bundle *at the same time* metaphorically (Cortes, 2004: 413). The following example was extracted from the students’ corpus for analysis:

> He was tactful in his letter, being firm, but ***at the same time*** polite.

Cortes argues that the student used the bundle in a more creative way, which is conveyed differently than what would naturally be found in academic writing. Furthermore, students did not use many of the lexical bundles that were frequently found in the history and biology journals. Cortes concludes that there is a gap between students’ writing and published writing in the use of lexical bundles, supporting findings from previous studies such as Hyland (2008a). Hyland’s results reveal that there are considerable variations in frequency, structure and function across student and professional writing.

In addition, in a series of studies on learners’ corpora (Ädel and Erman, 2012; Bychkovska and Lee, 2017; Chen and Baker, 2010, 2016; Karabacak and Qin, 2013; Levy, 2008; Wei and Lei, 2011) comparing lexical bundles in academic native and non-native writing and professional writing, it was found that both types of writers employ different sets of lexical bundles and to different degrees. For example, Chen and Baker (2010) investigated the use of lexical bundles in three groups by comparing one corpus of published academic texts and two corpora of native and non-native English learners’ academic writing. The results revealed a gap between native speaker professional academic writing and university student writing (both native and non-native speakers), which seems to be an indication of immature writing. University students tend to frequently use verb-based bundles and discourse organizer bundles, which are employed to structure texts (e.g. *this essay is going to*). The native expert writers, on the other hand, were found to use referential bundles, which are used to determine attributes of many kinds such as:

> Framing: *in the context of*, *the nature of the*
> Quantifying: *a wide range of*, *in a number of*
> Place/time/text-deictic: *are shown in fig*, *at the same time*

Chen and Baker (2010: 34) also suggest that this gap or these variations may be due to genre differences between "published academic essays and university assignments" but also it could be more related to writing proficiency. This line of research on lexical bundles in students' academic writing shows a gap between both university native writers (Cortes, 2004) and non-native writers and professional writing (Ädel and Erman, 2012; Chen and Baker, 2010; Karabacak and Qin, 2013; Wei and Lei, 2011).

More importantly, Chen and Baker (2010) demonstrate that this gap could be attributed to a lack of introduction to such sequences in EAP textbooks. They argue that with the development of corpus techniques, interest in using corpus tools in identifying sequences, and the wide recognition of the importance of using sequences as building blocks in constructing discourse, ELT publishers and teachers do not appear motivated to focus on computer-retrieved sequences or bundles in their materials. This conclusion is in line with Burton (2012) and Harwood's (2002) assumptions on textbooks; Burton (2012) conducted a questionnaire study on current coursebook authors, suggesting that many textbook authors do not consult available corpora. Chen and Baker (2010) also add that formulaic sequences employed in native expert writing can be very helpful to learner writers in order for them to reach a native-like academic writing style. Thus, these bundles or sequences should be integrated into ESL curricula.

To add to research on academic discourse, Hyland (2008a, 2008b) examined three electronic corpora of written texts, which included research articles, PhD dissertations and MA/MSc theses. The corpora were compiled from four disciplines: electrical engineering, business studies, applied linguistics, and microbiology. The research article corpus was composed of 120 published papers, selected from leading journals suggested by experts, and formed a total of 730,000 words. The PhD corpus had 20 texts in each discipline and contained 1.9 million words, while the master's corpus also covered 20 texts in each discipline but with a total of 825,000 words. The PhD and master's students were mainly Cantonese-speaking first-language learners studying at five Hong Kong universities.

In academic writing, Hyland's corpus provided 130 different four-word bundles, including *on the other hand* as the most frequent bundle, occurring 100 times per million words. In addition, Hyland noted that the top 10 bundles were dominated by prepositional and noun phrase structures with *of* fragments. In terms of research articles, Hyland reported that it had 71 different bundles; while the PhD and master's theses had 95 and 149 different bundles, respectively. Hyland also found that many of the bundles used by postgraduate students were not found in research articles or occurred less frequently due to topic specificity (e.g. students wrote texts on topics related to Hong Kong unlike the other corpora). Not only did postgraduate writers use a greater variety of bundles, but they also used them at great frequency. Hyland (2008b) also reports that there are noteworthy disciplinary differences. For example, in terms of the number of bundles, electrical engineering (with 213 four-word bundles) had a wider range of bundles than biology (with 131 different bundles). Hyland claims that the reasons are unclear for these marked differences, but, he interprets it as relating to the distinctive styles that biology writers adopt to argue problems.

2.4.6.2 Sequences and bundles in EAP

Researchers and language instructors are realizing the importance of formulaic sequences and lexical bundles for EAP learners (Coxhead, 2008). EAP learners wanting to continue studying in English-speaking countries are surrounded by an overwhelming amount of academic language, either during EAP courses or in university settings. To adapt to the requirements of the academic environment and to succeed in their studies, learners need to be equipped with certain academic language in the form of vocabulary, formulaic sequences, or lexical bundles. Many researchers have also argued that academic sequences/bundles are important to writers. As stated repeatedly in this thesis, Biber *et al.* (2004: 371) view lexical bundles as "basic building blocks of discourse". Sequences or bundles are considered as defining markers of fluent language use of a particular register such as academic writing (Cortes, 2004). Hyland (2012: 153) notes that bundles help "to facilitate pragmatically efficient communication"; and they "function to structure discourse by guiding readers through a text (*in the next section*) or by linking ideas (*is due to the use*)". In addition, Coxhead and Byrd (2007: 134–135) state that these frequencies/bundles are important for academic writing instructors and for learners such as second-language learners (L2) for at least three reasons:

1. "the [lexical bundles] are often repeated and become a part of the structural material used by advanced writers, making the students' task easier because they work with ready-made sets of words rather than having to create each sentence word by word;
2. as a result of their frequent use, such [lexical bundles] become defining

markers of fluent writing and are important for the development of writing that fits the expectations of readers in academia;

3. these [lexical bundles] often lie at the boundary between grammar and vocabulary; they are the lexico-grammatical underpinnings of a language so often revealed in corpus studies but much harder to see through analysis of individual texts or from a linguistic point of view that does not study language-in-use".

In the field of EAP, some scholars have investigated what formulaic sequences to teach by using corpus-based research in developing teaching materials. According to Harwood (2005, 2010, 2014), corpus consultation can indeed help us in developing materials for EAP (e.g. Martinez, 2013; O'Keeffe *et al.*, 2007; Swales and Feak, 2000; Swales, 2002; Tomlinson, 2011). One way to explore formulaic sequences in EAP research is to look at recent research into EAP vocabulary at the word level. A considerable number of attempts have been made to compile lists of key academic terms to guide course material writers and teachers and to assist students in planning their learning more effectively. Computers and electronic corpora have revolutionized the making of new dictionaries by providing new ways to handle and analyse lexicographic data and by introducing new sources of linguistic evidence in EAP.

A recent and significant corpus-based project is Coxhead's (2000) Academic Word List (AWL), based on a 3.4-million-word corpus of academic writing from a wide range of disciplines. It is considered as a useful resource in EAP development research, and had a great impact on EAP teaching because it managed to collect high-frequency words from academic written discourse. Because EAP requires its own set of lexis and patterns and since academic discourse is marked by formal language not frequently used in fiction or newspapers (Coxhead, 2000), the creation of the AWL, as a teaching resource, filled a significant gap in language education. It offers a corpus-based list of lexical items directed for academic purposes that both teachers and learners can benefit from (Coxhead and Nation, 2001).

Although the AWL is very useful for EAP learners, at the same time, it can be overwhelming for learners to fully understand and use when studying individual words separately or when they are taught words excluded from their context and the patterning in which these words occur. As for teachers, these individual words are considered time-consuming to fully cover and explain in classrooms. These were some of the reasons why researchers were interested in looking for new ways to document key academic terms, which resulted in innovations such as forming lists of formulaic sequences. These lists looked beyond the word level and focused on multi-word sequences.

The AWL (Coxhead, 2000) influenced and inspired other researchers to develop lists of multi-word sequences. For example, Liu (2012), in his corpus-based study, identified and examined the most frequently used multi-word constructions (MWCs) of different types of combinations such as idioms, lexical bundles, and phrasal/prepositional verbs in general academic writing across the academic division and through using two mega sub-corpora from the Corpus of Contemporary American English (COCA) and the British National Corpus (BNC). The academic writing sub-corpus of COCA included 82.91 million words and the academic writing sub-corpus of BNC had a total of 15.33 million words. The study yielded a list of the 228 most frequently used academic written formulaic sequences of different types, including lexical bundes, idioms (e.g. *give rise to*), and phrasal/prepositional verbs (e.g. *deal with*).

In addition, Liu (2012) confirmed previous finding, such as those in Biber *et al.* (1999), that *noun and prepositional constructions* constitute the two largest types of sequences in academic written English. He introduced new findings, which may help learners in developing their academic writing. For example, the study noted that the *NP + linguistic action verbs (e.g. suggest) + that constructions* had high frequency in academic writing. He noted that the first three of the following six verbs constructions (**suggest** / **show** / **argue** / indicate / believe / claim) were among the most frequent constructions found at the top of his list. These constructions included forms such as *it has been suggested that* and *it has been shown that*. Liu's (2012) list of MWCs is based on general academic writing instead of discipline-specific writing, and is suitable for university academic writing classes as well as EAP classrooms.

Alongside the AWL, Byrd and Coxhead (2010) created what they claim to be a short but powerful list of 21 four-word bundles that occurred across four disciplines: arts, commence, law, and science. They suggest that it is challenging for teachers to take lexical bundle data into EAP classrooms. Similarly, Simpson-Vlach and Ellis (2010) created a list called the Academic Formula List (AFL), previously mentioned, for inclusion in EAP teaching. However, unlike Liu's

(2012) study which focused on identifying multi-word combinations in general academic writing across the academic divisions of the corpora and exploring their patterns, Simpson-Vlach and Ellis (2010: 490) derived a list that focused on one type of multi-word combination: formulaic sequences. Simpson-Vlach and Ellis (2010) included 2.1 million words each of academic speech and writing in their corpora of academic discourse. They note that their academic writing corpus comprised Hyland's (2004) 1.2-million-word research article corpus and selected British National Corpus (BNC) files taken from across academic disciplines (consisting of a total of 931,000 words), employing Lee's (2001) genre categories for the BNC. The written academic texts from the BNC included research articles and textbooks. The writing corpus was broken down into four sub-corpora that included the following academic disciplines: humanities and arts, social science, natural science/medicine, and technology and engineering.

The AFL was produced from an innovative combination of quantitative and qualitative criteria, corpus statistics and linguistic analyses, psycholinguistic processing metrics, and instructor insights (Simpson-Vlach and Ellis, 2010) (see Chapter 4, section 4.5.4.1 for a full explanation) to create a pedagogically useful list of formulaic sequences for academic speech and writing. The AFL provides manageable sets of formulaic sequences for use in classroom applications and teaching materials development. I was inspired to use the AFL in the present study as a measuring tool for investigating lexical bundles in EAP materials, focusing only on the written AFL (see Appendix B for the written AFL top 200 formulas for both written and spoken English). However, the present research will utilize the written AFL list and not the spoken one; I will focus on four-word bundles from this list (called the written AFL sub-list) (see Appendix C for the modified written AFL sub-list, including only four-word bundles). Chapter 4 describes the reasons for use, procedure, and modification made to the AFL and its sub-functional list and structural list (see Appendices D, E, F, and G).

An interesting study conducted by Jablonkai (2010) investigated the most frequent lexical bundles in English documents issued by the European Union (EU) for English for Specific Purposes (ESP) and pedagogic purposes. The aim of the study was to draw conclusions for language courses on English for EU purposes. The results revealed that there are similarities between the structural and functional classifications of EU bundles and the language of university textbooks and academic prose. In addition, the study suggests that EU texts are fairly formulaic and highlights the importance of the explicit teaching of lexical bundles in English for EU purposes courses.

From an EAP teaching perspective, with the aim to address issues such as "how learners should acquire formulaic sequences", Jones and Haywood (2004: 269) conducted an exploratory study to examine the teaching of formulaic sequences to a group of non-native EAP learners in a six-month intensive pre-sessional EAP course. Jones and Haywood decided to teach formulaic sequences that were presented according to Biber *et al.*'s (1999) research on lexical bundles, choosing and creating a useful list of approximately 80 lexical bundles and focusing on sequences related to academic reading and writing. The findings showed that most students gained awareness of formulaic sequences but few of them were able to use them appropriately in their essays due to the restricted timescale of the study (10 weeks only). The short duration of the course was considered a limitation in the observation and assessment of formulaic sequences in the students' writing.

2.4.6.3 Treatment of sequences and bundles in EAP textbooks

Developments in corpus-based research have allowed for lexical investigation regarding the treatment of formulaic sequences in the content of EAP textbooks. Studies and books on EAP textbooks and English-language teaching have been conducted to reveal the different types of formulaic sequences found in textbooks (e.g. lexical bundles, collocations; e.g. Shahrokhi and Moradmand, 2014), phrasal verbs (e.g. Zarifi and Mukundan, 2013), and multi-word constructions (e.g. Liu, 2012) or to report on the absence or presence of multi-word sequences or lexical bundles in textbooks, and to provide empirical interpretations of these findings (see Gouverneur, 2008; Harwood, 2014).

Regarding investigating sequences in textbooks, Jones and Haywood (2004), as previously mentioned, examined four widely known academic writing textbooks to investigate their use of formulaic sequences in the EAP context. Their study is considered to be among the few studies that have examined the development of EAP textbooks in relation to presenting and teaching lexical bundles. Based on their review of the EAP textbooks, they found that such textbooks select bundles for teaching by relying mainly on teacher intuition. However, it is important to note that this finding was not based on corpus-based interpretations. Instead, they analysed the content of these textbooks by focusing on the concept of formulaic sequences and how they are portrayed in these coursebooks, and by summarizing the phrasal language found. These textbooks included what they called "*Structural and Vocabulary Aid*" pages found at the end of each chapter (Jones and Haywood, 2004: 270). These pages included words, linking expressions, and academic phrases.

Unlike Jones and Haywood's (2004) research, the present study will conduct a frequency-driven approach to investigate lexical bundles, and will perform a structural and functional analysis of the lexical bundles to report on the treatment of bundles in EAP materials. Hyland (1994) examined epistemic modality in scientific discourse, focusing in particular on the treatment of hedging devices in a range of EAP and EST writing textbooks. The study suggested that textbook writers do not make use of empirical studies when handling hedging devices, and that textbook materials should rely on authentic data.

Koprowski (2005) examined the use of lexical bundles and other multi-word combinations in three contemporary ELT coursebooks by using range and frequency criteria in his corpus. Koprowski compared the sequences found to data in the COBUILD corpus. The findings revealed that of the 822 multi-word items gathered, only seven items were shared by any of the three coursebooks. Furthermore, it was found that not even a single combination was shared among the content of the three coursebooks although the ELT coursebooks were developed as general British English materials. Koprowski confirmed Jones and Haywood's (2004) findings, which concluded that the material developers do not follow any standard method when selecting their lexical items, arguing that ELT authors base their choices on personal experience and intuition.

Likewise, Wood (2010) investigated the most frequently used formulaic sequences which are typically found in the most popular textbooks in EAP teaching courses in North America. The aim of the study was to establish the kind of exposure EAP learners get in terms of the important features of academic writing and whether these textbooks use the formulaic sequences which are most relevant to academic reading and writing. The corpus was compiled from six EAP textbooks published by well-known publishers, covering intermediate to advanced-level learners. The EAP textbooks included two types of texts: reading and writing skill-focused texts and instructional texts (e.g. texts that included exercises and instructions for activities, and so forth). The study incorporated a frequency-driven method and the frequency cut-off was set to 20 instances per million words, and focused on four-word bundles. Wood (2010) concluded that the textbooks that the study examined were not effective in handling formulaic sequences and lexical bundles. It was noted that most of the lexical bundles that EAP learners would encounter would be from the instructional material found in textbooks, which are more related to the language used in the classroom but not the language found in academic prose.

Another key corpus-based study that was conducted by Chen (2010) examined lexical bundles in an electrical engineering textbooks corpus and compared them to those identified in English for Specific Purposes (ESP) textbooks aimed at teaching electrical engineering. A total of 105 four-word bundles were identified in the electrical engineering corpus. The identified bundles were classified functionally according to Biber *et al.*'s (2004) framework. Referential bundles constituted the largest proportion of bundles (78%), followed by stance bundles (19%), and discourse organizers (3%). The results, moreover, showed that the ESP materials were lacking a large amount of the functional sub-categories that were generated from the electrical engineering corpus. It was also found that the distributions of categories were different: referential bundles made up 88% of the total, stance bundles only 9%, and discourse organizers 3%. These results suggest that EAP materials fail to fully portray the language found in textbooks that aims to introduce students to their discipline.

To advance our understanding of formulaic sequences in EAP textbooks, Wood and Appel (2014) conducted a valuable corpus research study to investigate the most frequently used formulaic sequences, or what the researchers called multi-word constructions (MWC). These MWC were extracted from a corpus compiled of first-year textbooks, which represented the most popular academic disciplines: business and engineering, at a large Canadian university for EAP learners. The EAP corpus was constructed from five EAP textbooks aimed at intermediate and advanced learners. The EAP corpus was divided into two sub-corpora. The first sub-corpus contained texts that included the actual running texts such as articles and short readings found in EAP textbooks while the second sub-corpus contained "instructional language", which is the language usually found when introducing comprehension exercises or vocabulary, etc. According to Wood and Appel (2014), having two separate EAP sub-corpora is useful in order to better examine the pedagogical treatment of MWC in EAP textbooks. However, unlike Wood and Appel (2014), the present study examines the EAP textbooks as one corpus, combining reading and instructional information, to provide results from a different angle or approach as represented to EAP learners.

In Wood and Appel's (2014) study, the MWC retrieved from the EAP corpora were then compared to those in the first-year business and engineering textbooks corpus (FYBETC). The findings were analysed based on two levels of data: frequency and functional data. At the frequency level, with figures of occurrences ranging from 35% to 47%, it was reported by Wood and Appel (2014) that the majority of MWC found in the FYBETC did not appear at all in the EAP textbooks. This means that MWC, in numerous situations, occurred only one or two times in an EAP textbook, concluding that MWC were not being presented to learners in high-level EAP textbooks, relating well to the finding of Chen (2010) mentioned above. In other words, the MWC found in the FYBETC were underrepresented in intermediate and advanced EAP textbooks.

At the functional category level, the findings showed that a total of 44 MWC found in the business and engineering textbooks were not found in four out of five EAP textbooks. To illustrate, Wood and Appel's (2014) list revealed that these 44 MWCs included: 27 referential bundles (61%), 8 stance bundles (18%), and 9 discourse organizing bundles (21%). They concluded that from these percentages, it seems that the three functional categories were covered by the MWC and were not missing from any of the five EAP textbooks. Their final conclusion was that EAP materials lack the most common formulaic sequences that are found in first-year business and engineering textbooks. The analysis, in addition, indicated that although some formulaic sequences were found in EAP textbooks, none of the MWC were being introduced as teachable sequences, or presented to learners in any significant manner. The present study would add to the line of research on formulaic sequences such as lexical bundles by examining the structural category of the EAP textbooks and instructors' materials and then compare those found to the written AFL sub-list (see Chapter 8).

In sum, the literature outlined in this chapter provides an inclusive descriptive overview of formulaic sequences in general and lexical bundles in particular. It also demonstrates the important role of corpora in analysing academic discourse and EAP materials, and advocates that mastery of these bundles is vital for EAP learners. For example, Sánchez (2014) investigated lexical bundles in both biology textbooks and research articles. The study highlighted the pedagogical implications of using both biology textbooks and biology research articles to explicitly teach lexical bundles in courses aimed at biologists in a second-language context.

Furthermore, it attempts to provide a line of reasoning that lexical bundles are a prevalent and important feature of academic discourse, focusing on academic writing. Much research has been done in an attempt to uncover and explore the use of lexical bundles in, for example, university or EAP textbooks (Biber *et al*., 2004; Chen, 2010; Koprowski, 2005; Wood, 2010; Wood and Appel, 2014), and university and EAP teaching (Biber *et al*., 2004; Jones and Haywood, 2004), which has enhanced our understanding of the use of lexical bundles in course material design. However, there have been some limitations. Only investigating textbooks (e.g. Wood, 2010; Wood and Appel, 2014) can be far from straightforward. Although such studies relied on a frequency-driven approach, they are limited in that they only focused on examining textbooks rather than investigating different EAP materials (e.g. instructors' handouts) presented to learners in EAP settings, or relied on subjective analysis rather than corpus-based accounts in evaluating the textbooks, such as the study by Jones and Haywood (2004).

From the studies mentioned in the literature review above, I identified the need for research examining lexical bundles in EAP materials provided in an EAP pre-sessional programme particularly focusing on academic writing. Thus, the present study extends the research on lexical bundles by adopting a corpus-based approach and examining the use of lexical bundles not only in EAP textbooks, but also in instructors' handouts. Such studies on instructors' materials appear to be very minimal in the field of formulaicity.

Most previous research has focused on textbooks, classroom teaching and written work by learners and professionals. Moreover, most of the previous research has tended to concentrate on different sets of textbooks: academic writing textbooks or EAP textbooks that differ from those examined in the present study (see Chapter 3, section 3.2.4.1). It would be useful to examine whether EAP authors and instructors on a selected course select and use the most frequently occurring lexical bundles, as evidenced from previous corpus studies. One of the aims of the present study is to profile the lexical bundles that are most frequently presented to EAP learners. This was achieved by examining the EAP materials of two different genres to ascertain whether EAP textbook authors and instructors' materials are using corpus-informed materials such as the AFL. The comparisons shed light on the types, structures and functions of lexical bundles used in the EAP materials provided in such pre-sessional programmes. In addition, the study examined the location of the retrieved four-word bundles in the texts and checked whether the exercises or tasks covered teaching on lexical bundles. The results obtained will inform both EAP instructors and material designers, contributing to the pedagogical knowledge on lexical bundles in relation to EAP materials.

On a final note, it is noteworthy to mention that in the methodology phase of this research, important issues surfaced that needed to be closely profiled. Despite constraints related to time and the area of focus, I took the view that I should address these issues in the present study because they would serve to enhance the understanding of EAP materials. These important matters were subdivided into two separate questions, forming research questions 4 and 5.

The present study attempts to address the following research questions:

1. What are the most frequent four-word bundles profiled in textbooks and instructors' materials, in an EAP pre-sessional programme in the UK particularly aimed at teaching academic writing?

2. How are the most frequent four-word bundles classified functionally and structurally?

3. To what extent are the identified bundles in the EAP textbooks and instructors' materials based on data retrieved from corpus-driven lists such as the Academic Formulas List (AFL) (Simpson-Vlach and Ellis, 2010), in terms of the frequencies, structures and functions of the most frequent four-word lexical bundles?

4. Where in the EAP materials texts do the identified lexical bundles appear (for example, in readings and/or in instructions accompanying these readings)?

5. Do the EAP textbooks and instructors' materials present the identified four-word lexical bundles in tasks and exercises for EAP learners?

CHAPTER III

BACKGROUND AND CONTEXT OF THE STUDY

This chapter provides a full overview of the EAP pre-sessional course that has been selected in order to investigate its use of lexical bundles. The chapter will detail the decision-making process regarding the selection of the pre-sessional course for the investigation, including a number of reasons that influenced the eventual choice. A detailed description of the EAP pre-sessional course and the academic writing materials used will also be provided.

3.1 The rationale for selecting a leading pre-sessional course

The reasons for choosing this particular course are twofold: (1) to conduct an in-depth investigation on the use of lexical bundles in a leading pre-sessional course, by examining its published and unpublished materials; (2) to further our understanding of instructors' use of lexical bundles in the teaching of academic writing. A number of studies have investigated the effectiveness of EAP pre-sessional courses and how useful these courses are in helping learners develop the appropriate language and the academic study skills required to prepare them for future studies in higher education (e.g. McKee, 2012; Shaw and Liu, 1998; Storch and Tapper, 2009; Terraschke and Wahid, 2011). For example, Shaw and Liu (1998: 241, 245), in their study, found evidence of academic writing development by international students attending an EAP course. They reported that students' English writings became "less speech like" and shared more characteristics with those found within the standards of academic written English. They argued that some of the learners' language features that occurred frequently were considered "quite formulaic"; they attributed this change to the probability of learners' acquisition of lexical phrases. For example, there was an increase in the use of [*evaluation* bundles] or "certainty markers" such as the academic bundle *it is clear that* in students' writing, as fully explored in Chapter 9.

Similarly, Storch and Tapper (2009) evaluated L2 learners over time and found significant improvements in their use of academic vocabulary in academic writing, by measuring the vocabulary against the Academic Word List (AWL) (Coxhead, 2000). They partially attributed this lexical gain to the students' exposure to academic texts in their concurrent studies. This is particularly related to postgraduate students who are required to do extensive reading in their main discipline.

Overall, leading EAP pre-sessional courses seem to show promising results in developing students' general academic skills as well as their lexical repertoire for use in their academic writing. However, few studies have focused on which formulaic sequences, such as lexical bundles, are presented in the course materials (textbooks and instructors' handouts) of leading EAP pre-sessional programmes, and how they are presented. In addition, while the literature has highlighted the effectiveness of attending well-known EAP pre-sessional courses due to their usefulness (Storch and Tapper, 2009; Terraschke and Wahid, 2011), little research has compared the occurrence of lexical bundles in EAP materials with those found in corpus-driven lists such as the AFL. This comparison would provide additional value and useful information to EAP programmes regarding the use of lexical bundles. Additionally, concordance analyses would indicate the types, frequencies, structures and functions of lexical bundles.

To fulfil the dual aims of this research, a decision was made to select a well-known EAP pre-sessional programme to investigate its use of lexical bundles. It was essential to gather and design specialized corpora from EAP texts, which corresponded with the aims of this study and contained lexical bundles that students were exposed to in their academic writing.

3.2 The context: overview of the pre-sessional course

When selecting a pre-sessional course for this study, it was important to consider four interlinked factors: the course overview and aims; entry requirements; syllabus and teaching guidelines; and the EAP materials used. By considering these factors, it was possible to obtain a clear overview of the EAP course setting.

3.2.1 Course aims

Like many well-known English Language Teaching Centres, the Centre selected for the present study provides a wide range of English-language courses. The Centre states that its aim is to help international students improve their English-language proficiency and to develop their overall academic skills. It also claims that it prepares students for university study in a specific discipline such as business or engineering, and helps learners who want to develop their English for General Academic and Professional Purposes (EGAP). The present study focuses on a full-time intensive EAP pre-sessional course delivered at one of the UK's prestigious and leading universities. The main reason for selecting this course is due to its focus on EGAP, which fits with the aims of the present study. This course is provided by the university's English Language Centre. The pre-sessional course is an academically focused course designed to help learners achieve the required level of English-language proficiency for university entry, and to help learners who want to pass the IELTS test.

The pre-sessional course offers a foundation in academic English skills: academic listening, academic speaking, academic reading, and academic writing. The course is three terms' long, starting in September and ending in June. The course usually runs on a "semester" basis, from September to June (approximately 7 to 10 months in duration), and has different starting times. Students are able to join at the beginning of any of the three terms, subject to them meeting the university entry requirements and meeting the required English proficiency score (IELTS or equivalent).

3.2.2 Entry requirements for the pre-sessional course

In terms of English-language proficiency, not all international students attain the required English-language score to enter university; accordingly, the Centre may provide an opportunity for them to practise their English and hone their academic skills. The entry requirements for the pre-sessional course and the length of the programme are largely linked to the university's entry requirements (based on IELTS scores), and the entry requirements of each discipline. This means that students holding conditional offers for places on degree programmes need to study this course to achieve the required IELTS score and to gain entry to their desired discipline.

The pre-sessional course only accepts students onto the programme if they have an overall minimum score of 4.5 in the IELTS or equivalent (with nothing less than 4.5 in all components). Students starting on this overall band should be able to reach a band of 6.5 in one year, that is 30 weeks, with 10 weeks in summer school. Students starting on an overall band of less than 4.5 will need to attend the other

English programme, which is a different course as reported in the EAP course syllabus.

As previously mentioned, international students need to have a minimum IELTS score of 4.5 to be accepted onto the pre-sessional course. This course takes learners with different IELTS band scores above 4.5 but who still need to obtain a higher IELTS score for acceptance onto their course. Learners who are no more than two IELTS bands below the entry requirements of their course are required to take a 40-week course, while students who are no more than 1.5 IELTS bands below the entry level enrol in January on a 28-week course, as mentioned on the pre-sessional course website. For example, to study computer science, students need to reach an overall IELTS score of 6.5 (with a minimum score of 6.0 in each component). If they currently have an IELTS score of 4.5 (with a minimum score of 4.0 in each component), then they would need to enrol in September on a 40-week course. Before commencing the pre-sessional course, students are required to take a placement test to assign them to the most appropriate course to suit their needs.

Some students may start the programme in September, while a second group of students at an advanced language level may enrol in January. Throughout the three terms, most continuing students will have the same instructor and the same class. As reported by the EAP course, if individual students' proficiency scores and class assessments show signs of development and progression, then these students will be placed in an advanced-level class, following meetings between the tutor/student and tutor/Academic Director. The pre-sessional course syllabus states that classes are required to make progress each term and for the syllabus to be designed to provide sufficient and repeated practice in academic skills and language.

Newcomers joining the pre-sessional course need to take a placement test to assign them to a class that suits their English level. These students are required to master skills covered in Term 1 through consolidation and evaluation in Term 2. In Term 3 in April, there is no new intake of students and the course only includes continuing students. As stated above, to be accepted to study computer science, for example, or any other discipline, students must achieve the level of English proficiency required by their prospective department. Therefore, during or at the end of this course, students have the choice to either re-take the IELTS test to attain the required pass grade, or to obtain the pass grade by taking the University English Proficiency Test.

3.2.3 The pre-sessional course syllabus and academic teaching guidelines

The pre-sessional course syllabus claims that it aims to improve the reading, listening, speaking and writing skills of students. The EAP course syllabus states that special attention is paid to familiarizing students with the standard academic writing expected in the British academic environment, by instructing students on academic language and providing them with ample practice. The pre-sessional course provides a Core EAP component focusing on reading and writing skills to enhance students' academic writing and reading abilities; classes last for a duration of six hours per week.

Another course component in the pre-sessional course is Extended Writing, where students are expected to produce at least one piece of extended writing each term before undertaking a study project. For example, in Terms 1 or 2, students can submit an *Extended Essay*, on a topic selected by their tutors. They could write an *Extended Summary*, or could produce a *Synthesis* of two or more texts focused on a similar topic with contrasting views. Students also could present a *Research Report* based on data gathered by them in the form of questionnaires or interviews. The Extended Writing component is taught in classes lasting a total of one and a half hours per week. In addition, in Term 3, students submit a 2,000–3,000-word *Study Project*, where learners are required to research a topic related to their field of study and report on their findings, with guidance from their instructors.

According to the pre-sessional course syllabus, the proposed curriculum should serve as a guide and be regarded as a descriptive rather than a prescriptive document. Hence, instructors are not required to follow one rigid method but are expected to elaborate or alter the syllabus to suit the students' academic writing needs. Writing instructors can pick and choose techniques and activities that are most appropriate to enable students to become independent learners and to be able to function well in the writing aspect of their academic studies. The *Group Tutors* who teach the writing classes are expected to foster students' autonomy in academic settings by providing tasks and interaction that promote critical thinking. Writing instructors are advised to provide students with supplementary materials suggested by the course or with the teacher's own materials if they identify that learners need more practice in certain problematic areas. Additionally, as claimed by the EAP course syllabus, on a weekly basis, learners are provided with a guided study period so that they can discuss any aspect of their *Study Project* with their instructor.

Academic reading and writing are integrated skills. The EAP pre-sessional course, like many other courses, merges academic reading and writing activities in its classes. Although these two academic skills are closely intertwined, the teaching of these skills involves providing EAP learners with a different set of activities and tasks. As stated by Grabe and Zhang (2016), studies conducted on summarizing skills is a feature of abilities that suggest reading-writing relations. Thus, as I was interested in examining materials aimed only at teaching writing, it seemed logical to collect handouts, sheets and readings or any other material used during writing classes. When instructors assign readings during writing classes, their students may be aware that these readings are given to help them accomplish a certain written task related to academic writing (e.g. paraphrasing, summarizing, arguing). Texts given when teaching academic reading were more likely to equip EAP learners with skills devoted to academic reading (e.g. skimming, scanning, reading with a purpose, comprehension questions, reading quickly). Since the focus of this study is to examine materials aimed at teaching writing, my aim was to gather materials given to learners during composition classes (see Chapter 1 for the reason behind only examining writing materials).

3.2.4 EAP teaching writing materials

3.2.4.1 The textbooks

Textbooks used: The following textbooks are used in the pre-sessional course:

1. *Cambridge Academic English: An Integrated Skills Course for EAP*, Student's Book, Intermediate (B1+), by Craig Thaine.
2. *Cambridge Academic English: An Integrated Skills Course for EAP*, Student's Book, Upper Intermediate (B2), by Martin Hewings.
3. *Cambridge Academic English: An Integrated Skills Course for EAP*, Student's Book, Advanced (C1), by Martin Hewings and Craig Thaine.
4. An Extended Writing Booklet (EWB) for Terms 1, 2, and 3, by an anonymous writer(s), was also offered to learners. According to the pre-sessional course syllabus, the textbooks and EWB were used by learners across the three terms, as shown in Table 3.1. The course expects to have different learners with different levels throughout the year, so changes and amendments to this

suggested distribution were possible.

Table 3.1. The distribution of textbooks and EWB used across the three terms

Terms	Textbooks	Level
Term 1	Cambridge Academic English	Intermediate Upper Intermediate
Term 2		Intermediate Upper Intermediate Advanced
Term 3	Cambridge Academic English	Upper Intermediate Advanced
All Terms	Extended Writing Booklet (EWB)	All levels

Target learners and levels: The Cambridge Academic English (CAE) series and Extended Writing Booklet (EWB) are for learners who need to use English in their academic studies. The CAE series consists of integrated skills coursebooks, developing learners' abilities in reading, writing, listening, and speaking at different levels (Intermediate/Upper-Intermediate/Advanced) within an academic context. They provide students with different topics and texts that are suited to all disciplines or subject areas. Each textbook is targeted at a certain level in higher education. The following is a description of the different levels (B1, B2, and C1), according to the *Common European Framework of Reference* (CEFR), which is found in the introductions of all three student textbooks in the CAE series (2012: 5):

> Student's Book B1 is aimed at students who need to improve their English significantly in order to guarantee success in higher education. If you are familiar with the Common European Framework of Reference (CEFR) proficiency levels, Student's Book B1 is likely to be most useful for Independent Users at level B1 and above. Student's Book B2 is aimed at students who will soon be starting undergraduate or postgraduate studies and are Independent Users at level B2 and above. Student's Book C1 is aimed at students who may already have begun their academic studies. It will also be of interest to non-native English-speaking academics who need to present and publish in English. It will be of most use to Proficient Users at level C1 and above.

The EWB is also aimed at learners of Intermediate/Upper-Intermediate/Advanced levels, preparing them and formalizing relevant writing skills expected in the academic environment.

Textbook organization: Despite the integrated nature of the four academic skills covered by this course, I will attempt to highlight aspects only linked to the component of academic writing. The books are organized into ten units, centring around a broad topic of interest. At the end of each of the ten units there is a Grammar and Vocabulary section, which is very important for academic written and spoken communication. As previously mentioned, I only focused on the written communication parts within each unit. The cross-references in the margins of each unit refer to any additional information, strategies or extra practice that can be found in the Grammar and Vocabulary section. Due to the main focus of the study, this study only considers additional information, strategies or extra practice that focus on academic writing skills found in the textbooks. Thus, in line with my research questions (see Chapter 2, section 2.4.6.3 and Appendix A), the aim was to explore EAP materials designed to teach academic writing.

The EWB is structured into seven sections, targeting the study skills needed to meet the Extended Writing requirements (e.g. a long argument essay and problem-cause-evaluation essay). Each section focuses on developing three to four writing skills, providing many exercises and extra practice to help learners improve their academic writing, such as using source materials and paraphrasing.

Each textbook from the CAE series consists of 176 pages, while the EWB consists of 123 pages. The components of the textbooks, in terms of the number of texts and their word count, can be found in Table 4.2 in the methodology chapter.

Language taught: According to the CAE series, the language that is being taught comes from authentic academic texts. They claim to use texts found in the textbooks and journal articles that subject tutors might ask or recommend their learners to read. They state that the vocabulary is carefully selected due to its particular importance in academic writing or reading, and also include many words from the Academic Word List designed by Averil Coxhead (2000) that learners may encounter in their academic studies. They also claim that when preparing the textbook materials, they made use of the Cambridge Academic English Corpus (CEAC), providing up-to-date information and materials. In addition, the pre-sessional course developed the EWB, which they claim advocates the standards and conventions of academic writing, using academic language and materials.

Additional features in the CAE series: In each unit, learners encounter different feature boxes, each with unique aims. *Study tip boxes* provide suggestions to improve learners' study methods. *Information boxes* provide useful background information related to language or academic culture. There are also *Focus on your subject boxes*, which encourage learners to apply what they have learned to their own subject of interest. Moreover, *Corpus research boxes* present and report findings from the CEAC – it is a 400-million-word resource corpus in two parts, spoken and written. The written academic language is taken from textbooks and journals, covering both British and American English. Additionally, the *Word list* at the end of the textbooks lists key academic words.

3.2.4.2 Instructors' materials

Participants: In the pre-sessional course, five instructors of academic writing agreed to be involved in this study. As previously mentioned in section 3.2.3, tutors are advised to meet learners' needs, and are permitted to make modifications to the syllabus as required. Thus, in addition to the core materials, instructors were found to use different types of materials to help EAP learners advance their academic writing.

Participants' teaching background: I obtained background information regarding the instructors' educational and teaching experiences in order to gain insightful knowledge on instructors' understating of lexical bundles. In order to add depth to my data, I formulated a short questionnaire to obtain further evidence on the experience of the academic writing instructors. I was interested in finding out whether teachers are aware of lexical bundles, and how these teachers feel about using lexical bundles or any academic lexical lists during their writing classes. The data was collected online, and four out of the five teachers completed the questionnaire (Appendix L). Table 3.2 provides background information on each teacher, including their teaching experience, nationality, native language and qualifications.

Table 3.2. Instructors' background information

Teachers	Teaching experience	Nationality	Language	Qualifications
Teacher #1	Over 10 years	British	English	MSc. Teaching English for Specific Purposes
Teacher #2				MA in Linguistics and TESOL
Teacher #3				MA in TESOL
Teacher #4	7 years	Polish	Polish	MA Applied Linguistics
Teacher #5	Did not hand in the questionnaire – no information was provided			

Material types: Instructors' materials were presented in different formats; some teachers provided their learners with journal articles, ranging in size and topic. For example, *Plagiarism*, *Navigating Computer Files*, *History of Sheffield*, and *Adapting to Climate Change* are examples of topics offered by teachers. A second type of material provided to learners consisted of exercise sheets related to grammar and vocabulary. Other teachers also used handouts, involving tasks such as *Read and answer the following questions*. These reading tasks were included in the corpus because they provided practice related to academic writing skills such as summarizing, paraphrasing, and referencing. Instructors also created and handed out *short quizzes* on vocabulary and reading and writing in their writing classes. In the discussion chapters, further details will be provided in relation to specific instructors and the type of materials they used in their classes. The components of the instructors' materials, consisting of the number of texts for each teacher and word counts, can be found in Table 4.3 in the methodology chapter.

3.3 Concluding remarks

It is important to note that the information included in this chapter was obtained and accessed via four means: the Centre's syllabus, *Guidelines and Suggested Materials for Tutors*; the Centre's website; the Academic Cambridge English textbook series; and via materials directly handed in by instructors who agreed to participate in this study

In this chapter I have provided an overview of the EAP pre-sessional course under investigation and its relevant materials. I will now turn to the design of the corpora itself, beginning with detailing the methodology and procedure in Chapter 4.

CHAPTER IV

METHODOLOGY, CORPORA, AND PROCEDURES

The present study adopts a corpus-based approach (see Chapter 2, section 2.4.6) to retrieve and closely examine a set of lexical bundles. The study uses a frequency-driven approach (see Chapter 2, section 2.2.2.1) to identify and highlight the most frequently found four-word bundles. The identified bundles are taken from a collection of texts compiled from teaching materials (textbooks and instructors' materials) used in an EAP pre-sessional course at an English Centre at one of the UK's prestigious and leading universities (see Chapter 3).

Section 4.1 describes the data collection steps. Sections 4.2, 4.3 and 4.4 detail the three main methodological phases in retrieving the lexical bundles from the EAP materials. Phase 1 involved the process of building the two EAP material corpora. Phase 2 involved setting the bundle identification criteria, while Phase 3 focused on the exclusion specifications applied to the bundles found in this study. The retrieved bundles were created to answer the first, second and third research questions (see Chapter 2, section 2.4.6.3 for the three research questions and Appendix A).

Section 4.5 presents the procedure adopted to carry out the analysis. Sections 4.5.1, 4.5.2 and 4.5.3 explain how the frequency, structural, and functional analysis were conducted, respectively. The purpose of identifying and classifying bundles according to their functions and structures was twofold: (1) to profile the most frequent types, structures and functions of bundles, gaining an understanding of the preferred bundle use by material designers; (2) to ascertain whether or not EAP materials are written on the basis of the AFL. This part of the study answers research questions 2. For comparison purposes, section 4.5.4 presents the rationale behind using the AFL and shows the procedure that was carried out to modify the AFL and its corresponding functional and structural list to suit the present study, answering research question 3.

Two additional questions (see Appendix A for the research questions) surfaced during the concordance line checks, which were adapted and integrated into the present study to provide helpful information to assist me in drawing final conclusions on the EAP materials. Hence, sections 4.5.5 and 4.5.6, respectively, detail the procedures adopted to facilitate the analysis of the identified bundles in terms of in-context information (bundles found in readings and/or in the instructions accompanying these readings) and for pedagogical analysis (were the identified bundles presented to learners pedagogically?). This section is important to address the fourth and fifth research questions.

4.1 Data collection process

Having obtained research ethics approval from the School of English to conduct my research, I contacted the head of the EAP programme at the English Centre. Through a series of emails, I requested access to the relevant EAP materials. The head of the EAP programme provided me with the guidelines for researchers who are interested in conducting research at the Centre, which provides information on gathering data and contacting and approaching staff members and teachers.

Upon the Centre's request, I provided a Microsoft Word document which included a description of my research, listing a set of points requested by the English Centre (see Appendices H and I for the information sheets that I provided). The information sheet included information on: the purpose of the study, the type of participants (e.g. staff and instructors), the agreement and consent forms (Appendix J), benefits arising from this study, information on confidentiality, and my contact details. It is important to note that the EAP textbooks and instructors' handouts/materials included in this study can only be used for the purposes of this research and will not be available to other researchers.

The head coordinator provided me with the EAP textbooks and relevant materials (in the form of hard copies) from the EAP course selected. However, due to the instructors' busy schedules, it took five to six months to collect all relevant materials. Instructors' materials were gathered and received electronically (e-documents) through emails directly sent to me by all five teachers who had agreed to participate in my study and who were teaching writing on the EAP course. Instructors were requested to only send in all materials that were targeted at teaching academic writing and used during the writing class (see Chapter 2 for reasons why the focus was on academic writing).

4.2 Phase 1: the corpora

As one of the main purposes of this research study is to identify the most frequent four-word lexical bundles in EAP teaching materials, corpora of EAP texts were needed. The first step involved building two corpora to generate the EAP lexical bundles: one from the textbooks and the other from the instructors' materials. The bundles retrieved from the textbooks corpus and the instructors' materials will be listed and called the *textbooks bundles list* and the *instructors' materials bundles list*, respectively.

4.2.1 The textbook texts: the design and construction process

A total of three EAP textbooks and a writing booklet were received and used to create the textbooks corpus of this study. The textbooks were aimed at intermediate / upper intermediate / advanced learners of English. The context and nature of these books are discussed in Chapter 3, section 3.2.4.1. The EAP materials consisted of:

- *Extended Writing Booklet (EWB);*
- *Cambridge Academic English – Intermediate (CAEI);*
- *Cambridge Academic English – Upper Intermediate (CAEUI);*
- *Cambridge Academic English – Advanced (CAEA).*

In order to convert the above-mentioned textbooks (hard-copy versions) into digital format (soft-copy versions), I first needed to scan them. Advances in the printing market and the easy-to-use interface of office devices has made it possible for many researchers and linguists to easily operate scanners. In the present study, an HP printer (model: Deskjet Ink 4625 Advantage) was used. Additionally, Optical Character Recognition (OCR) technology was required to complete the scanning tasks and enabled me to convert different types of documents into editable and searchable data. I used Readiris Pro 15, which is an OCR program that is included as part of the HP printer's full-feature driver. This software was chosen because it provides advanced OCR features, which facilitated the scanning process. The freeware program Readiris Pro 15 was downloaded along with the full-feature driver from the following link: https://support.hp.com/emea_middle_east-en/drivers/selfservice/hp-deskjet-ink-advantage-4620-e-all-in-one-printer-series/5157451/model/5157452.

For this research, the aim was to use OCR to convert the textbook pages into enhanced digitalized document files in plain text (*.txt*) format and save them in folders. Editable *.txt* format was selected because the AntConc tool only works with plain text format and not with PDF or Word files. As a result, four digitalized and converted textbooks in *.txt* format files were stored and saved on my personal computer in the *Textbook .txt* folder; this folder contained the following files:

- *EWB.txt file*
- *CAEI.txt file*
- *CAEUI.txt file*
- *CAEA.txt file*

4.2.1.1 Manual checking and clean-up process

The data in the four saved *.txt* files were checked and cleaned in two distinct stages, using two different approaches. At this stage of the manual checking process, manual inspection was a significant requirement when building the textbooks corpora in order to trace, identify, and delete or change errors that occurred during the scanning process, establishing clear and reliable data for analysis.

Stage 1: manual checking

Along with all of the inherent advantages that scanning technologies offer, they also bring in inevitable and redundant errors. Any process that involves electronic conversion may require some level of manual checking and editing to ensure that the original hard-copy document is best represented. To ascertain whether the information scanned from the textbooks had been accurately transferred to all *.txt* files, I conducted a line-by-line manual inspection. This first stage involved manually checking the *.txt* files "*longitudinally"*, that is, going through the files word-by-word, "sentence-by-sentence, and text-by-text" from beginning to end (Nelson *et al.*, 2002: 17).

I applied line-by-line amendments to the *.txt* files as they emerged, checking and correcting any errors throughout the four *.txt* files, one file at a time. To achieve as accurate a representation of the textbooks as possible, I tried to maintain a systematic data-checking technique to guarantee a high degree of consistency.

Although running a manual check on the *.txt* files was of major importance to ensure reliability, the level of scrutiny required was also time-consuming and effort-intensive.

The *.txt* files consisted of two types of content: *lexical texts* in the form of words, phrases, sentences, paragraphs, exercises, titles, headings, and readings; and *enclosed texts* in the form of words, phrases or numbers located within tables, pictures, charts, figures, diagrams, illustrations, drawings, etc. Regarding the first type, it was evident that large portions of the content of these texts had been properly scanned. However, minor but frequent errors appeared regularly throughout the files. For example, in some instances certain letters were depicted as numbers (such as the *letter l* being scanned as the number *1*). Also, certain words or expressions had not been scanned properly, such as those found in handwritten form (e.g. time /on the other hand) or crossed-out words (e.g. ~~organize / on the other hand~~). In these cases, if such an error was detected, I had to manually fix the mistake in the *.txt* file by deleting and/or re-keying the correct item, trying as far as possible to maintain an error-free representation.

The second type of data, *enclosed texts*, were very messy and the most problematic to process. Scanning problems meant that manual checking and re-keying were again required to ensure reliability. Typically, all tables, charts, pictures and figures were not transferred to the *.txt* file format, but their enclosed or surrounding lexical and numerical data were. However, such lexical items were commonly misplaced, jumbled or incomplete, providing scrambled and vague information. In corpus linguistics, this kind of data is referred to in the literature as *messy data* (Mair, 2006), meaning that it is difficult to obtain clear and interpretable information from them due to their apparent ambiguous nature. Thus, this checking process required longer and special manual handling, making it impractical and meaningless to fix and re-key every error. Thus, at this point, to overcome such problems and ensure greater consistency, a decision was made to see which items needed to be retained and which needed to be discarded, depending on the type of data inside the tables and so forth. In this regard and to facilitate the data analysis, the decision to clean up and delete insignificant and irrelevant information within this data type was a functional strategy based on logical judgement.

My decision to ignore this type of data was based on two important factors. First, all numerical data were scanned in a disorganized manner, which needed deleting and re-entering for accuracy purposes, requiring more accurate keying-in. This process was time-consuming and demanding, and the fact that I was under time constraints meant that keying in unrelated numerical information was not the best use of my time. The second factor was related to the goal of the study, which requires the investigation of lexical bundles. Thus, excluding numerical data from the tables in my corpus will not affect the results and the aims of the study. Removing numerical information may have an impact on the total word count but if a study does not have a predetermined target for the overall corpus size (Nelson, 2010), then it is common practice to focus on the more relevant data (see e.g. Cortes, 2004). Thus, I made sure that the omitted data did not significantly affect the results of this study.

If the tables or pictures displayed only numerical information, showing numbers, statistics, years, ages, percentages, calculations, etc., I decided to delete this kind of data. Table 4.1 presents an excerpt taken from the *Cambridge Academic English* (*Upper Intermediate)* textbook, displaying percentages such as 39%, 15%, 13%, etc. that were scanned in a disorderly manner. I omitted all numbers from the *.txt* file and retained the surrounding words or phrases, such as *me*, *women*, *age group*, *retirement* and *proportion not working*.

Table 4.1. Excerpt taken from the *Cambridge Academic English: An Integrated Skills Course for EAP (Upper Intermediate)* textbook, Chapter 8 [b], p. 110

Age group	Proportion not working	
	Men	*Women*
18–24	39%	42%
24–34	15%	29%
35–49	13%	24%
50 to retirement	28%	30%

In the tables or figures where the data contained both lexical and numerical information, the data were checked, retyped or fixed to best resemble the original data. In charts or tables, if such data were misplaced or disorganized due to scanning error, I manually removed the whole chunk of the messy data and re-entered the numerical data with the lexical items directly into the *.txt* file, checking line-by-line for transparency and consistency. For example, for misplaced numerical data inside tables or charts such as the expression *in the year 2000* which were scanned incorrectly, e.g. *2000 in the year*, I manually fixed the data as the lexical items are considered relevant to the study. Finally, pictures without text and pictures and tables with very small font were manually removed for the obvious reason that they were too small to read.

In this stage, a second clean-up process was crucial in order to closely examine the lexical bundles in materials aimed at a specific register: academic writing. The term “cleaned” was used by Cortes (2004: 403) and others to refer to the act of leaving out certain items and pages or other information that are considered irrelevant to the study and which do not contribute towards the aim and goals of the research. Therefore, the clean-up process in the second phase consisted of two complementary objectives: (1) focusing on material targeted at teaching academic writing, including reading materials; (2) excluding irrelevant materials directed at teaching speaking and listening.

To achieve these objectives, a further clean-up process involved going over all of the data and eliminating information targeted at teaching listening and speaking, but keeping all of the data focusing on teaching academic writing. This is because, as stated by Biber *et al.* (2004) and Biber (2009: 275), “it turns out that multi-word patterns/lexical bundles typical of speech are fundamentally different from those typical of academic writing”. Speech has a different set of lexical bundles to those found in academic writing (see Chapter 2). From these studies and similar ones, it has been established that the lexical bundles used in academic writing are generally distinct from those found in speech.

For the purposes of transparency, consistency and representativeness, it is crucial to clearly explain these issues within the design and construction phase. At the heart of modern corpus linguistics lies the issue of data transparency, which involves reporting on what type of data is included in the corpora being studied. In this regard, and in addition to the above-mentioned manual checking and clean-up process, I also removed all exercises or pages that included lecture skills and audio scripts. In addition, I excluded cover pages, author pages, copyright information pages, acknowledgements, introductions, table of content pages and end matter pages from inclusion in the corpora. As has been mentioned several times, my aim was to maintain a close focus on materials targeted at teaching writing, providing targeted data for analysis. In the context of the present study, consistency requires the instructors’ materials corpus and textbooks corpus, and the AFL written sub-list, to be compatible in terms of sampling features, with a focus on four-word bundles targeted towards academic writing (see section 4.3).

Regarding representativeness (see Chapter 2, section 2.4.2 and Chapter 11, section

11.1.3), the present study can be useful for exploring lexical bundles found in the teaching of writing materials within the specified textbooks used by instructors within a specified course, "thus reinforcing the interrelationship of research question, representativeness, corpus design and size" (Reppen, 2010: 32). However, for studies seeking to generalise the findings on lexical bundles found within EAP materials, EAP courses or academic writing, corpora with millions of words, a wider scope of EAP materials, and a selection of EAP courses are needed to ensure that reliable conclusions are obtained.

4.2.1.2 The textbooks corpus

At this point, I incorporated the four checked and cleaned *.txt* files into one folder called the *textbooks corpus*. These texts were ready to be imported into the AntConc program to generate the initial list of lexical bundles. The total number of the four *.txt* files are as follows: the *EWB* consisted of *34,649* words; the *CAEI* totalled *59,671* words; the CAEUI totalled *56,240* words; while the *CAEA* totalled *61,675* words. Table 4.2 below provides information on the main components of the textbooks corpus ready for investigation and its finalized word count (212,235 words) after the checking and clean-up processes were completed.

Table 4.2. Constituents of the textbooks corpus

Representation	Word count	Number of texts
AEPC Extended Writing	*34,649*	*1*
Cambridge Academic English: (Intermediate)	*59,671*	*1*
Cambridge Academic English: (Upper Intermediate)	*56,240*	*1*
Cambridge Academic English: (Advanced)	*61,675*	*1*
Totals	*212,235 words*	*4 texts*

4.2.2 The instructors' texts: the design and construction process

A series of steps were followed to create the instructors' corpus, each containing a sequence of sub-steps, from the process of conversion to plain text files to the manual clean-up process. The second corpus included instructors' materials (e.g. handouts such as exercise sheets, readings, etc.) (see Chapter 3, section 3.2.4.2 for information on the instructors' materials).

As mentioned in the data collection section (see section 4.1), each instructor emailed his/her materials that were used during the writing classes. Similar to the textbook pages, these electronic documents (e-documents) needed to be converted to plain text files to be handled successfully by the AntConc program. As with the textbooks, the Readiris Pro 15 program was used to convert the e-documents to text files.

4.2.2.1 Manual checking and clean-up process

Stage 1: manual checking

The manual checking process for the instructors' materials was slightly different from the process used for the textbook materials. Although most of the manual checking process was done at the same time as the conversion, an additional clean-up process was needed to guarantee transparency and accuracy. As part of the corpora design, an OCR error post-correction was required in order to guarantee consistency and to obtain error-free lexical bundles. Thus, an additional manual checking process was conducted, which required line-by-line inspection of the *.txt* files to clean up any messy data that had been overlooked during the previous step. Similar to the clean-up phase in the textbooks corpus, words or items that had been incorrectly recognized were removed from or changed in the *.txt* files during the conversion process (see section 4.2.1.1 for information on the manual checking process).

In addition, unlike with the textbook materials, the second stage of the clean-up process was not performed on this set of data. This is because the materials provided by the teachers were aimed at teaching academic writing. Therefore, they did not require further cleaning up such as removing speaking activities or audio scripts.

4.2.2.2 The instructors' materials corpus

The instructors' materials corpus consisted of five *.txt* files; each of these five *.txt* files represented a teacher's entire materials, which ranged in document number and size. The number of e-documents sent by the five teachers ranged from 1 to 29 e-documents and fluctuated between 611 and 45,925 words. For example, one teacher provided 29 e-documents, totalling 18,278 words. A detailed description of each instructor's materials and associated word counts are provided in the appendices (see Appendix K). Table 4.3 below provides information on the main components of the instructors' materials corpus and the number of texts; its finalized word count after clean-up was 86,693 running words. Each teacher's chunk of materials was considered as a separate whole text and was technically treated as one complete text since he/she was regarded as the author or owner of the material. For this reason, each instructor's material is grouped and presented as a complete work provided under a single long text. The instructors' materials corpus was completed and prepared for transfer to AntConc to generate the second lexical bundle list.

Table 4.3. Constituents of the instructors' materials corpus

Representation	Total number of texts provided	Word count
Teacher # 1	6	45,925
Teacher # 2	1	12,654
Teacher # 3	8	9225
Teacher # 4	1	611
Teacher # 5	29	18,278
Total	45	86,693

4.3 Phase 2: generating and identifying the lexical bundle lists

4.3.1 Software tool: AntConc

The aims of this study played a key role in deciding on the most appropriate program for generating the lexical bundle lists. I found AntConc to be well suited for helping me find specific answers to my primary research questions since the main focus of this study is to take a frequency-driven approach. AntConc (Anthony, 2018b) was developed by the corpus linguist and professor Laurence Anthony. AntConc is considered to be among the three most well-known software packages for users working in corpus analysis (Römer and Wulff, 2010; Laviosa *et al.*, 2017). In addition to the advantage of this program being freeware, there is a user support file introducing this multiplatform tool which can be used to carry out lexical investigations in corpus linguistics research (Anthony, 2018a). It is reliable in producing fast-frequency outputs and concordance lines when investigating lexical bundles in texts.

In the present study, each corpus was processed independently to generate the lexical bundle lists. AntConc was able to provide ready-made lists from my corpora on the basis of automatic calculations. In addition, the concordance lines provided me with different, easy and fast ways to access the texts. Without AntConc, or corpus tools, my corpus may be of no use. However, AntConc has weaknesses and limitations. One weakness that I encountered is the following: due to the fact that AntConc handles the corpora quickly and with a user-friendly interface as stated, it occasionally generates ambiguous sequences (e.g. sequences that consist of individual letters appearing alone within the sequence). For example, during the initial lists, one line displayed this unclear combination: "*chambers j k sociolinguistics*". This kind of combination was deemed as ambiguous and was not included in the present study. Due to this weakness, I had to go over my lists manually to exclude each ambiguous incident; this will be explained further in section 4.4. I found this drawback time-consuming and tedious, especially when working with long lists.

4.3.2 Lexical bundle identification

To begin the process of generating the lexical bundle lists from the textbook and instructors' materials, I began by applying traditional corpus-based identification techniques, setting important measures regarding lexical bundle list creation. I needed to make critical decisions in reference to three key criteria (see Chapter 2):

1. the length of bundle combinations (e.g. three-, four-, or five-word bundles)
2. range (number of texts in which each bundle occurs)
3. the frequency threshold (number of occurrences of bundles within a corpus)

4.3.2.1 Key criterion 1: setting the length of lexical bundles

The first decision involved setting the length of the lexical bundles, which determined the length of the word combination to be investigated (usually two-, three-, four-, five-, or six-word bundles). A decision was made to focus on four-word lexical bundles for a number of reasons. First, previous research studies on lexical bundles have regularly selected four-word bundles over two- or three-word sequences (e.g. Ädel and Erman, 2012; Biber *et al.,* 2004; Chen and Baker, 2010; Cortes, 2004, 2006; Hyland, 2008b). The reason behind this is that four-word bundles can hold two- and three-word bundles within their structures (as in *as a result of*, which contains *as a result*) (Cortes, 2004; Hyland, 2008a, 2008b; Wood and Appel, 2014). Second, four-word bundles are much more frequent than five- or six-word bundles, which are found to be rare sequences, (Cortes, 2004: 401; Hyland, 2012), thus providing a solid basis for analysis.

In addition, the present study aims to provide a full and detailed analysis of the lexical bundles found in an EAP pre-sessional course, but due to time limitations four-word bundles are a good starting point for manual categorization (structurally and functionally) and concordance checks (Chen and Baker, 2010). Finally, four-word bundles were selected because they "offer a wider variety of structures and functions to analyse" (Cortes, 2004; Hyland, 2012: 151) (see Chapter 6 and 7). Therefore, the bundle length on AntConc's *N-gram* was set to "4". In corpus analysis tools, *N-gram* is a term used to designate the number of lexical items to be investigated, so 4-gram and 5-gram are equivalent to four-word bundles and five-word bundles, respectively; these terms will be used in the present study.

4.3.2.2 Key criterion 2: setting the distribution of lexical bundles

The second key criterion for bundle identification involved setting a range criterion for lexical bundles, which entails that they have to occur across different texts. Corpus-based studies employ different measures to determine the range of bundles (see Biber *et al.*, 2004; Hyland, 2008b). Wood and Appel (2014: 5) state that the range criteria adopted depend on the "research goals of the study and on the number of texts included in each corpus". Even though measures of range are closely associated with the size and type of corpora used in a particular study, the purpose of measuring the range is to ensure consistency. This consistency is important to "ensure that the lexical bundles are representative of the corpus as a whole, and not confined to only a high number of occurrences in a small amount of text or by individual writers" (Biber *et al.*, 2004; Chen and Baker, 2010; Cortes, 2006; Hyland 2008b; Wood and Appel, 2014: 5), as explained in section 2.2.2.2.

Thus, a distributional range across the instructors' materials and textbooks corpora had to be established. In this regard, it was important to set a suitable and realistic distribution threshold to generate the lexical bundle lists from the instructors' materials and textbooks corpora to be included in the analysis, based on the aims of this study and on the size of each corpora. A minimum distributional threshold was implemented to generate a sufficient number of four-word bundles to analyse and to ensure that reliable results could be obtained.

In the present study, the textbooks corpus contains four textbooks written by three different authors. The three *CAE* textbooks are written by two authors, with each author writing one complete textbook and co-authoring the third textbook. The fourth textbook, the *EWB*, was produced for the English course in the English Centre, and written by unknown author(s). The instructors' materials corpus contained 45 different-length texts, which were originally gathered from five teachers (see Chapter 3, section 3.2.4.2 for descriptions of these materials and the instructors' backgrounds). Each of the texts included the materials of one instructor who was teaching on the pre-sessional course (see Appendix K for the word count of each text in the instructors' materials).

The distribution threshold for determining four-word bundles was set to "2" for both corpora. At this stage, it was important to achieve a balance between representative lists of the pre-sessional programme as a whole (rather than being confined to only a limited number of teachers or textbooks) and meeting the distribution parameter. In addition, this distribution threshold avoided the possibility of the inclusion of frequent four-word bundles that are confined to only one academic textbook or the materials of one teacher, as reported in similar studies (e.g. Wood and Appel, 2014). However, two important concerns must be noted here which relate to the issue of representativeness (see Chapter 2, sections 2.4.2; Chapter 4, section 4.2.1.1; and Chapter 11, section 11.1.3 for details on representativeness). Firstly, the textbooks corpus is not very representative because three textbooks were produced by two authors from the same publisher, while one textbook was produced from the EAP course. Secondly, any results generated from the instructors' materials corpus is likely to be skewed due to the committed contribution of a single tutor (e.g. one instructor provided 29 documents while another only provided one). Acknowledging such limitations from an early stage is vital; this limitation can be addressed by restating the main purpose of the study (see Chapter 1) and reflecting on the reality of the EAP course being examined. Although in theory it would have been useful to have more than three authors from different publishers, and for instructors to have submitted an equal number of texts (and a higher quantity than was achieved in the present study), in practice, this was not possible.

4.3.2.3 Key criterion 3: setting the frequency threshold

Once the range of texts was set and the length of bundles was fixed, the final decision involved determining an appropriate frequency threshold to use as a cut-off to determine the bundles to be included in the lists. A significant concern to be addressed when selecting the cut-off frequency is the issue of corpus size and its relation to the research questions (see Chapter 2, section 2.4.2 and Appendix A). Regarding the corpora size, there is a large degree of consensus among many linguists that small corpora consist of "250,000 words" or less (Flowerdew, 2004: 19). The size of the textbooks corpus is 212,235 words and the instructors' materials corpus is 86,693 words; both are considered relatively small as they are under 250,000 words, but are different in size. It is important to note here that Flowerdew (2004) states, like many others, that the size of a corpus is not centred on how big or small it is; instead, it largely depends on the type of data it contains as well as on the linguistic items being investigated (see Chapter 2 for information on corpus size).

For the present study, the cut-off frequency for generating lexical bundle lists was calculated within the conventional normalized frequency threshold used for small corpora. Normally, lexical frequency measures are designed around the range of 2 to 10 times per hundred or million words; this range has been applied in previous studies of this type (e.g. Altenberg, 1998; De Cock, 1998). Since cut-off frequencies are "somewhat arbitrary" (Hyland, 2008b: 8), a minimum cut-off frequency of 2 to 10 times per million words needed to be tested in order to choose the appropriate frequency level for the two corpora.

In the "trying and testing" phase, and to generate the lexical bundle lists, a bundle length of 4-grams and a range of 2 were set for both corpora, as previously stated. I continued to adjust the frequency levels, starting at a frequency of 10 and moving downwards. In both corpora, when I reached the cut-off frequency of 2, I noticed major discrepancies between these results and those produced when the frequency level was set at 3. As shown in Table 4.4 below, for the textbooks corpus, when the frequency cut-off was adjusted to 2, AntConc generated 3,414 4-gram bundles, while when the frequency rate was set at 3, it yielded only 1,475 4-grams. This marked variation also occurred in the instructors' materials corpus.

Table 4.4. Results of frequency counts for 4-gram types across the two corpora

Corpora	Cut-off freq.	4-gram	Cut-off freq.	4-gram types
Textbooks corpus	3	1,475	2	3,414
Instructors' materials corpus	3	184	2	2,502

Lower-frequency cut-offs generate a large number of 4-grams; this is correlated with the nature of 4-grams providing a richer range of bundles, as previously mentioned. Therefore, if the n-gram size increased from 4 to 5, 6, or 7, the higher n-grams would result in an extremely limited range of bundles for each type, having the characteristic of being rare. For example, by setting the bundle length of bundle to 6 and the cut-off frequency to 3, AntConc only generated 34 bundles in the instructors' materials corpus.

I re-examined the threshold procedure for adjusting the cut-off frequency so that a sufficient number of lexical bundles would be included in the initial lists. Repeated experiments involving adjusting the frequency rate were crucial, requiring deciding to set the minimum cut-off frequency of four-word bundles to between 2 and 10 occurrences, to offer a reliable and straightforward measure for creating the two lexical bundle lists. For both corpora, setting the minimum frequency cut-off to 2 produced a large set of data from the instructors' materials corpus and textbooks corpus (2,502 and 3,414 lexical bundles, respectively) of the most frequent four-word bundles (as explained in the previous paragraph). Although this number of lexical bundles provides very rich data to analyse and explore, it would have been extremely overwhelming to work with this amount of data due to the time restrictions of this study.

In addition, I was faced with the major problem of using a fixed frequency cut-off for corpora of different sizes. For example, bundles with a cut-off frequency of 3 yielded 184 bundles in the instructors' materials corpus, but 1,475 bundles in the textbooks corpus (see Table 4.4 above). This shows major discrepancies among the bundles across the two corpora. In accordance with the methodology proposed by Biber and Barbieri (2007) and Chen and Baker (2016), I adopted the technique of using a dynamic threshold as a way of dealing with frequency variations in the bundles among corpora of different sizes. As a solution, in the instructors' materials corpus, I decided to examine four-word bundles which occur three times or more, while in the textbooks corpus the minimum frequency was set to nine occurrences. This dynamic threshold approach enables an optimum number of bundles to be derived from each corpus, ranging between 70 and 120 bundles. This range criterion yielded an adequate number of bundles to examine; thus, this overall total number of bundles (70–120) is considered to be sufficiently representative of the pre-sessional course and comparable with the AFL (see Chapter 2, section 2.4.2 for information on representativeness).

Setting a dynamic threshold for frequency helped to overcome the challenging issue of having a static cut-off point between sub-corpora of different sizes. This method is practical for "lexical bundle use when compared between sub-corpora of different sizes, ranging from over 1 million words to fewer than 40,000" (Chen and Baker, 2016: 854). According to Koester (2010), small and specialized corpora provide valuable evidence of patterns of language use in certain settings. When analysing high-frequency sequences, relatively small and specialized corpora may generate important and powerful results (Koester, 2010). In addition, they are compiled to provide answers to specific research questions (Baker, 2010: 99), such as "just texts that were published in Singapore or just newspaper reporting".

Additionally, according to Chen and Baker's (2010) study, when handling corpora of different sizes, a very important aspect that needs to be taken into consideration when setting the frequency cut-off is to explicitly report both the raw cut-off frequency and the equivalent normalized frequency. This is important in order to ensure transparency and to provide accurate measurements. In corpus linguistics, the normalization frequency can be calculated as follows:

$$Normalised\ Frequency\ (per\ hundred\ thousand) = \frac{Frequency\ in\ corpus \times Hundred\ thousand}{Total\ word\ count\ of\ corpus}$$

By calculating the normalized frequencies for both corpora, I found it very useful to report on the selected raw frequency cut-offs to provide additional information and produce an optimum number of bundles that are considered representative and sufficient for both corpora being investigated. As shown in Table 4.5 below, the normalized frequencies were 4.7 in the textbooks corpus and 4.6 in the instructors' materials corpus, corresponding to a raw cut-off frequency of 10 and 4, generating 118 and 106 bundles, respectively. From this normalized frequency count, it seems that these cut-off frequency rates are compatible, yielding an optimum number of bundles for examination. However, due to further exclusion specifications on the generated bundles (which will be fully discussed in the next section), 118 and 106 bundles, respectively, were not considered an ideal number of bundles to be examined.

Table 4.5. Normalized frequency and raw frequency calculations

Corpora	Cut-off freq.	Norm. freq.	Bundles generated	Cut-off freq.	Norm. freq.	Bundles generated	Word count
Textbooks corpus	10	4.7	118	9	4.2	145	212,235
Instructors' materials corpus	4	4.6	106	3	3.5	184	86,693

I examined very carefully the threshold procedure for adjusting the cut-off frequency so that a sufficient number of lexical bundles would be included in the initial lists. Thus, as previously stated, it was decided to opt for a lower-frequency cut-off to generate a more optimal number of bundles within the range of 200, having cut-off frequencies of 9 for the textbooks corpus and 3 for the instructors' materials corpus. Initial lists of four-word bundles meeting the aforementioned minimum frequency point and distribution creation were generated. The results led to a total of 145 bundle types in the textbooks corpus and 184 in the instructors' materials corpus, as previously reported (see Table 4.5). From these two initial lists, further exclusion specifications on the generated bundles were important to create the finalized lists: *textbooks bundles list* and *instructors' materials bundles list*.

4.4 Phase 3: exclusion specifications

Looking at the initial lists generated from AntConc, it was evident that important eliminations needed to be made to create two practical and reliable lists of four-word bundles for further examination. Thus, the initial lists from the instructors' materials corpus and textbooks corpus needed to undergo a weeding-out process, involving excluding certain sequences of words that are not considered valuable due to their unrelated or ambiguous characteristics (see Appendices M and O for full lists). The exclusion process carried out in this study followed the same weeding-out process implemented by Salazar (2014) on her master list of lexical bundles, aimed at teaching purposes. It should be noted, however, that the filtering process had to be adjusted according to the types of sequences found in the initial lists of the current study. This means that not all of the exclusion criteria found in Salazar's (2014: 45-50) study were applicable here; only relevant ones were used, based on the types of sequences generated by AntConc.

The following section will describe in detail the exclusion process:

Meaningless bundles. The first procedure that was performed involved filtering out all items or sequences of words from the initial lists that were regarded as meaningless sequences or units. This is related to the previously explained weakness of AntConc (see section 4.3.1). I explained that AntConc often produces ambiguous sets of bundles. These are bundles containing individual letters with other words and lack identifiable meaning; these are usually generated by corpus tools. To illustrate, from the instructors' materials initial list and textbooks initial list, examples such as *n c u the* and *n c something that*, respectively, are two bundles with undiscernible meaning and, therefore, were excluded from the final lists.

Faulty bundles. AntConc generated a group of sequences that were incorrectly counted as four-word bundles, producing false positive results. For example, some combinations had the same fixed wording but incorporated different punctuation marks within the bundle, which AntConc regarded as a type of four-word bundle. To illustrate, the bundle *to in order to* occurred six times, but some of its incidences incorporated distinctive kinds of punctuation marks such as a comma, dash and parentheses, including bundles such as *to*, *in order to*, *to/in order to*, and *to () in order to*. These types of combinations were excluded because although AntConc listed them as the same four-word bundles, these sequences were in fact different from one another.

Hyphenated words were a problematic issue and were also considered to produce false positive results. For example, compound words such as *non-native* and *first-year* in bundles such as *non-native speakers of* and *first-year students at*, respectively, were regarded as two words by AntConc while compound words are usually closed and were counted as one word. This means that the bundles *non-native speakers of* and *first-year students at* are three-word bundles rather than four-word bundles. Since the study focuses on four-word bundles, these types of sequences were eliminated from the analysis. In addition, apostrophes ('s) in possessives in bundles such as *someone else s words* were considered a separate item by AntConc, leading to three-word bundles, which in turn were also eliminated from the final list.

Lexical bundles interrupted by punctuation marks (e.g. hyphens, dashes, slashes, apostrophes, brackets, commas, semicolons, quotation marks, etc.) or by numbers and percentages within the four-word bundle span, and which gave false positive results, led to fault frequency information due to their inconsistencies; therefore, they were removed from the final lists. These types of bundles were thus regarded as unrepresentative. AntConc's software generated these inconsistent bundles; therefore, they should not be included in the present study. This was a limitation generated by AntConc that was normally resolved through manual checking. In addition, although these eliminated bundles have distinctive definable characteristics, they do not match the working definition of lexical bundles used in this study (see Chapter 2, section 2.2). However, it is important to mention that some of these excluded bundles (see Appendix N and P) could be considered useful and interesting for studies of a different nature with different aims and methods. As noted by Biber (2009: 290), "there are different kinds of multi-word sequences, and different quantitative methods are needed to identify them".

Web noise. Removing sequences referred to as web noise is the third procedure in the exclusion process. Some sequences were found to be part of website links or references found in the downloaded articles provided by teachers to their students, or presented in the textbooks for students to use. Bundles produced from these fragmented sequences were dealt with in two manners. Sequences that were found to be four-word bundles and were written within website links or in references were taken into consideration, such as the four-word bundle *perspectives on plagiarism and.* Bundles such as *www sciencedirect com science* were deemed as web noise and were excluded from the final lists.

Context-dependent bundles or topic-specific bundles. The fourth procedure used to identify which four-word bundles would be included in the final lists was to closely examine context-dependent bundles or topic-specific bundles (Chen and Baker, 2010; Salazar, 2014). These terms refer to sequences containing context-related word(s), usually incorporating proper nouns such as *Humanities and Social Sciences*. Alternatively, they are sequences that appeared in the same texts provided by two or more teachers/authors on a particular topic (e.g. *of the Japanese students* and *plagiarism is not considered*). These four-word bundles were found in readings or paragraphs provided to students by teachers on a specific topic such as *Plagiarism*. The decision was made to omit all context-dependent bundles from the lists. This is because the focus of this study is on academic bundles. Context-dependent bundles were mostly found in the instructors' materials bundles list but not in the textbook list, (see Appendix N for the excluded context-dependent bundles).

Overlapping. The final and most critical procedure in the exclusion process is the treatment of overlapping bundles. According to Chen and Baker (2016: 855), overlapping bundles are "four-word bundles which are actually part of a longer expression and yet, as a result of automatic retrieval, the longer expression is split into two or three shorter units". For example, the four-word bundles *at the end of* and *the end of the* are considered overlaps of the five-word bundle *at the end of the* (see Appendices N and P for full lists).

Looking at how previous corpus-based studies dealt with the exclusion of overlapping bundles showed that it is a problematic matter. Some researchers investigating lexical bundles have approached this difficulty by allowing bundles with significant overlap to be included in the analysis. Returning to the bundles *at the end of* and *the end of the*, in order to provide a thorough examination of bundles, Cortes (2004) included these two overlapping bundles in her four-word bundle list created from published academic writing in biology.

However, a few studies (e.g. Chen and Baker, 2010, 2016; Salazar, 2014; Wood and Appel, 2014) have attempted to address this issue by finding a solution that best fits their research goals. Thus, different exclusion criteria on lexical bundles have been employed in corpus-based studies. On the one hand, the exclusion process is considered "methodologically tricky and open to claims of subjectivity", as reported by Simpson-Vlach and Ellis (2010: 4). This means that the intuitive exclusion process is a questionable, challenging, and demanding issue facing researchers investigating lexical bundles. On the other hand, overlapping bundles are considered repeated items which are incorporated into a longer bundle; keeping them could be misleading and provide inaccurate analysis, which could lead to inflated results (Chen and Baker, 2010).

To overcome this dilemma, combining overlapping bundles into one longer bundle is one tactic that has been applied to solve this problem (Chen and Baker, 2010; Wood and Appel, 2014). For example, Wood and Appel (2014) handled the bundles *at the end of* and *the end of the* by combining the two bundles into one condensed combination (*at) the end of (the)* to create a condensed listing. Bundles sharing similar structures but with different meaning were not combined, such as the bundles *is one of the* and *one of the most.*

The decision over whether to keep or remove the overlapping bundles was also carefully considered in the present study. Similar to Salazar (2014), Wood and Appel (2014), and Chen and Baker (2010, 2016), I found the exclusion process for overlapping bundles to be a necessary step, in order to produce clear, organized and comparable lists. Thus, I decided to deal with overlapping bundles by finding a solution that combines the two approaches. The process of weeding-out overlapping bundles should be handled efficiently and with caution to suit the study aims. Overlapping bundles were manually checked via concordance checks to make important judgements.

In the following, the three types of exclusion decision on overlapping bundles will be detailed, providing information on the weeding-out process.

Type 1: Two four-word bundles with a similar frequency level and shared three-word structures were both checked through concordance lines. Consider, for example, the bundles *are a number of* and *there are a number*, which both occurred four times across two texts in the instructors' materials corpus. For these types of bundles, one bundle was kept and the other was removed from the final list to avoid unnecessary repetition when examining bundles. Similar to the previous example, from the textbook list, the bundles *complete the following sentence* and *the following sentence using* both occurred 13 times in two different texts. In such cases, one four-word bundle was included in the final list; while the other bundle was omitted. Concordance lines for these two bundles showed that not only did the bundle *complete the following sentence* include the structure of the bundle *the following sentence using*, but also provided additional useful information compared to the other bundle.

Type 2: This exclusion decision focused on two four-word bundles that share very similar frequency levels and incorporate three-word structures. In cases like these, bundles with higher occurrences prevailed and were kept while overlapping lower-frequency bundles were excluded. Closer inspection of the concordance lines showed that the bundle *due to the fact* (frequency of 3) is part of the bundle *to the fact that* (frequency of 4). This led to including the higher-frequency bundle due to its added informational feature.

Type 3: Unlike the previous two types of exclusion decision, this category incorporates two four-word bundles that have different frequencies. In this type, both bundles were preserved for their high frequency rate and because each bundle provides useful information that the other bundle does not provide. From the textbook list, the overlapping bundles *at the end of* and *the end of the* occurred 27 and 16 times, respectively; both bundles were retained as they both provide further information.

Table 4.6 shows the above-mentioned exclusion specifications along with examples. From an operational point of view, in the present study, overlapping and context-dependent bundles were removed from the final lists and were not included in the analysis (see removed overlapping and context-dependent bundles in Appendix N and P, respectively). The elimination of overlapping bundles from the final lists was inevitable to avoid unnecessary repetition or redundant bundles in the analysis. Regarding context-dependent bundles, similar to Chen and Baker's (2010) work on bundles, they were removed because they are more bound to a certain topic than to an academic feature.

Table 4.6. Exclusion specifications with examples

Excluded Bundles	Examples
Meaningless bundles	*n c something that, j the effect of, n c u the*
Faulty bundles	*the writer's position, in pairs, and compare, of first-year students*
Web noise	*a m roy eds, http www sciencedirect com, Cambridge Cambridge university press*
Context-dependent bundles or topic-specific bundles	*English for Academic Purposes, plagiarism and intellectual property, students in Hong Kong*

Overlapping	*in japan and the, is the best way, are going to read*

The number of bundles was reduced after the weeding-out and exclusion steps were applied (see Figure 4.1 and Table 4.7). This exclusion process produced clear, effective and concise lists that could be examined quantitatively (statistical measures) and qualitatively (concordance checks). These final four-word bundle lists appear to genuinely reflect the language presented in the EAP pre-sessional course examined in this study, yielding reliable and representative results of bundles found in EAP materials used within the examined institution, as stated earlier in this section. However, it is vital to note that it is not possible to achieve complete representativeness of EAP materials in general, which would require more than one institution and a wide range of EAP materials (textbooks, instructors' materials and syllabuses) to be examined. Thus, the present study does not aim to achieve complete representativeness of lexical bundles used in EAP pre-sessional courses and EAP materials. In accordance with the present study's research questions (see Chapter 2, section 2.4.6.3 and Appendix A), the representativeness issue in this study was resolved by collecting enough texts to accurately represent the type of language being examined (see Chapter 2, section 2.4.2 and Chapter 11, section 11.1.1 for more details).

In addition, it is a manageable size for analysis within the time frame of the present study. The finalized list for the instructors' materials corpus contains 79 bundles (see Chapter 5, section 5.1 for a full list of bundles), while the finalized list for the textbooks corpus consists of 102 bundles (see Chapter 6, section 6.1). These two lists include frequency and range (number of texts) data.

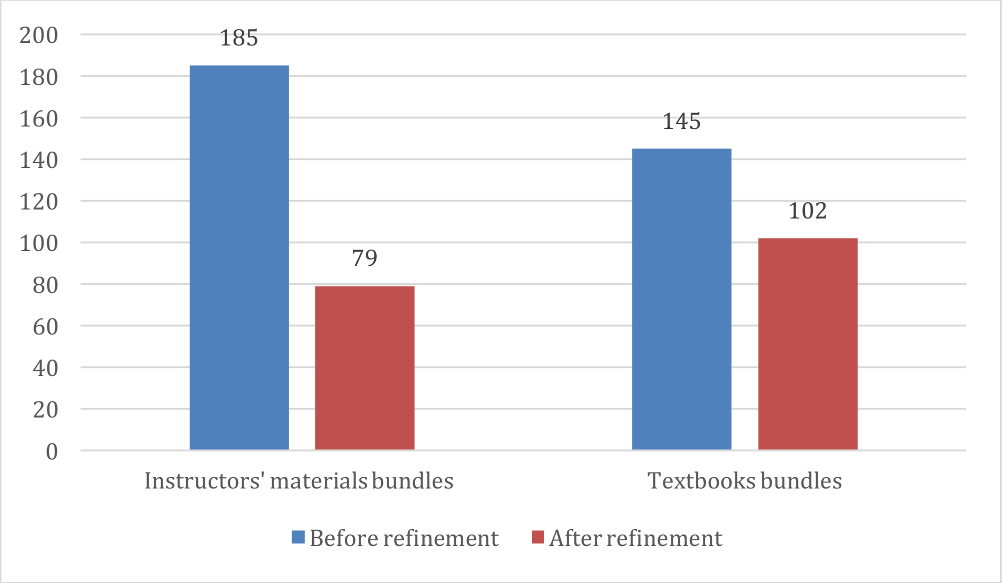

Figure 4.1. Number of bundles before and after the exclusion and weeding-out process

Table 4.7. Number of bundles before and after the weeding-out and exclusion process

	Description of bundle lists	Corpus	
		Instructors' materials corpus	Textbooks corpus
		No. of bundles	No. of bundles
Before refinement	Initial	184	145
	Weeded out	43	20
	Excluded	62	23
After refinement	Final	79	102

Using an adequate number of bundles from both corpora (79 and 102 bundles, respectively), I managed to extract sufficient examples of the most frequent four-word bundles. This operational number of bundles allowed me to provide a full explanation and reliable descriptions of the bundles' functions and structures with examples. Also, it made it possible to compare those bundles with the AFL through concordance, thus answering research question 3.

In addition, it is important to note that I did not perform type/token ratio calculations because it was found to be irrelevant to the present study. A *type* refers to a single and distinct form of lexical bundle in a corpus, such as *on the other hand* and *as a result of*, while a *token* refers to the occurrences or repeat uses of the same bundle. For example, the bundle *on the other hand* might occur 18 times in one corpus and 15 times in another. The type/token ratio is simply “the number of types divided by the number of tokens expressed as percentages” (Baker, 2006: 52). Conducting a type/token ratio calculation is useful when comparing bundles across corpora (Baker, 2006; Chen and Baker, 2010). However, the AFL (Simpson-Vlach and Ellis, 2010) that is used in the present study as a point of comparison did not provide a type/token ratio in their study. It is for this reason that frequency counts for bundle *types* are reported without reporting *token* counts and without performing type/token ratio calculations.

4.5 Procedure followed for lexical bundle analysis

Once a manageable number of four-word bundles had been produced, the next step was to address the measures needed to perform the analysis. Thus, in this section, I will provide information on the procedures followed to present the analysis in the present study.

4.5.1 Frequency analysis

To answer research question 1, where I was interested in ascertaining the most frequent four-word bundles in EAP materials aimed at teaching academic writing, I obtained quantitative information about the occurrence of the retrieved bundles. This was accomplished by reporting frequency counts of the most frequent four-word bundles across the textbooks bundles list and the instructors’ materials bundles list (see Chapter 5 and 6).

4.5.2 Functional classification analysis

In addition to undertaking a quantitative analysis of the occurrences of bundles, I conducted a qualitative analysis after manually categorizing the functional relations of these bundles. In order to accomplish this, an in-context examination of bundles was conducted, ensuring that each four-word bundle was functionally classified through concordance checks to help answer research question 2 (how are the most frequent four-word bundles classified functionally)?

Thus, an important step in the analysis of the four-word bundles was to classify them functionally, according to their corresponding meaning in context, using the concordance program AntConc. I credit the work of Simpson-Vlach and Ellis (2010) (see Chapter 2, section 2.2.3.2), whose classification scheme was found to be particularly useful for this study, with a slight adjustment.

Simpson-Vlach and Ellis's (2010) broad groupings were maintained for both lists: *stance expressions*, *discourse organizing functions*, and *referential expressions.* Two sub-categories, one in the *referential expressions* and the other in *stance expressions*, were eliminated due to the types of bundles found in my study. This meant that the sub-categories *vagueness markers* and *intention/violation prediction* in *referential expressions* and *stance expressions*, respectively, were discarded. Thus, Table 4.8 only includes the functional sub-categories from Simpson-Vlach and Ellis's taxonomy that were relevant to my study, along with examples from the textbook and instructors' materials corpora.

Table 4.8. Functional taxonomy (adapted from Simpson-Vlach and Ellis, 2010: 498–502) for instructors' materials bundles and textbooks bundles

Referential expressions	Stance expressions	Discourse organizing functions
(1) Specification of attributes (a) Intangible framing attributes *the extent to which*, *in a way that* (b) Tangible framing attributes *as part of the*, *as part of a* (c) Quantity specification *there are a number of*, *a wide range of*	(1) Hedges *are more likely to*, *is more of a*	(1) Metadiscourse and textual reference *the words in the*
(2) Identification and focus *of his or her*, *in your own words*	(2) Epistemic stance *is not considered a*, *what do you think*	(2) Topic introduction and focus *look at the following*
(3) Contrast and comparison *the relationship between*, *what is the difference*	(3) Obligation and directive *you may need to*, *focus on your subject*	(3) Topic elaboration (a) non-causal *the purpose of this,* (b) Topic elaboration: cause and effect *in order to avoid*
(4) Deictics and locatives *at the end of*, *in the United States*	(4) Expressions of ability and possibility *can be used to*, *can you think of*	(4) Discourse markers *at the same time*
	(5) Evaluation *it is important to*, *is the most important*	

4.5.3 Structural classification analysis

Besides analysing the bundles functionally, a structural categorization of bundles in the textbooks bundles list and instructors' materials bundles list, in terms of their grammatical types, was applied. To accomplish this, an in-context check of bundles was conducted, ensuring that each four-word bundle was structurally classified through concordance checks to help answer research question 2 (how are the most frequent four-word bundles classified structurally)? The framework used for structuring lexical bundles is commonly known and used in corpus linguistics on bundles found in academic prose. Biber *et al.* (2004) established that lexical bundles have strong grammatical structural correlates, as repeatedly mentioned in this thesis. The structural classification of the identified four-word bundles, summarized in Table 4.9 and Table 4.10, was organized and grouped according to Biber *et al.*'s (1999) study (see Chapter 2, section 2.2.3.1). The current structural taxonomy is an adaptation of the structural classification scheme outlined in their study, but with differences to reflect the structures found in the textbooks and instructors' materials bundles lists.

For each list (textbooks and instructors' materials), bundles were placed into categories after I examined them in context by carrying out concordance line checks using the AntConc tool. Unlike the functional classifications, new and different categories were added to both lists. As illustrated in Table 4.9, three new sub-categories were added to the instructors' materials bundles list with examples for each category:

- Prepositional phrase with *to*-clause
- Passive verb + noun phrase fragment
- Verb phrase with pronouns.

Table 4.9. Structural patterns (adapted from Biber *et al.*, 1999: 1014–1021) for instructors' materials bundles list

Structural patterns	Examples
Noun phrase with *of*-phrase fragment	*the end of the*, *the results of the*
Noun phrase with other post-modifier fragments	*the extent to which*
Prepositional phrase with embedded *of*-phrase fragment	*at the end of*, *as a result of*

Other prepositional phrase (fragment)	*in the United States*, *on the other hand*
Prepositional phrase with *to*-clause	*in order to avoid*
Anticipatory *it* + verb phrase/adjective phrase	*it is important to*, *it is possible that*
Passive verb + prepositional phrase fragment	*can be used in*
Copula *be* + adjective phrase	*be aware of the*
(Passive) verb phrase + *to*-clause fragment	*can be used to*
To-clause fragment	*to be one of*
Passive verb + noun phrase fragment	*is not considered a*
(verb phrase +) *that*-clause fragment	*have shown that the*
that-clause fragment	*that he or she*
(verb/adjective +) *to*-clause fragment	*are more likely to*
Verb phrase with pronoun	*he was unable to*
Pronoun/noun phrase + be (+…)	*there are a number*
Other expressions	*by no means the*

In the textbooks bundles list, seven new sub-categories were added to the structural classification, which are, as shown in Table 4.10:

- Verb phrase/noun phrase + *that*-clause fragment
- Verb phrase + determiner phrase fragment
- Verb phrase + prepositional phrase fragment
- Verb phrase + noun phrase fragment
- WH-question phrases
- 2nd person pronoun you + VP fragment
- Other expressions

Table 4.10. Structural patterns (adapted from Biber *et al.*, 1999, pp. 1014–1021) for textbooks bundles list

Structural patterns	Examples
Noun phrase with *of*-phrase fragment	*the end of the*, *a great deal of*
Noun phrase with other post-modifier fragments	*the words in the*, *the ways in which*
Prepositional phrase with embedded *of*-phrase fragment	*at the end of*, *as a result of*
Other prepositional phrase (fragment)	*from the text in*, *in the United*

	States
Anticipatory *it* + verb phrase/ adjective phrase	*it is important to*, *it is possible to*
Passive verb + prepositional phrase fragment	*used in the text*, *be followed by a*
Copula *be* + adjective phrase	*is the most important*
(verb phrase/noun phrase +) *that*-clause fragment	*research shows that in,*
Noun + verb phrase + *that*-clause	*research shows that the*
That-clause fragment	*that something is true*
(Verb /adjective +) *to*-clause fragment	*are more likely to*
Predicative adjective + *to*-clause	
(Passive) verb phrase + *to*-clause fragment	*have been asked to*, *can be used to*
Verb phrase/noun phrase +) *that*-clause fragment	*research shows that in*
To-clause fragment	*to check your answers*
Verb phrase + determiner phrase fragment	*focus on your subject*
Verb phrase + prepositional phrase fragment	*work in pairs and*, *look again at the*
Verb phrase + noun phrase fragment	*look at the following, and answer the questions*
WH-question phrases	*what do you think*, *what you have read*
2nd person pronoun you + verb phrase fragment	*you are going to*, *you have been given*
Other expressions	*common in academic writing*, *each of the following*

The full classification of the structural and functional framework for the 79 bundles found in the instructors' materials bundles list can be seen in Chapter 5, sections 5.2 and 5.3. The 102 bundles listed in the textbooks are classified structurally and functionally and fully presented in Chapter 6, section 6.2 and 6.3.

4.5.4 Comparison of corpora

The next procedure in the analysis involved comparing the use of lexical bundles in the instructors' materials bundles list and the textbooks bundles list to the AFL. Firstly, I will describe the AFL and provide justification for why it was selected. Secondly, I will describe the modification I made to the AFL and its functional list to make it suitable for use in the present study.

4.5.4.1 The rationale behind using the AFL

The academic writing corpus used in the AFL is composed of Hyland's (2004) research article corpus, comprising 1.2 million words, in addition to selected BNC files, comprising 931,000 words. The writing corpus of more than 2 million words was divided into four sub-corpora by academic discipline, including Humanities and Arts (360,520), Social Sciences (893,925), Natural Sciences/Medicine (513,586), and Technology and Engineering (349,838). This stage in the analytical procedure is crucial for four reasons. It is important to briefly summarize why I selected the AFL and its corresponding functional list to compare with the EAP material lists created. Simpson-Vlach and Ellis's (2010) AFL has been selected to serve as a comparative tool in the present study for the following reasons:

- The AFL is a corpus-driven research tool which lists formulaic sequences known as "frequent recurrent patterns" or lexical bundles in written and spoken academic language corpora. While the AFL presents sequences that are common in academic speech and writing, the present study will only focus on the written AFL top 200 list related to academic writing. The written AFL top 200 list was modified (as explained in the following section) for the purpose of this study. Appendix B presents the written AFL top 200 list before my modifications (available at Applied Linguistics online as supplementary material; Simpson-Vlach and Ellis, 2010).
- The AFL occurs more commonly in academic discourse than in non-academic discourse. Because the AFL covers a wide range of academic genres, it is considered relevant to and useful for the EAP context rather than being restricted to a specific discipline. In addition, the AFL was influenced by Coxhead's (2000) AWL, which served as a significant teaching resource in the EAP context (Simpson-Vlach and Ellis, 2010; Gardner and Davies, 2014). As a result, the AFL may indeed provide the same impact and significance for teaching and research in the EAP context.
- The AFL was derived empirically through using frequency counts for ranking sequences extracted from a corpus and also by employing an important

statistical measure called Mutual Information (MI), which estimates the degree of association between words in a phrase occurring together more frequently than by chance (see Chapter 2, section 2.2.2.4). As stated previously, a higher MI score indicates that there is a strong association between the pair of words, while a lower MI score shows that the co-occurrence is more possibly due to chance. In addition, Simpson-Vlach and Ellis did not want to rely solely on frequency counts to rank the bundles in their list. This is because lower-frequency bundles will not make it into the top frequency-ordered list. Thus, they wanted another corpus metric to rank the formulas in their list. High MI sequences tend to have more coherence than expected by chance, relating to distinctive functions or meaning in context. However, the MI score tends to recognize rare sequences comprised of rare component words. This led Simpson-Vlach and Ellis to combine the information generated by both metrics (frequency and MI) to rank the academic formulas to ensure legitimate results which are not based on intuition and to present the bundles to instructors to obtain a FTW score, which will be explained next.

- Besides using statistics, Simpson-Vlach and Ellis (2010: 488) used a psychological measure of utility called 'formula teaching worth' (FTW) to rank their bundles, in which the researchers asked 20 experienced instructors and language testers to rate the formulas in terms of teaching worth. The FTW scores were based on instructors' judgments in relation to whether the formulas are worth teaching. The formulas were provided to them in a random order, and they were asked to indicate their disagreement or agreement over whether the formulas are worth teaching on a scale of 1 (disagree) to 5 (agree). This method enables the production of a reliable and valid ranked arrangement of formulas and serves as a guide for instructors and materials writers in arranging formulaic sequences for instructional uses.

4.5.4.2 Modifications to the AFL and its corresponding functional and structural list

This section details what sort of modifications had to be applied to the AFL and its corresponding functional list to create a practical measuring tool for this study. It also laid the groundwork for performing the functional and structural comparisons, which will answer research question 3.

Simpson-Vlach and Ellis (2010) presented and organized the AFL into three sublists. The first list is the core AFL list, which includes three-, four-, and five-word bundles from both the spoken and written AFL. The second list includes three-, four-, and five-word bundles from the top 200 formulas of the spoken AFL, and the third list includes three-, four-, and five-word bundles from the top 200 formulas of the written AFL. As mentioned previously, the present study will only use the third list, which focuses on the written AFL sequences, sorted by frequency (see Appendix B). However, several important adaptations needed to be made to the written AFL top 200 list before it could be used, in order to be able to perform comparisons and conduct an in-depth analysis that is particularly relevant and suitable for this study.

It was previously mentioned that the written AFL includes three-, four-, and five-word bundles; however, since the present study involves identifying and examining four-word bundles, the first modification involved manually excluding all but four-word bundles, forming a modified list (called the *written AFL sub-list*) containing a little over 57 four-word bundles. This means that the top 200 formulas of the written AFL was reduced to a third of its original size. This reduction is natural because the list only included four-word bundles, and it yielded a manageable amount of lexical bundles to analyse. Focusing on four-word bundles allowed me to conduct an in-depth examination and comparison of the sequences found in the EAP material through concordance checks, as mentioned earlier. The *written AFL sub-list* also includes the frequency measures of four-word bundles that appear in Simpson-Vlach and Ellis (2010), as shown in Appendix C, which was useful for examining frequencies (see discussion chapter).

The second modification relates to the AFL functional list provided by Simpson-Vlach and Ellis (2010: 498–502). Their functional list was used because it describes the different functions of the bundles in the top 200 formulas of the written AFL. It is important to understand how Simpson-Vlach and Ellis presented their AFL functional categorization list before attempting to discuss the modifications to the list made in the present study. Simpson-Vlach and Ellis organized the bundles under three groups according to their functions in texts: *Referential expressions*, *Stance expressions*, and *Discourse organization functions* (see section 2.2.3.2). Within each of these groups, additional sub-categories were included. For example, the *Stance expressions* group had a sub-category called *Specification of attributes*, which also had another sub-category called *Intangible framing attributes*.

In addition, Simpson-Vlach and Ellis organized these bundles into three sets and according to register, specifying whether the bundles were from the core AFL set (including bundles from both the spoken and written registers), the primarily written set, or the primarily spoken set. To illustrate, the bundle *more likely to be* is found in the core AFL set (written and spoken register) under the category *Stance expressions* and under the sub-category *Hedges*. In contrast, the bundle *might be able to* is found only in the primarily spoken set and the bundle *is likely to be* is found only in the primarily written set, but both bundles are found in the same category and sub-category, as shown in Table 4.11.

Table 4.11. An example of the AFL functional classification of four-word bundles, according to Simpson-Vlach and Ellis's (2010: 498–502) categorization

Group B. Stance expressions **Hedges**
Core AFL set (written and spoken) *more likely to be*
Primarily spoken set *might be able to*
Primarily written set *is likely to be*

From Simpson-Vlach and Ellis's (2010: 498–502) AFL categorized by function, I looked for the corresponding functions of the four-word bundles that I included in the written AFL sub-list (57 bundles). I went over the functions in each of the three groups and their sub-categories, focusing on only selecting four-word bundles that appeared in the written set, and I created the AFL functional categorization framework list. Some of the four-word bundles that were not found in the primarily written set were found in the core AFL set, and I still included these in the AFL functional categorization framework list. This is because although they appeared in the core set, they are bundles that are frequently used in the written register as well as in the spoken register. Appendix D presents a table of the *Functional categorization of the four-word bundles in the written AFL sub-list in terms of frequency per million words* (57 bundles), ready to be used in the comparisons in this study.

One final addition that was necessary to fully prepare for the comparison was to have a *Structural Categorization o*f the four-word bundles in the *written AFL sub-list in terms of frequency per million words*, which was produced by using Biber *et al.*'s (1999) structural taxonomy (see Appendix F for the list and section 2.2.3.1 for the structural classification). Table 4.12 displays some examples of the *AFL Structural Categorization Framework* of four-word bundles that was used in the comparison.

Table 4.12. Examples of the *Structural Categorization Framework* of four-word bundles, according to Simpson-Vlach and Ellis's (2010) taxonomies

Structure	Examples
Noun phrase with *of*-phrase fragment	*the nature of the*, *the purpose of the*
Noun phrase with other post-modifier fragments	*such a way that*, *an increase in the*
Prepositional phrase with embedded *of*-phrase fragment	*as part of the*, *in the form of the*
Other prepositional phrase (fragment)	*with respect to the*, *on the other hand*
Anticipatory *it* + verb phrase/adjective phrase	*it is likely that*, *it should be noted*
Passive verb + prepositional phrase fragment	*is based on the*, *be related to the*
(Verb phrase +) *that*-clause fragment	*is that it is*, *that there is no*
(Verb/adjective +) *to*-clause fragment	*is likely to be*, *to ensure that the*
Adverbial clause fragment	*as can be seen*
Pronoun/ noun phrase + *be* (+ …)	*there are a number*

4.5.5 In-context analysis of bundle types

This part of the analysis was informed by the studies by Wood and Appel (2014) and Wood (2010). Examining the bundles in their environment (in context) reveals information on whether the lexical bundles mostly occurred in instructional or reading parts (see Appendices Q and R for full lists). Table 4.13 provides a few examples of the in-context information. Wood and Appel (2014) constructed two corpora. The first corpora included bundles identified in reading parts, while the second corpus incorporated bundles from instructional parts. Adopting a different approach, but with a similar perspective to Wood and Appel (2014) and Wood (2010), I did not consider separating the reading text parts from the instructional text parts in my EAP corpora. This is because I wanted to find out if my corpora would yield similar or different results to Wood and Appel (2014) and Wood (2010). I was also determined to report on bundles in terms of in-context information and their overall proportions. This gives another dimension to the examination of EAP materials, enabling a complete picture to be produced of the bundles EAP learners are most frequently exposed to. This includes establishing the percentage occurrences of the two types of textual environments in which the bundles appear, as fully explored in the discussion chapter.

Table 4.13. Some examples of bundles in terms of in-context information in the instructors' materials and textbooks bundles lists

Instructors' materials bundles list	Textbooks bundles list
Bundles in textual parts *in the United States*	**Bundles in textual parts** *as a result of*
Bundles in instructional parts *title of the article*	**Bundles in instructional parts** *focus on your subject*
Bundles in both parts *at the end of*	**Bundles in both parts** *it is important to*

4.5.6 Bundle types as teachable units analysis

For pedagogical purposes, when analysing the most frequent four-word lexical bundles in EAP lists from an EAP pre-sessional programme in the UK particularly aimed at teaching academic writing, a key task was to examine the extent to which the bundles are considered as "teachable units", a phrase used by Wood and Appel (2014), and which requires *explicit teaching*. This phrase refers to the inclusion of sets of tasks and exercises which focus learners' attention on lexical bundles and give them opportunity to practice them. This entails raising students' awareness of the nature of bundles and including a sufficient number of exercises on lexical bundles.

This is important to highlight if EAP instructors' materials and textbooks include useful types of academic lexical bundles for learners to be aware of or to use in their written projects. This study only examines EAP materials; however, any other EAP pedagogical aspects related to bundle teaching are not within the focus of the study. This means that EAP learners could be exposed to lexical bundles from other resources, but it is not within the scope of the present study to look into their usefulness.

This task was accomplished by reflecting on the in-context information of the types of bundles, focusing on them in their contextualized environment in both the instructors' materials and textbooks, through checking concordance lines. I examined bundles to ascertain if there were any exercises or tasks that students needed to accomplish. Unlike Wood and Appel's (2014) study, the present study includes concordance line checks to report on the pedagogical aspect of lexical bundles in textbooks and instructors' materials to establish whether the exercises explicitly examined lexical bundles. For example, the bundle *on the other hand* was explicitly highlighted to learners in both EAP corpora, as will be discussed in the discussion chapter.

4.6 Concluding remarks

This methodological section explained the process of building the required corpora. Then, it elaborated on how a combination of frequency, distribution, and length of bundles criteria were used to create two lists for examining bundles from the two corpora of instructors' materials and textbooks, which are: *the textbooks bundles list* and *instructors' materials bundles list*. The chapter also presented the exclusion and weeding-out specifications needed to create reliable lists.

In addition, the procedure section elaborated on the functional and structural classifications of the most frequent four-word bundles found in the EAP lists and in the written AFL sub-list. It provided a basis for comparisons of types, structures, and functions of bundles to be conducted between the three corpora. As an analytical step, this chapter also provided a justification for the modification steps made in the present study on the AFL in order to prepare for the comparisons. Finally, it accounted for examining in-context information and the pedagogical treatment of lexical bundles.

In the next chapters (5 and 6), I will concentrate on the lexical bundles in the instructors' and textbooks materials corpora, respectively, and explore their types, frequency, functions and structural characteristics.

CHAPTER V

QUANTITATIVE RESULTS: INSTRUCTORS' MATERIALS CORPUS – THE LIST OF BUNDLES

This chapter is concerned with addressing research questions 1 ("What are the most frequent four-word bundles profiled in textbooks and instructors' materials, in an EAP pre-sessional programme in the UK particularly aimed at teaching academic writing?") and 2 ("How are the most frequent four-word bundles classified functionally and structurally?") in regard to the *instructors' materials corpus*. The chapter presents the quantitative results from the analysis of the most frequently occurring four-word bundles in the list, which was derived from the instructors' corpus. It focuses on three major aspects of bundle analysis. It describes the overall frequency counts, in terms of bundle types, details the structural patterns, and provides information on the functional categorization of the identified four-word bundles. At this stage, the empirical findings will be described in relation to the raw frequency counts, involving the actual occurrence of lexical bundles in the corpus. For practical purposes, when introducing a bundle, I will use parentheses to enclose information related to its frequency and range information (when needed). Example: The bundle *on the other hand* (15 occurrences in 3 texts) occurs in the instructors' corpus.

5.1 Quantitative results: types and frequency distribution

Table 5.1 below lists the most frequently occurring four-word bundles in the instructors' materials bundles list, in order of frequency, including range information, which here represents teachers' texts. The final four-word list consists of 79 different bundle types in the 86,693-word corpus, after removing excluded, overlapping, and context-based bundles (see Appendix M and N for these lists), as reported in Chapter 4. These 79 bundles amount to a total of 28,519 individual cases, which account for 33% of the more than 80,000 words in the instructors' corpus.

Table 5.1. Lexical bundles in instructors' materials bundles list by order of frequency with range information

Total No. of N-Gram Types: 79			
Rank	Frequency	Range	Bundles
	18	3	*in the United States*
	15	3	*at the end of*
	15	3	*on the other hand*
	12	4	*the end of the*
	10	4	*it is important to*
	9	2	*the extent to which*
	8	4	*as a result of*
	8	2	*in the field of*
	7	2	*with the help of*
	6	2	*and the United States*
	6	2	*at the beginning of*
	6	2	*in the middle of*
	6	2	*title of the article*
	5	3	*a great deal of*
	5	2	*are more likely to*
	5	2	*form of the verb*
	5	2	*is not considered a*
	5	2	*it is possible that*
	5	2	*it is quite possible*
	5	2	*the beginning of the*
	5	2	*the best way to*
	5	2	*the results of the*
	5	2	*to know each other*
	4	3	*as one of the*
	4	2	*as opposed to only*
	4	3	*at the time of*
	4	2	*at the university of*
	4	3	*can be used to*
	4	4	*for a long time*
	4	2	*from the department of*
	4	2	*have shown that the*
	4	2	*in the development of*
	4	2	*is important to note*
	4	2	*it is interesting that*
	4	2	*the fact that the*
	4	2	*the purpose of this*
	4	3	*the relationship between the*
	4	2	*there are a number*
	4	2	*to be one of*
	4	2	*to the fact that*
	3	2	*a third of the*
	3	3	*all over the world*
	3	3	*as a consequence of*
	3	2	*as part of the*
	3	3	*be aware of the*
	3	2	*be done by a*
	3	2	*by no means the*
	3	2	*can be used in*
	3	2	*considered a major issue*

3	3	*few and far between*
3	2	*he was unable to*
3	3	*in countries such as*
3	2	*in order to avoid*
3	2	*in order to learn*
3	2	*in order to test*
3	2	*in relation to the*
3	2	*in the case of*
3	2	*in which it is*
3	2	*is more of a*
3	3	*it is difficult to*
3	3	*it is possible to*
3	2	*lack of understanding of*
3	2	*may or may not*
3	2	*of different types of*
3	3	*of his or her*
3	2	*on a regular basis*
3	2	*on the one hand*
3	2	*part of a larger*
3	2	*purposes of this study*
3	3	*that he or she*
3	3	*that it is a*
3	2	*the form of the*
3	2	*the name of the*
3	2	*the passive form of*
3	2	*this has been the*
3	2	*this is not necessarily*
3	2	*through the use of*
3	2	*to be the most*
3	2	*you may need to*

Table 5.1 shows that the most frequent four-word bundles have different occurrences in the list (18, 15, 12, 10, and 7 occurrences). For example, the bundle *in the United States* occurs the most frequently (18 occurrences in 3 texts). Furthermore, the bundles *at the end of* and *on the other hand* both occur 15 times in three texts. In addition, *the end of the* (12 occurrences), *it is important to* (10 occurrences), *the extent to which* (9 occurrences), *as a result of* and *in the field of* (8 occurrences each), and *with the help of* (7 occurrences) are all among the top-ranked most frequently occurring four-word bundles in the instructors' corpus.

The remaining four-word bundles from the instructors' corpus have different frequency rates, ranging from six to three occurrences. Some examples of these bundles include the following: *in the middle of* (6 occurrences), *are more likely to* (5 occurrences), *as one of the* (4 occurrences), and *a third of the* (3 occurrences); see Table 5.1.

A larger number of four-word bundles found in the instructors' materials bundles list are of lower frequency (3 or 4 occurrences). Further calculations reveal essential information regarding their overall percentages in terms of bundle types. Figure 5.1 presents the overall distribution of frequencies of N-gram types in the instructors' materials bundles list. Bundles with a frequency of four account for 22% of the bundle types, including bundles such as *at the time of*, *can be used to*, *the fact that the*, and *the purpose of this* (see Table 5.1). Furthermore, Figure 5.1 shows that the majority of bundles in the instructors' materials bundles list have a frequency of three, accounting for an overall 49% of bundle types. This means that almost half of the bundles found in the instructors' materials bundles list occur three times. Although most instructors' materials bundles have a low frequency of three, they are considered the most varied in the corpus. The bundles include *a third of the*, *all over the world*, *as a consequence of*, *as part of the*, and *can be used in*, as shown in Table 5.1.

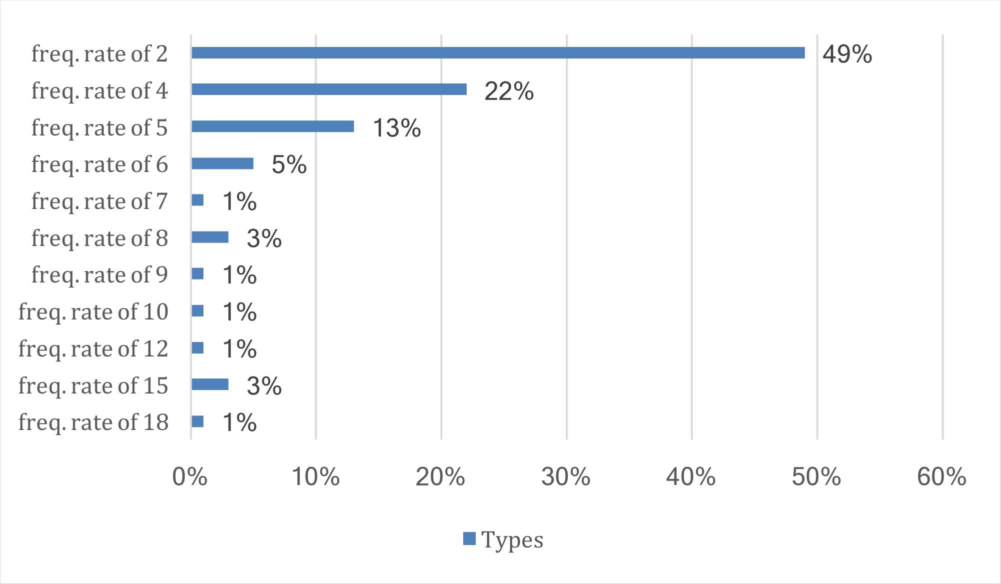

Figure 5.1. Overall distribution of the 79 bundles in the instructors' materials bundles list (types)

As for the bundles with occurrences of 18, 15, 12, and 10, although these bundles occur the most frequently in the instructors' materials bundles list, they only account for 1%, 3%, 1%, and 1% of the total number of bundle types, respectively, as shown in Figure 5.1. This means that bundles with a high frequency tend to be narrow in terms of types. While the types of the top-ranked four-word bundles are limited, occurring at low percentages compared to bundles with a frequency of 3 and 4, this result indicates that the top-ranked four-word bundles have significant usage, as will be discussed further.

5.2 Structural classification of four-word bundles in the instructors' materials bundles list

The structural classification of the 79 most frequent four-word bundles in the instructors' materials bundles list follows the taxonomy provided in the Longman Grammar of Spoken and Written English (LSWE) (the academic prose part) (Biber *et al.*, 1999: 1014–1024), as explained in the procedure chapter (see Chapter 4). The lexical bundles were grouped into four main structural patterns: Noun Phrases (NP-based, e.g. *the end of the*), Prepositional Phrases (PP-based, e.g. *at the end of*), Verb Phrases (VP-based, e.g. *it is important to*), and Other Expressions (such as *and the United States* and *as opposed to only*). Table 5.2 presents the 79 most frequently occurring four-word bundles in the instructors' materials bundles list according to their structural categories, with corresponding frequencies and range information. All of Biber's sub-categorizations of lexical bundles were similar to those in my corpus, with slight differences (as explained in the procedure section; see Chapter 4, section 4.5.3). As stated in Chapter 4, this structural framework has been widely used and applied to lexical bundles to explore their basic structural units and to understand their characteristics in an EAP register.

Table 5.2. Structural patterns of the four-word bundles in the instructors' materials bundles list

Categories	Frequency	Range	Lexical bundles
(1) NP-based			
1 Noun phrase with *of*-phrase fragment			
	12	4	*the end of the*
	6	2	*title of the article*
	5	3	*a great deal of*
	5	2	*form of the verb*
	5	2	*the beginning of the*
	5	2	*the results of the*
	4	2	*the purpose of this*
	3	2	*a third of the*
	3	2	*lack of understanding of*
	3	2	*part of a larger*
	3	2	*purposes of this study*
	3	2	*the form of the*
	3	2	*the name of the*
	3	2	*the passive form of*
2 Noun phrase with other post-modifier fragments			
	9	2	*the extent to which*

		5	2	*the best way to*
		4	2	*the fact that the*
		4	3	*the relationship between the*
(2) PP-based				
1	Prepositional phrase with embedded *of*-phrase fragment			
		15	3	*at the end of*
		8	4	*as a result of*
		8	2	*in the field of*
		7	2	*with the help of*
		6	2	*at the beginning of*
		6	2	*in the middle of*
		4	3	*as one of the*
		4	3	*at the time of*
		4	2	*at the university of*
		4	2	*from the department of*
		4	2	*in the development of*
		3	3	*as a consequence of*
		3	2	*as part of the*
		3	2	*in the case of*
		3	2	*of different types of*
		3	2	*through the use of*
2	Other prepositional phrase (fragment)			
		18	3	*in the United States*
		15	3	*on the other hand*
		4	4	*for a long time*
		4	2	*to the fact that*
		3	3	*in countries such as*
		3	2	*in relation to the*
		3	2	*in which it is*
		3	3	*of his or her*
		3	2	*on a regular basis*
		3	2	*on the one hand*
(3) VP-based				
1	Anticipatory *it* + verb phrase/ adjective phrase			
		10	4	*it is important to*
		5	2	*it is possible that*
		5	2	*it is quite possible*
		4	2	*it is interesting that*
		3	3	*it is difficult to*
		3	3	*it is possible to*
2	Passive verb + prepositional phrase fragment			
		3	2	*be done by a*
		3	2	*can be used in*
3	Copula *be* + noun or adjective phrase			
		5	2	*is not considered a*
		3	3	*be aware of the*
4	(Verb phrase +) *that*-clause fragment			
		4	2	*have shown that the*

5	*That*-clause fragment			
		3	3	*that he or she*
		3	3	*that it is a*
6	(Verb/adjective +) *to*-clause fragment Predicative adjective + *to*-clause			
		5	2	*are more likely to*
		4	2	*is important to note*
7	(Passive) verb phrase + *to*-clause fragment			
		4	3	*can be used to*
8	*To*-clause fragment			
		5	2	*to know each other*
		4	2	*to be one of*
		3	2	*to be the most*
9	Pronoun/noun phrase + *be* (+…)			
		4	2	*there are a number*
		3	2	*this has been the*
		3	2	*this is not necessarily*
10	Pronoun phrase + verb			
		3	2	*he was unable to*
		3	2	*you may need to*
11	Adverbial clause fragment by the subordinator phrase (in order to)			
		3	2	*in order to avoid*
		3	2	*in order to learn*
		3	2	*in order to test*
(4)	Other expressions			
		6	2	*and the United States*
		4	2	*as opposed to only*
		3	3	*all over the world*
		3	2	*by no means the*
		3	2	*considered a major issue*
		3	3	*few and far between*
		3	2	*is more of a*
		3	2	*may or may not*

5.2.1 VP- and PP-based structures

Figure 5.2 shows that in terms of N-gram types, the VP- and PP-based forms are the most common structures, accounting for 34% and 33% of instructors' materials bundles, respectively.

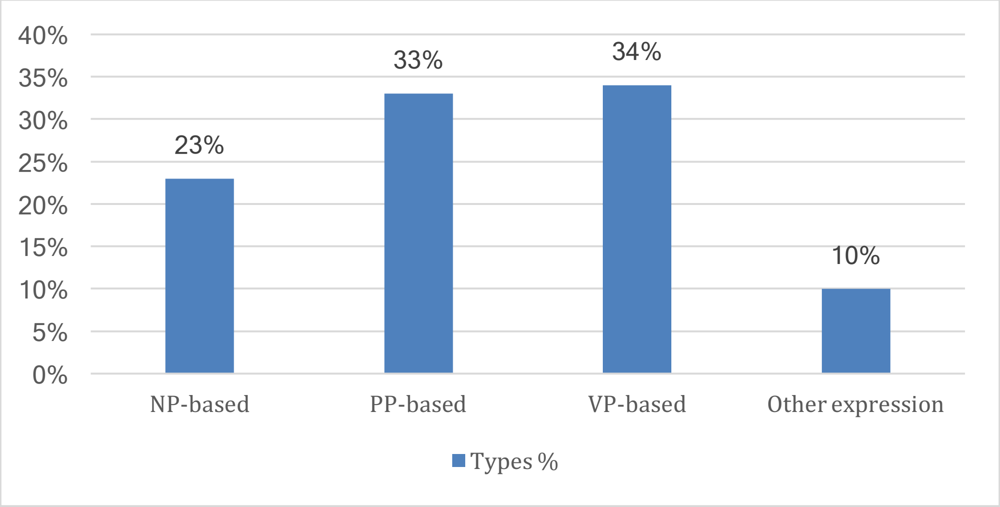

Figure 5.2. Structural distribution of the four-word bundles in the instructors' materials bundles list (types)

The percentages of the underlying sub-categories of each of the three major structures are displayed in Table 5.3. Looking more closely at the sub-categories of PP- and VP-based structural groups, further percentage differences can be identified. For example, while the majority of PP-based bundles are within the category *prepositional phrase with embedded of-phrase fragment*, accounting for 20% of bundle types, and *other prepositional phrase (fragment)*, accounting for 13% of bundle types, the VP-based bundles have more varied sub-categories but with low percentage occurrences, accounting for as little as 1% of bundle types. VP-based bundles have 11 sub-categories, yet there are limited bundle types within each verb structure; this excludes *anticipatory it + verb phrase/ adjective phrase*, which is the most frequently occurring within the verb category, totalling 7.5% of bundle types (see Table 5.3).

Table 5.3. Structural distribution of the 79 most common bundles in the instructors' materials bundles list (types)

Structure	Types	%
NP-BASED		
1-Noun phrase with *of*-phrase fragment	14	18%
2-Noun phrase with other post-modifier fragments	4	5%
PP-BASED		
1-Prepositional phrase with embedded *of*-phrase fragment	16	20%
2-Other prepositional phrase (fragment)	10	13%
VP-BASED		
1-Anticipatory *it* + verb phrase/ adjective phrase	6	7.5%
2-Passive verb + prepositional phrase fragment	2	2.5%
3-Copula *be* + noun or adjective phrase	2	2.5%
4-(Verb phrase +) *that*-clause fragment	1	1%
5-*That*-clause fragment	2	2.5%

6-(Verb/adjective +) *to*-clause fragment		
Predicative adjective + *to*-clause	2	2.5%
7-(Passive) verb phrase + *to*-clause fragment	1	1%
8-*To*-clause fragment	3	4%
9-Pronoun/noun phrase + *be* (+…)	3	4%
10-Pronoun phrase + verb	2	2.5%
11-Adverbial clause fragment by the subordinator	3	4%
OTHER EXPRESSIONS	8	10%
Totals	79	100%

Regarding the PP-based structures, the bundles *in the United States* (18 occurrences in 3 texts) and *on the other hand* (15 occurrences in 3 texts) are examples found within the sub-category *other prepositional phrase (fragment)*, and are also among the top-ranked most frequently occurring four-word bundles (see Tables 5.1 and 5.2). The sub-category of the PP-based *prepositional phrase with embedded of-phrase fragment* similarly comprises bundles from the top-ranked most frequently occurring bundles such as *at the end of* (15 occurrences in 3 texts), *as a result of* and *in the field of* (both occurring 8 times in 4 and 2 texts, respectively), and *with the help of* (7 occurrences in 2 texts). Although, overall, *prepositional phrase with embedded of-phrase fragment* and *other prepositional phrase (fragment)* account for the widest variety of lexical bundle structures among all other sub-categories in the instructors' materials bundles list (see Table 5.3), one structure contains more bundles than the other.

Table 5.3 shows the two structures of PP-based bundles with corresponding percentages of bundle types, revealing that the sub-category *prepositional phrase with embedded of-phrase fragment* includes a larger proportion of bundle types (20%) than *other prepositional phrase (fragment)* (accounting for 13% of bundle types). As shown in Figure 5.3, the structural sub-category *prepositional phrase with embedded of-phrase fragment* accounts for a larger proportion of bundle types than *other prepositional phrase (fragment)*, and all the other sub-categories.

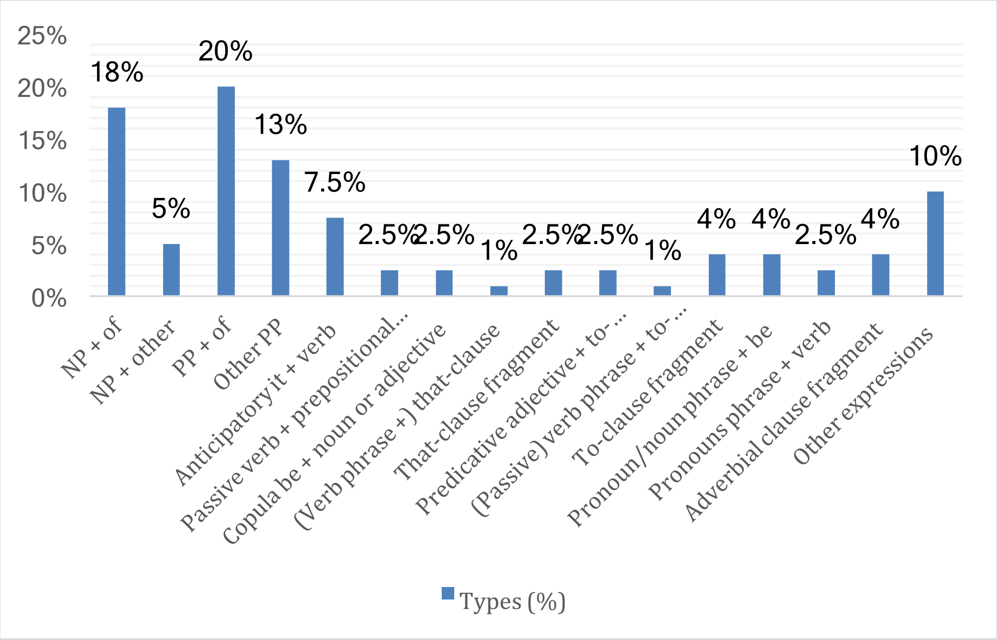

Figure 5.3. Distribution of the structural sub-categories in the instructors' materials bundles list (types)

VP-based structures incorporate 11 sub-categories. Figure 5.3 shows that the most notable sub-category in the VP-based structure is *anticipatory it + verb phrase/adjective phrase*, accounting for 7.5% of bundles, as shown in Table 5.3. This sub-category accounts for the highest number of bundles (e.g. *it is important to*) in this structure, with a frequency of 10 occurrences in four texts; it is also among the most frequent bundles (see Table 5.1). The VP-based sub-category also includes bundles with frequencies of five occurrences such as *it is possible that* and *it is quite possible*, each appearing across two texts (see Table 5.2).

The other sub-categories in the VP-based pattern are *to-clause fragment*, *pronoun/noun phrase + be (+…)*, and *adverbial clause fragment by the subordinator phrase*, which have similar percentages in terms of bundle types (4%), as shown in Table 5.3. The bundles include *to know each other* (5 occurrences in 2 texts), *there are a number* (4 occurrences in 2 texts), and *in order to avoid* (3 occurrences in 2 texts), respectively. Although VP-based structures are considered one of the major groupings in the instructors' corpus, they mostly include bundles with low frequencies, except for the *anticipatory it* bundle *(it is important to)* (see Table 5.2). In contrast, Figure 5.3 reveals that *(verb phrase +) that-clause fragment* and *(passive) verb phrase + to-clause fragment* are the least common sub-structural categories. An example is the bundle *can be used to* (4 occurrences in 3 texts) in the *(passive) verb phrase + to-clause,* structure, which accounts for only 1% of bundle types, as shown in Tables 5.2 and 5.3.

5.2.2 NP-based structures

Figure 5.1 shows that NP-based bundles form the third main structural category in the instructors' materials bundles list, comprising an overall proportion of 23% of bundle types. In terms of NP-based sub-categories, noun structures with *noun phrase with of-phrase fragment* account for 18% of bundle types. However, the bundles in the *noun phrase with other post-modifier fragments* sub-category only account for 5% of bundle types (see Table 5.3). The sub-category noun phrase with *of-phrase fragment* includes a variety of examples with different frequencies and ranges (e.g. *the end of the*, *title of the article*, *a great deal of, the beginning of the,* and *the purpose of*). Also, bundles (such as *the extent to which*, *the best way to*, and *the fact that the*) are included in *noun phrase with other post-modifier fragments*, as presented in Table 5.2. The bundles *the end of the* (12 occurrences in 4 texts) and *the extent to which* (9 occurrences in 2 texts) are among the top-ranked four-word bundles, with higher frequencies, as shown in Table 5.1. Additionally, Figure 5.3 shows that *noun phrase with of-phrase fragment* is the second-ranked sub-category after *the prepositional phrase with embedded of-phrase.* This shows that bundles in the instructors' materials greatly rely on bundles incorporating phrases with an *of-phrase* fragment, which will be examined in detail in the discussion chapter.

5.2.3 Other expressions

Other expressions include the lowest overall proportion of bundles compared to PP-, NP-, and VP-based structures, accounting for 10% of bundle types in the instructors' materials bundles list (see Figure 5.1). The sub-category other expressions features eight different types of bundles that tend not to fit into other categories. For example, the bundle *all over the world* (3 occurrences in 3 texts) is an *adverbial phrase without –of fragment*, while the bundle *and the United States* (6 occurrences in 2 texts) is *a conjunction fragment*. There are also bundles such as *as opposed to only* (4 occurrences in 2 texts), and *by no means the* (3 occurrences in 2 texts) which show the diversity of the forms within this sub-category, with most bundles having a lower frequency of three occurrences (see Table 5.2).

5.3 Functional classification of four-word bundles in the instructors' materials bundles list

The 79 most frequently occurring four-word bundles in the instructors' materials bundles list were classified according to an adapted version of Simpson-Vlach and Ellis's (2010) functional taxonomy, as discussed in the procedure chapter (see Chapter 4, section 4.5.2). Table 5.4 below displays the functional classification and sub-functional categories of the 79 most frequently occurring four-word bundles with their respective frequencies and range information. The applied taxonomy provided a basis upon which to categorize the lexical bundles according to their typical meaning and uses in context and to examine how these bundles are used in the EAP materials for teaching academic writing, thereby gaining specific insights into teachers' pedagogical materials. In order to avoid multifunctionality difficulty issues that may arise (see Chapter 2, section 2.2.3.2), one functional category was assigned for potential problematic bundles. This is one aspect that will be discussed in the limitations section (see Chapter 11, section 11.1.1).

Table 5.4. Functional categorization of the four-word bundles in the instructors' materials bundles list

Categories			
	Frequency	Range	Lexical bundles
1 Referential expressions			
	(1) Specification of attributes		
	(a) Intangible framing attributes		
	9	2	*the extent to which*
	8	2	*in the field of*
	7	2	*with the help of*
	6	2	*title of the article*
	5	2	*form of the verb*
	4	2	*in the development of*
	4	2	*to the fact that*
	3	2	*in relation to the*
	3	2	*in the case of*
	3	2	*in which it is*
	3	2	*on a regular basis*
	3	2	*the form of the*
	3	2	*the name of the*
	3	2	*through the use of*
	3	2	*the passive form of*
	(b) Tangible framing attributes		
	3	2	*as part of the*
	(c) Quantity specification		
	5	3	*a great deal of*

	4	3	*as one of the*
	4	2	*there are a number*
	3	2	*a third of the*
	3	3	*few and far between*
	3	2	*part of a larger*
	(2) Identification and focus		
	4	2	*to be one of*
	3	2	*of different types of*
	3	3	*of his or her*
	3	3	*that he or she*
	3	3	*that it is a*
	3	2	*this has been the*
	(3) Contrast and comparison		
	15	3	*on the other hand*
	4	2	*as opposed to only*
	4	3	*the relationship between the*
	3	2	*on the one hand*
	(4) Deictics and locatives		
	18	3	*in the United States*
	15	3	*at the end of*
	12	4	*the end of the*
	6	2	*and the United States*
	6	2	*at the beginning of*
	6	2	*in the middle of*
	5	2	*the beginning of the*
	4	3	*at the time of*
	4	2	*at the university of*
	4	4	*for a long time*
	4	2	*from the department of*
	3	3	*all over the world*
	3	3	*in countries such as*
2 Stance expressions			
	(1) Hedges		
	5	2	*are more likely to*
	3	2	*is more of a*
	3	2	*may or may not*
	(2) Epistemic stance		
	5	2	*is not considered a*
	4	2	*have shown that the*
	3	3	*be aware of the*
	3	2	*be done by a*
	3	2	*by no means the*
	3	2	*considered a major issue*
	(3) Obligation and directive		

	3	2	*you may need to*
	(4) Expressions of ability and possibility		
	5	2	*it is possible that*
	5	2	*it is quite possible*
	4	3	*can be used to*
	3	2	*can be used in*
	3	2	*he was unable to*
	3	3	*it is possible to*
	(5) Evaluation		
	10	4	*it is important to*
	5	2	*the best way to*
	4	2	*is important to note*
	4	2	*it is interesting that*
	3	3	*it is difficult to*
	3	2	*lack of understanding of*
	3	2	*this is not necessarily*
	3	2	*to be the most*
3 Discourse organizing functions			
	(1) Topic elaboration: cause and effect		
	8	4	*as a result of*
	5	2	*the results of the*
	5	2	*to know each other*
	4	2	*the purpose of this*
	4	2	*to the fact that*
	3	3	*as a consequence of*
	3	2	*in order to avoid*
	3	2	*in order to learn*
	3	2	*in order to test*
	3	2	*purposes of this study*

5.3.1 Referential expressions

As can be seen from the results presented in Figure 5.4, of the three main functional categories, *referential expression* bundles are the most frequently occurring, accounting for 57% of bundle types. *Referential expressions* account for more than half of the 79 most frequently occurring four-word bundles in terms of type, making them the most widely used of the three main functional classifications in the instructors' materials bundles list.

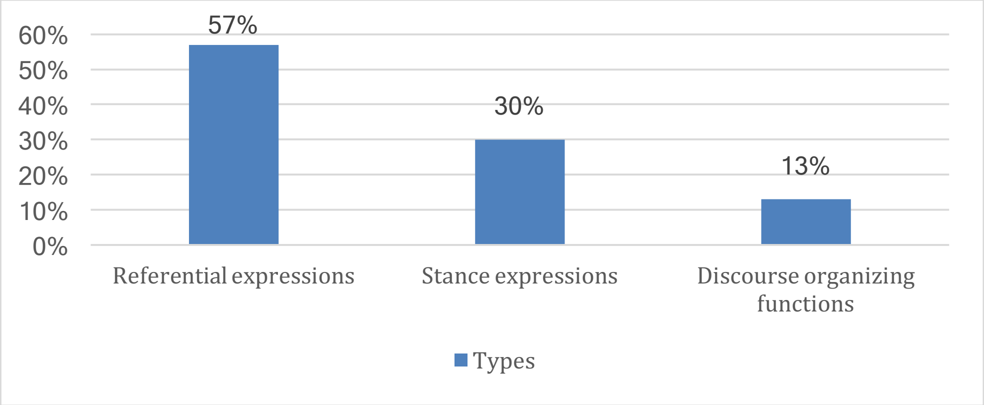

Figure 5.4. Functional distribution of the four-word bundles in the instructors' materials bundles list (types)

By looking more closely at the sub-functions in the *referential expressions* category, it can be observed that *(specification of attributes) intangible framing attributes* and *deictics and locatives* account for the highest proportion of bundles within this main group. For example, as shown in Table 5.4, the sub-function *intangible framing attributes (specification of attributes)* contains bundles such as *the extent to which* (9 occurrences in 2 texts), *in the field of* (8 occurrences in 2 texts), and *with the help of* (7 occurrences in 2 texts). This sub-category contains 19% of the overall functional distribution of bundle types in the instructors' materials bundles list (see Table 5.5). Table 5.5 provides a detailed classification of the discourse functions, showing the distribution of sub-functions in terms of the total number of bundle types.

Table 5.5. Functional distribution of the 79 most common bundles in the instructors' materials bundles list (types)

Functions	Types	%
REFERENTIAL EXPRESSIONS		
Intangible framing attributes (Specification of attributes)	15	19%
Tangible framing attributes	1	1%
Quantity specification	6	8%
Identification and focus	6	8%
Contrast and comparison	4	5%
Deictics and locatives	13	16%
STANCE EXPRESSIONS		
Hedges	3	3%
Epistemic stance	6	8%
Obligation and directive	1	1%
Expressions of ability and possibility	6	8%
Evaluation	8	10%

DISCOURSE ORGANIZING FUNCTIONS		
Topic elaboration: cause and effect	10	13%
Totals	79	100%

As shown in Table 5.5 and in Figure 5.5, *deictics and locatives* is a sub-function, accounting for the second highest percentage of bundle types (16%) after *intangible framing attributes* in the *referential expressions* group and among all sub-functions. The bundles in the *deictics and locatives* sub-function include *in the United States* (18 occurrences in 3 texts), *at the end of* (15 occurrences in 3 texts), and *the end of the* (12 occurrences in 4 texts) (see Table 5.4). When observing Table 5.4, although *deictics and locatives* comprises fewer types of bundles than *intangible framing attributes*, it is associated with bundles with a high number of occurrences. This includes the three most frequent bundles found at the top of the instructors' materials bundles list: *in the United States*, *at the end of*, and *the end of the* (see Table 5.1). Besides the bundles already mentioned above in the *deictics and locatives* and *intangible framing attributes* sub-functions, the referential expressions group also includes bundles such as *on the other hand* (15 occurrences in 3 texts) under the sub-function *contrast and comparison*.

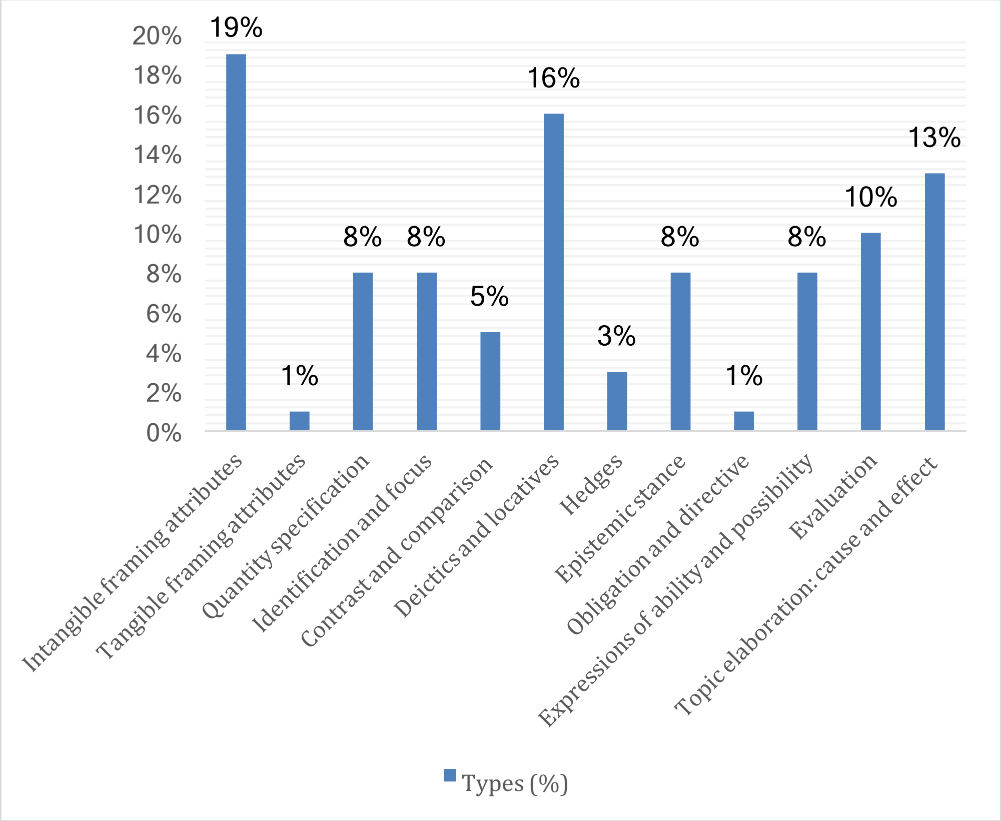

Figure 5.5. Distribution of the functional sub-categories in the instructors' materials bundles list (types)

5.3.2 Stance expressions

Figure 5.4 reveals that *stance expressions* constitute the second most frequent classification among the three primary functions served by lexical bundles, accounting for 30% of bundle types (depending on their meaning in context in the instructors' materials bundles list). All sub-functions of the instructors' materials bundles list, in this category (*hedges*, *epistemic stance*, *obligation and directive*, *expressions of ability and possibility*, and *evaluation)* account for different functional percentage totals of bundle types, ranging from as low as 1% to high as 10%, as shown in Table 5.5 and Figure 5.5. Bundles found in the *evaluation* sub-function account for the largest proportion (10%) of bundle types. For example, bundles such as *it is important to* (10 occurrences in 4 texts) and *the best way to* (5 occurrences in 2 texts) are included in this group.

Furthermore, *epistemic stance* bundles, account for 8% of bundle types (e.g. *is not considered a*, 5 occurrences in 2 texts; and *have shown that the*, 4 occurrences in 2 texts). Similarly, *expressions of ability and possibility* account for 8% of bundle types. This sub-category includes bundles such as *it is possible that* (5 occurrences in 2 texts), *it is quite possible* (5 occurrences in 2 texts), and *can be used to* (4 occurrences in 3 texts). Furthermore, the sub-functions *hedges* and *obligation and directive* account for the lowest percentage of bundle types (3% and 1%, respectively) (see Table 5.5 and Figure 5.5). Bundles such as *are more likely to* (5 occurrences in 2 texts) are included in *hedges*; while *obligation and directive* include one type of bundle, which is *you may need to* (3 occurrences in 2 texts), as shown in Table 5.4.

5.3.3 Discourse organizing functions

The last major functional classification, and the least frequently occurring of the three, is *discourse organizing functions*, accounting for 13% of the total bundle types, as displayed in Figure 5.4. In this category, there is one sub-function that serves as *topic elaboration: cause and effect*, which accounts for 13% of bundle types, as shown in Table 5.5 and Figure 5.5. In this sub-function, bundles included are *as a result of* (8 occurrences in 4 texts), *the results of the* (5 occurrences in 2 texts), *to know each other* (5 occurrences in 2 texts), *the purpose of this* (4 occurrences in 2 texts), and *to the fact that* (4 occurrences in 2 texts), as presented in Table 5.4.

5.4 Concluding remarks

This chapter has reported on the overall quantitative findings in relation to types, frequency counts and distributional accounts of the 79 most frequently occurring bundles identified in the instructors' corpus. It has shown that the bundles *in the United States*, *at the end of*, and *on the other hand* are ranked among the most frequently occurring bundles. VP- and PP-based bundles form the main structural classification of the instructors' materials bundles, including highly frequent bundles such as *in the United States* and *it is important to*. Furthermore, *referential expressions* constitute the major pragmatic function for the instructors' materials bundles, again including the bundle *in the United States.*

In the next chapter, I will examine the second corpus in this study, the textbooks corpus. The same structural and functional frameworks applied in this chapter will be used to explore the second corpus, allowing for comparable results to be reported in a systematic manner.

CHAPTER VI

QUANTITATIVE RESULTS: TEXTBOOKS CORPUS – THE LIST OF BUNDLES

This chapter will present the results of two research questions; research question 1 ("What are the most frequent four-word bundles profiled in textbooks and instructors' materials, in an EAP pre-sessional programme in the UK particularly aimed at teaching academic writing?") and 2 ("How are the most frequent four-word bundles classified functionally and structurally?"). This chapter is concerned with research questions 1 and 2 above, with regard to the *textbooks corpus*. I will present the quantitative results from the analysis of the most frequent four-word bundles found in the textbooks corpus. Similar to the presentation of the quantitative results of the *instructors' materials corpus* data in Chapter 5, three major aspects of the bundle analysis will be included. The overall distribution of frequency counts, relating to bundle types, and the structural and functional patterns of the four-word bundles, will be reported. Again, at this stage the empirical judgements are based on the raw frequency counts of actual occurrences of lexical bundles.

6.1 Quantitative results: types and frequency distribution

After removing excluded, overlapping, and context-based bundles, the final list consisted of 102 four-word bundle types in the 212,235-word textbooks corpus. There are 154,122 individual cases in the textbooks bundles list, accounting for 73% of the more than 200,000-word textbooks corpus. Table 6.1 displays the 102 most frequently occurring four-word bundles in the textbooks list as previously stated in Chapter 4, in order of frequency and with range information.

Table 6.1. Lexical bundles in the textbooks bundles list by order of frequency with range information

Total No. of N-Gram Types: 102			
Rank	Frequency	Range	Bundles
	38	3	*focus on your subject*
	34	4	*as a result of*
	33	4	*it is important to*
	32	3	*look at the following*
	31	4	*and answer the questions*

29	4	*the words in the*
29	3	*work in pairs and*
28	3	*essay with the title*
27	4	*at the end of*
27	4	*what do you think*
25	3	*from the text in*
25	3	*use a dictionary to*
23	3	*in the text in*
23	4	*in the United States*
23	3	*you are going to*
21	4	*the information in the*
20	3	*the words in bold*
19	4	*can you think of*
19	2	*the text on page*
18	4	*on the other hand*
17	4	*in a way that*
17	3	*of the text in*
17	3	*research shows that in*
17	3	*the subject of the*
17	4	*the ways in which*
17	2	*what you have read*
17	3	*you have been given*
16	3	*answer the following questions*
16	2	*answer the questions about*
16	4	*at the beginning of*
16	3	*from the text on*
16	4	*the end of the*
16	3	*the following extracts from*
15	3	*common in academic writing*
15	3	*have been asked to*
15	4	*in your own words*
15	4	*make a note of*
15	3	*the meaning of the*
14	4	*a wide range of*
14	4	*do you think the*
14	3	*in bold in the*
14	2	*to check your answers*
14	3	*why do you think*
13	3	*at the same time*
13	2	*complete the following sentences*
13	3	*decide which of the*
13	4	*do you think are*
13	3	*look again at the*
13	4	*make notes on the*
13	4	*the way in which*
13	2	*the way we do*
13	3	*the written academic corpus*
12	3	*are more likely to*
12	3	*it is possible to*
12	4	*one of the most*
12	3	*read the text again*
12	4	*the beginning of the*
12	4	*the rest of the*
12	4	*what is the difference*
12	2	*with a similar meaning*

11	3	*a similar meaning to*
11	2	*complete the sentences using*
11	2	*inferring the meaning of*
11	4	*look back at the*
11	3	*of the words in*
11	2	*the following extract from*
11	3	*the relationship between the*
11	3	*used in the text*
11	2	*which of the following*
10	3	*as part of a*
10	3	*be more than one*
10	2	*check your answers in*
10	3	*discuss the following questions*
10	4	*do you agree with*
10	2	*each of the following*
10	2	*in the box to*
10	3	*in the form of*
10	4	*is the most important*
10	3	*research shows that the*
10	2	*scientists and their work*
10	2	*that something is true*
10	4	*the first paragraph of*
10	3	*the first part of*
10	3	*the phrases in bold*
10	4	*the title of the*
10	2	*used to refer to*
10	3	*you think are the*
9	4	*a great deal of*
9	3	*a large number of*
9	3	*a piece of writing*
9	2	*article in a journal*
9	3	*be followed by a*
9	3	*can be used to*
9	2	*from the same family*
9	2	*in the correct order*
9	2	*it is a good*
9	3	*the advantages and disadvantages*
9	3	*the correct form of*
9	4	*the use of computers*
9	2	*the verbs in the*
9	4	*to write an essay*
9	3	*what you already know*

Having examined the list in Table 6.1, I realized that the 102 four-word bundles can be arranged into three groups based on their occurrences. Table 6.1 shows that the bundles *focus on your subject* (38 occurrences in 3 textbooks), *as a result of* (34 occurrences in 4 textbooks), *it is important to* (33 occurrences in 4 textbooks), *look at the following* (32 occurrences in 3 textbooks), *and answer the questions* (31 occurrences in 4 textbooks) have the top-ranked frequencies in the textbooks bundles list, with 30 occurrences and above. The second highest ranked four-word bundles include *the words in the* and *work in pairs and* (29 occurrences each in 4 and 3 textbooks, respectively), *essay with the title* (28 occurrences in 3 textbooks), *at the end of* and *what do you think* (27 occurrences each, both in 4 textbooks), *from the text in* and *use a dictionary to* (25 occurrences each, both in 3 textbooks), *in the text in*, *in the United States*, and *you are going to* (23 occurrences each, in 3, 4, and 3 textbooks, respectively), *the information in the* (21 occurrences in 4 textbooks), and finally *the words in bold* (20 occurrences in 3 textbooks); all have frequencies in the 20s and above, as shown in Table 6.1.

The third group of four-word bundles have frequencies ranging between 19 and 9 occurrences in the textbook list, as shown in Table 6.1. This group includes bundles such as *can you think of* (19 occurrences in 4 textbooks), *on the other hand* (18 occurrences in 4 textbooks), *in a way that* (17 occurrences in 4 textbooks), *answer the following question* (16 occurrences in 3 textbooks), *common in academic writing* (15 occurrences in 3 textbooks), *a wide range of* (14 occurrences in 4 textbooks), *at the same time* (13 occurrences in 3 textbooks), *are more likely to* (12 occurrences in 3 textbooks), *a similar meaning to* (11 occurrences in 3 textbooks), *as part of a* (10 occurrences in 3 textbooks), *and a great deal of* (9 occurrences in 4 textbooks).

In Table 6.1 above, it can be seen that although the top-ranked four-word bundles have the highest frequencies (in the 30s and 20s), they also have the least variation in bundle types, including one or two types for each frequency rate. For example, there are five types of bundles with frequencies in the 30s, while there are 12 bundle types with frequencies in the 20s. In contrast, four-word bundles with low frequency levels (such as 13, 11, 10, and 9 occurrences) display more variation in bundle types. For example, as shown in Figure 6.1, four-word bundles with low frequencies of 10 and 9 occurrences have different types, accounting for 18% and 15% of all bundle types, respectively. In contrast, the bundles in the top-ranked four-word bundles (with occurrences of 38, 34, 33, 32, and 31) have a range value of 1% each with regard to bundle types. Figure 6.1 displays the percentages of all frequencies of bundle types in the textbooks bundles list.

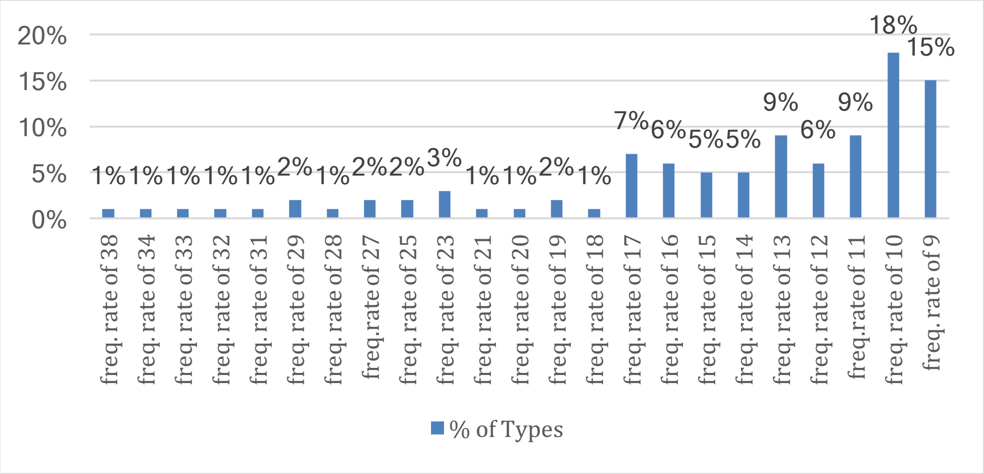

Figure 6.1. Overall distribution of the 102 bundles in the textbooks bundles list (types)

6.2 Structural classification of the four-word bundles in the textbooks bundles list

The 102 four-word bundles in the textbooks bundles list were structurally classified using the same framework used for the instructors' materials bundles list. Therefore, the 102 four-word bundles in the textbooks bundles list were categorized structurally using the same taxonomy used by Biber *et al.* (1999: 1014–1024), in terms of their grammatical types. In addition, the structural characteristics of those bundles follow those used for the bundles identified in previous studies of academic prose (Biber and Conrad, 1999; Biber *et al.*, 1999; Cortes, 2004; Simpson-Vlach and Ellis, 2010), as previously stated. The 102 four-word bundles identified in the textbooks corpus also fall into various structural sub-categories under four main grammatical patterns: Prepositional Phrases (PP-based, e.g. *as a result of*), (NP-based, e.g. *the subject of the*), Verb Phrases (VP-based, e.g. *it is important to*), and Other Expressions (such as *common in academic writing*), as shown in Table 6.2. As mentioned in the procedure chapter (see Chapter 4, section 4.5.3), further sub-categorizations were added, which are associated with the types of four-word bundles found in the textbooks bundles list. Table 6.2 displays the 102 four-word bundles with their structural correlates in the textbooks bundles list.

Table 6.2. Structural patterns of the four-word bundles in the textbooks bundles list

Categories	No.	Frequency	Range	Lexical bundles
(1) NP-based				
1	Noun phrase with *of*-phrase fragment			
		17	3	*the subject of the*
		16	4	*the end of the*
		15	3	*the meaning of the*
		14	4	*a wide range of*
		12	4	*one of the most*
		12	4	*the beginning of the*
		12	4	*the rest of the*
		10	4	*the first paragraph of*
		10	3	*the first part of*
		10	4	*the title of the*
		9	4	*a great deal of*
		9	3	*a large number of*
		9	3	*a piece of writing*
		9	3	*the correct form of*
		9	4	*the use of computers*
2	Noun phrase with other post-modifier fragments			
		29	4	*the words in the*
		28	3	*essay with a title*
		21	4	*the information in the*
		20	3	*the words in bold*
		19	2	*the text on page*
		17	4	*the ways in which*
		16	3	*the following extracts from*
		13	4	*the way in which*
		11	3	*a similar meaning to*
		11	2	*the following extract from*
		11	3	*the relationship between the*
		10	3	*the phrases in bold*
		9	2	*article in a journal*
		9	2	*the verbs in the*
3	Other noun phrase expressions			
		9	3	*the advantages and disadvantages*
(2) PP-based				
1	Prepositional phrase with embedded *of*-phrase fragment			
		34	4	*as a result of*
		27	4	*at the end of*
		16	4	*at the beginning of*
		10	3	*as part of a*
		10	3	*in the form of*
2	Other prepositional phrase (fragment)			
		25	3	*from the text in*
		23	3	*in the text in*
		23	4	*in the United States*
		18	4	*on the other hand*
		17	4	*in a way that*

		17	3	*of the text in*
		16	3	*from the text on*
		15	4	*in your own words*
		14	3	*in bold in the*
		13	3	*at the same time*
		12	2	*with a similar meaning*
		11	3	*of the words in*
		10	2	*in the box to*
		9	2	*from the same family*
		9	2	*in the correct order*
(3) VP-based				
1	Anticipatory *it* + verb phrase/ adjective phrase			
		33	4	*it is important to*
		12	3	*it is possible to*
		9	2	*it is a good*
2	Passive verb + prepositional phrase fragment			
		11	3	*used in the text*
		9	3	*be followed by a*
3	Copula *be* + noun or adjective phrase			
		10	4	*is the most important*
4	(Verb phrase +) *that*-clause fragment *That*-clause fragment			
		10	2	*that something is true*
5	Noun + verb phrase + *that*-clause			
		17	3	*research shows that in*
		10	3	*research shows that the*
6	(Verb/adjective +) *to*-clause fragment Predicative adjective + *to*-clause			
		12	3	*are more likely to*
7	(Passive) verb phrase + to-clause fragment			
		15	3	*have been asked to*
		10	2	*used to refer to*
		9	3	*can be used to*
8	*To*-clause fragment			
		14	2	*to check your answers*
		9	4	*to write an essay*
9	Verb phrase + determiner phrase fragment			
		38	3	*focus on your subject*
		10	2	*check your answers in*
10	Verb phrase + prepositional phrase fragment			
		29	3	*work in pairs and*
		13	3	*look again at the*
		13	4	*make notes on the*
		11	4	*look back at the*
11	Verb phrase + noun phrase fragment			
		32	3	*look at the following*
		31	4	*and answer the questions*
		25	3	*use a dictionary to*
		16	3	*answer the following questions*
		16	2	*answer the questions*

				about
		15	4	*make a note of*
		13	2	*complete the following sentences*
		13	3	*decide which of the*
		12	3	*read the text again*
		11	2	*complete the sentences using*
		10	3	*discuss the following questions*
12	WH-questions phrases			
		27	4	*what do you think*
		17	2	*what you have read*
		14	3	*why do you think*
		12	4	*what is the difference*
		11	2	*which of the following*
		9	3	*what you already know*
13	*Yes-no* question fragments			
		14	4	*do you think the*
		13	4	*do you think are*
		10	4	*do you agree with*
14	2nd person pronoun you+VP fragment			
		23	3	*you are going to*
		17	3	*you have been given*
		10	3	*you think are the*
15	Modal verb question fragments			
		19	4	*can you think of*
(4) Other expressions				
		15	3	*common in academic writing*
		13	2	*the way we do*
		13	3	*the written academic corpus*
		11	2	*inferring the meaning of*
		10	3	*be more than one*
		10	2	*each of the following*
		10	2	*scientists and their work*

6.2.1 VP-based structures

The most notable structure emerging from Figure 6.2 is the verb-based structure, which includes the majority of four-word bundle types in the textbooks bundles list, accounting for 44% of bundle types. This means that the textbooks bundles list incorporates a larger proportion of verb phrase structures compared to the NP-based, PP-based, and other expression types.

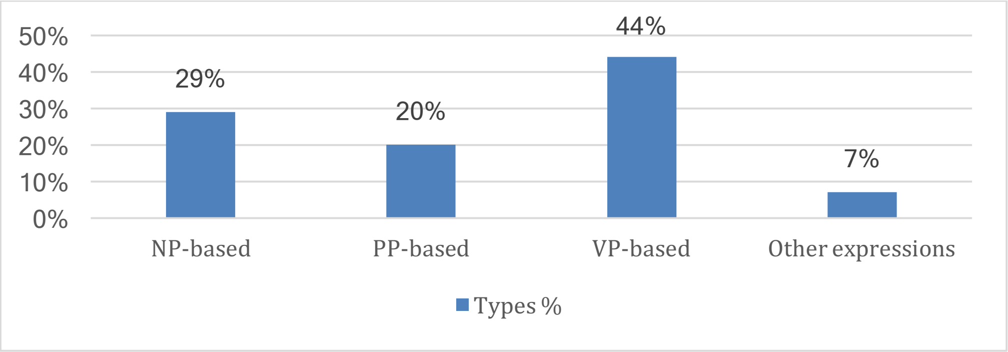

Figure 6.2. Structural distribution of the four-word bundles in the textbooks bundles list (types)

In the present study, under the VP-based structural category, I identified 15 sub-categories that incorporate *verb phrase bundles*, as shown in Table 6.2. Having examined Table 6.3 below, it can be observed that most of the sub-categories have percentages ranging from 1% to 5% with regard to bundle types. For example, the sub-category structures *anticipatory it + verb phrase/adjective phrase*, *(passive) verb phrase + to-clause fragment*, *yes-no questions fragments*, *and 2nd person pronoun you + VP fragment* have the same percentage with regard to bundle types (3%).

The *anticipatory it + verb phrase/adjective phrase* sub-category includes the bundle *it is important to* (33 occurrences in 4 textbooks), while the sub-category *yes-no questions fragments* contains the bundle do *you think the* (14 occurrences in 4 textbooks), comprising a limited number of types. Also, the form *verb phrase + prepositional phrase fragment* accounts for 4% of bundle types, including the bundle *work in pairs and* (29 occurrences in 3 textbooks); also, the form *WH-questions phrases includes* accounts for 5% of bundle types, including the bundle *what do you think* (27 occurrences in 4 textbooks). These sub-categories include just a few types of bundles with high frequencies, as shown in Table 6.1 and 6.2.

Table 6.3. Structural distribution of the 102 most common bundles in the textbooks bundles list (types)

Structure	Types	%
NP-BASED		
1-Noun phrase with *of*-phrase fragment	15	15%
2-Noun phrase with other post-modifier fragments	14	14%
3-Other noun phrase expressions	1	1%
PP-BASED		
1-Prepositional phrase with embedded *of*-phrase fragment	5	5%
2-Other prepositional phrases (fragment)	15	15%
VP-BASED		
1-Anticipatory *it* + verb phrase/adjective phrase	3	3%
2-Passive verb + prepositional phrase fragment	2	2%
3-Copula *be* + noun or adjective phrase	1	1%
4-(Verb phrase +) *that*-clause fragment *That*-clause fragment	1	1%
5-Noun + verb phrase + *that*-clause	2	2%
6-(Verb/adjective +) *to*-clause fragment Predicative adjective + *to*-clause	1	1%
7-(Passive) verb phrase + *to*-clause fragment	3	3%
8-*To*-clause fragment	2	2%
9-Verb phrase + determiner phrase fragment	2	2%
10-Verb phrase + prepositional phrase fragment	4	4%
11-Verb phrase + noun phrase fragment	11	11%
12-WH-questions phrases	6	5%
13-*Yes-no* questions fragments	3	3%
14-2nd person pronoun you+VP fragment	3	3%
15-Modal verb question fragments	1	1%
OTHER EXPRESSIONS	7	6%
Totals	102	100%

Figure 6.3 shows that the sub-category *verb phrase + noun phrase fragment* has the highest percentage of bundle types (11%); see also Table 6.3. This sub-category includes more bundle types with high frequencies such as *look at the following* (32 occurrences in 3 textbooks), *and answer the questions* (31 occurrences in 4 textbooks), and *use a dictionary to* (25 occurrences in 3 textbooks) (see Table 6.2).

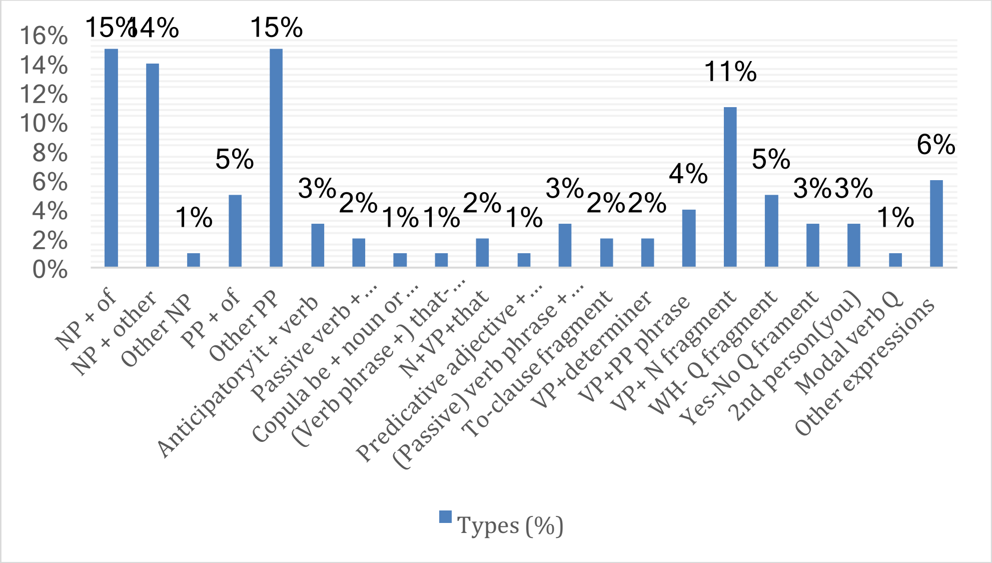

Figure 6.3. Distribution of the structural sub-categories in the textbooks bundles list (types)

6.2.2 NP- and PP-based structures

Figure 6.2 revealed that the second most frequent structural group is NP-based structures, accounting for 29% of bundle types. In this sub-category, the majority of bundles are found within the groups *noun phrase with of-phrase fragment* (15% of bundle types) and *noun phrase with other post-modifier fragments* (14% of bundle types), as shown in Table 6.2, 6.3 and Figure 6.3. For example, the form *noun phrase with of-phrase fragment* includes bundles such as *the subject of the* (17 occurrences in 3 textbooks), *the end of the* (16 occurrences in 4 textbooks), and *the meaning of the* (15 occurrences in 3 textbooks), as shown in Table 6.2. Furthermore, the sub-category *noun phrase with other post-modifier fragments* includes bundles with higher frequencies such as *the words in the* (29 occurrences in 4 textbooks), *essay with a title* (28 occurrences in 3 textbooks), *the information in the* (21 occurrences in 4 textbooks), and *the words in bold* (20 occurrences in 3 textbooks), as shown in Table 6.2. In addition, these two sub-categories have more varied types than the third form, *other noun phrase expressions*, which accounts for only 1% of bundle types, as displayed in Table 6.3 and Figure 6.3, including the bundle *the advantage and disadvantage* (9 occurrences in 3 textbooks) (see Table 6.2).

Besides the NP-based structures, PP-based structures account for a lower percentage of bundle types (20% of bundle types), as shown in Figure 6.2. As can be observed from Table 6.3, the sub-category *other prepositional phrases* *(fragment)* accounts for a higher percentage of bundle types (15% of bundle types) than the form *prepositional phrase with embedded of-phrase fragment*, which only accounts for 5% of bundle types. Table 6.2 shows that bundles such as *from the text in* (25 occurrences in 3 textbooks), *in the text in* (23 occurrences in 3 textbooks), and *in the United States* (23 occurrences in 4 textbooks) within the sub-category *other prepositional phrases (fragment)* have high frequencies and more diverse lexical bundles. However, bundles within the sub-category *prepositional phrase with embedded of-phrase fragment* such as *as a result of* (34 occurrences in 4 textbooks), and *at the end of the* (27 occurrences in 4 textbooks) include bundles with a high frequency rate, but contain a restricted number of bundle types.

6.2.3 Other expressions

With regard to other expressions, Figure 6.2 shows that it is the least frequent pattern, accounting for 7% of bundle types. This sub-category has the most limited number of bundle types and includes bundles with a lower frequency than those which are VP-, PP-, and NP-based. It includes bundles such as *common in academic writing* (15 occurrences in 3 textbooks), *the way we do* (13 occurrences in 2 textbooks), and *inferring the meaning of* (11 occurrences in 2 textbooks). These bundles found in this category do not fit into other structures.

6.3 Functional classification of four-word bundles in the textbooks bundles list

Using an adapted version of Simpson-Vlach and Ellis's (2010) functional taxonomy, I classified the 102 most frequent four-word bundles in the textbooks bundles list, as discussed in the procedure chapter (see Chapter 4, section 4.5.2), and similar to the functional classifications applied to bundles in the instructors' materials bundles list. Table 6.4 below shows the 102 most frequent four-word bundles with their respective frequencies and range information, together with additional sub-categories. The framework of the pragmatic functions is needed to interpret how the language of textbooks, especially lexical bundles, is used in the EAP context during writing classes.

Table 6.4. Functional categorization of the four-word bundles in the textbooks bundles list

Categories			
No.	Frequency	Range	Lexical bundles
1 Referential expressions			
(1) Specification of attributes			
(a) Intangible framing attributes			
	17	4	*in a way that*
	17	3	*the subject of the*
	17	4	*the ways in which*
	15	3	*the meaning of the*
	13	4	*the way in which*
	10	3	*in the form of*
	10	4	*the title of the*
	9	3	*the correct form of*
(b) Tangible framing attributes			
	12	4	*the rest of the*
	10	3	*as part of a*
(c) Quantity specification			
	14	4	*a wide range of*
	12	4	*one of the most*
	10	3	*be more than one*
	10	2	*each of the following*
	9	4	*a great deal of*
	9	3	*a large number of*
(2) Identification and focus			
	10	2	*used to refer to*
(3) Contrast and comparison			
	18	4	*on the other hand*
	12	3	*with a similar meaning*
	11	2	*a similar meaning to*
	12	4	*what is the difference*
	11	3	*the relationship between the*
	9	2	*from the same family*
	9	3	*the advantages and disadvantages*
(4) Deictics and locatives			
	28	3	*essay with the title*
	27	4	*at the end of*
	23	4	*in the United States*
	16	4	*at the beginning of*
	16	4	*the end of the*
	13	3	*the written academic corpus*
	12	4	*the beginning of the*
	10	2	*scientists and their work*
	9	3	*a piece of writing*

	9	2	*article in a journal*
	9	4	*the use of computers*
2 Stance expressions			
	(1) Hedges		
	12	3	*are more likely to*
	(2) Epistemic stance		
	27	4	*what do you think*
	14	4	*do you think the*
	14	3	*why do you think*
	13	3	*decide which of the*
	13	4	*do you think are*
	13	2	*the way we do*
	10	4	*do you agree with*
	10	3	*you think are the*
	9	3	*what you already know*
	(3) Obligation and directive		
	38	3	*focus on your subject*
	31	4	*and answer the questions*
	29	3	*work in pairs and*
	25	3	*use a dictionary to*
	23	3	*you are going to*
	16	3	*answer the following questions*
	16	2	*answer the questions about*
	15	3	*have been asked to*
	15	4	*in your own words*
	15	4	*make a note of*
	14	2	*to check your answers*
	13	2	*complete the following sentences*
	13	4	*make notes on the*
	12	3	*read the text again*
	11	2	*complete the sentences using*
	11	2	*inferring the meaning of*
	10	2	*check your answers in*
	10	3	*discuss the following questions*
	9	2	*in the correct order*
	9	4	*to write an essay*
	(4) Expressions of ability and possibility		
	19	4	*can you think of*
	12	3	*it is possible to*
	9	3	*can be used to*
	(5) Evaluation		
	33	4	*it is important to*
	15	3	*common in*

			academic writing
	10	4	*is the most important*
	10	2	*that something is true*
	9	2	*it is a good*
3 Discourse organizing functions			
(1) Metadiscourse/Textual reference			
	29	4	*the words in the*
	25	3	*from the text in*
	23	3	*in the text in*
	21	4	*the information in the*
	20	3	*the words in bold*
	19	2	*the text on page*
	17	3	*of the text in*
	17	2	*what you have read*
	16	3	*from the text on*
	16	3	*the following extracts from*
	14	3	*in bold in the*
	11	3	*of the words in*
	11	2	*the following extract from*
	11	3	*used in the text*
	11	2	*which of the following*
	10	2	*in the box to*
	10	4	*the first paragraph of*
	10	3	*the first part of*
	10	3	*the phrases in bold*
	9	2	*the verbs in the*
(2) Topic introduction and focus			
	32	3	*look at the following*
	13	3	*look again at the*
	17	3	*you have been given*
	11	4	*look back at the*
(3) Topic elaboration: **(a) Non-causal**			
	17	3	*research shows that in*
	10	3	*research shows that the*
	9	3	*be followed by a*
(b) Topic elaboration: cause and effect			
	34	4	*as a result of*
(4) Discourse markers			
	13	3	*at the same time*

6.3.1 Stance expressions

Figure 6.4 shows that the largest proportion of the four-word bundles in the textbooks bundles list function as stance expressions, which account for 37% of bundle types.

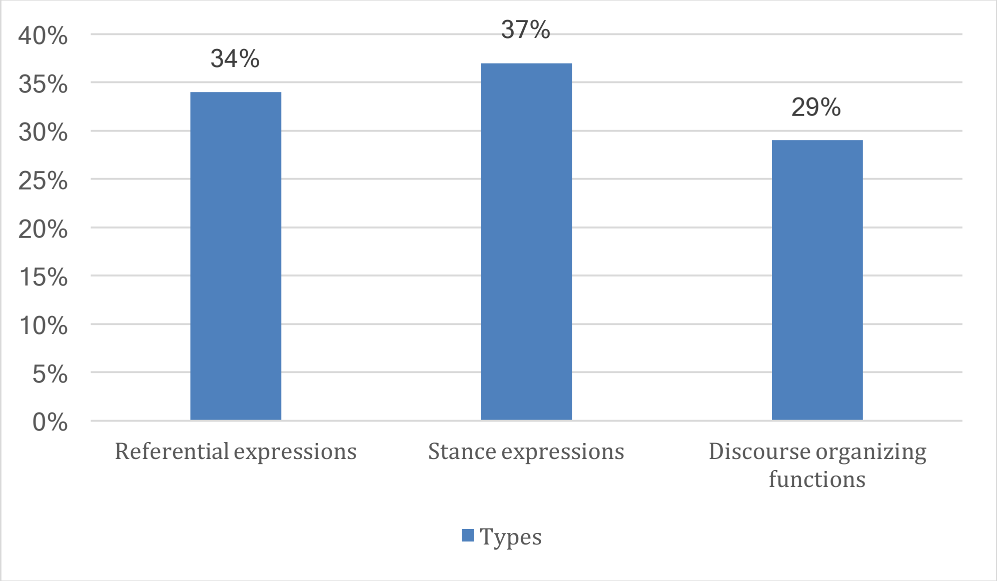

Figure 6.4. Functional distribution of the four-word bundles in the textbooks bundles list (types)

Within the stance expressions function, the *hedges* sub-category includes the lowest number of bundles, accounting for 1% of bundle types. However, the largest of the five sub-categories in the stance expressions function is *obligation and directive*, accounting for 20% of bundle types, as shown in Table 6.5 and Figure 6.5. This grouping includes bundles such as *focus on your subject* (38 occurrences in 3 texts), *and answer the questions* (31 occurrences in 4 texts), *work in pairs and* (29 occurrences in 3 texts), *use a dictionary to* (25 occurrences in 3 texts), and *you are going to* (23 occurrences in 3 texts) (see Table 6.4). As for *epistemic*, the second largest sub-category, it accounts for 8% of bundles, including bundles such as *what do you think* (27 occurrences in 4 texts), as displayed in Tables 6.4 and 6.5.

Table 6.5. Functional distribution of the 102 most common bundles in the textbooks bundles list (types)

Functions	Types	%
REFERENTIAL EXPRESSIONS		
(Specification of attributes) Intangible framing attributes	8	8%
Tangible framing attributes	2	2%
Quantity specification	6	6%
Identification and focus	1	1%
Contrast and comparison	7	7%
Deictics and locatives	11	10%
STANCE EXPRESSIONS		
Hedges	1	1%
Epistemic stance	9	8%
Obligation and directive	20	20%
Expressions of ability and possibility	3	3%
Evaluation	5	5%
DISCOURSE ORGANIZING FUNCTIONS		
Metadiscourse and textual reference	20	20%
Topic introduction and focus	4	4%
Topic elaboration: non-causal	3	3%
Topic elaboration: cause and effect	1	1%
Discourse markers	1	1%
Totals	102	100%

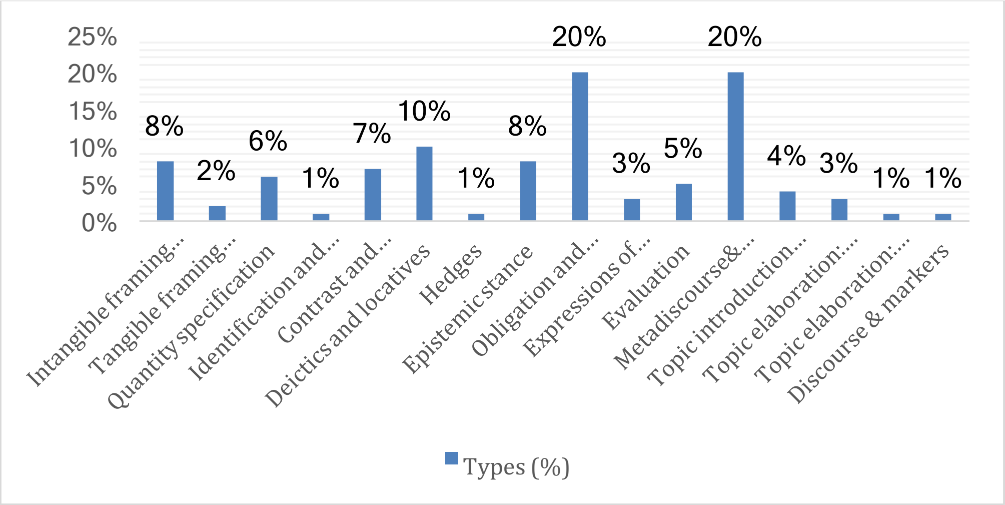

Figure 6.5. Distribution of functional sub-categories in the textbooks bundles list (types)

6.3.2 Referential expressions

Although stance expressions account for slightly higher percentages of bundle types than bundles with referential expression and a little more than discourse organizing functions, the three major pragmatic functions do not differ drastically, having closer distributions within the textbooks corpus, as shown in Figure 6.4. Referential expressions account for 34% of bundle types in the textbooks corpus, as shown in Figure 6.4 above.

Figure 6.5, with regard to referential expressions, shows that the *deictics and locatives* sub-category includes the largest number of four-word bundles, accounting for 10% of bundle types (see also Table 6.5). This sub-category includes high-frequency bundles such as *essay with a title* (28 occurrences in 3 texts), *at the end of the* (27 occurrences in 4 texts), and *in the United States* (23 occurrences in 4 texts) (see Table 7.4). Moreover, the lowest number of bundles can be found in the sub-categories *identification and focus* and *tangible framing attributes*, which accounts for only 1% and 2% of bundle types, respectively, as shown in Table 6.5 and Figure 6.5. *Intangible framing attributes, contrast and comparison,* and *quantity specification* have close percentages, accounting for 8%, 7%, and 6%, of bundle types, respectively (see Table 6.5). They include bundles such as *in a way that* (17 occurrences in 4 texts), functioning as *intangible framing attributes*. *Contrast and comparison bundles* consists of sequences such as *on the other hand* (18 occurrences in 4 texts) and *with a similar meaning* (13 occurrences in 3 texts), whereas *quantity specification* consists of bundles such as *a wide range of* (14 occurrences in 4 texts) and *one of the most* (12 occurrences in 4 texts), as shown in Table 6.4.

6.3.3 Discourse organizing functions

Comprising 29% of types, *discourse organizing functions* has the fewest bundle types of the three functions, as shown in Figure 6.4. *Metadiscourse and textual reference* is by far the most common sub-functional category in the textbooks bundles list, as shown in Table 6.5, and Figure 6.5, accounting for 20% of bundle types. This sub-functional category has similar percentages to the previously mentioned sub-functional category *obligation and directive* found within the stance expressions category (see Table 6.5 and Figure 6.5). Both sub-functional categories include 20 four-word bundles, which include some high-frequency sequences, marking the largest sub-functional groupings among all sub-functions. *The words in the* (29 occurrences in 4 texts), *from the text in* (25 occurrences in 3 texts), *in the text in* (23 occurrences in 3 texts), *the information in the* (21 occurrences in 4 texts), and *the words in bold* (20 occurrences in 3 texts) are examples from the *metadiscourse and textual reference* sub-functional categories (see Table 6.4).

In contrast, *topic elaboration: cause and effect* and *discourse markers* account for the lowest percentages, totalling 1% of bundle types (see Table 6.5 and Figure 6.5). There is only one bundle type in each of these sub-functional categories; for example, in *topic elaboration: cause and effect*, the bundle *as a result of* (34 occurrences in 4 texts) is included. Furthermore, *discourse makers* includes the bundle *at the same time* (13 occurrences in 3 texts), as shown in Table 6.4. *Topic introduction and focus* shares a closer percentage total with another sub-grouping found under this major classification, which is *topic elaboration: non-causal.* Table 6.5 shows that the sub-functional category *topic elaboration: non-causal* has 3% of bundle types, similar to *introduction and focus*, which includes bundles, with 4% of bundle types. In terms of the *introduction and focus* sub-function, it includes one of the most frequent bundles, which is *look at the following* (32 occurrences in 3 textbooks) alongside other bundles, ranging between 17 to 11 occurrences (see Table 6.4). The *topic elaboration: non-causal* sub-function include bundles such as *research shows that in* (17 occurrences in 3 textbooks) and *research shows that the* (10 occurrences in 3 textbooks), as shown in Table 6.4.

6.4 Concluding remarks

In this chapter, I have presented the results of the 102 most frequent four-word bundles in the textbooks corpus, reporting on their raw frequency counts and overall distribution in terms of the total number of bundle types. *Focus on your subject* and *as a result of* are among the top most frequently occurring bundles in the textbooks corpus. Verb-based bundles form the main structural classification of lexical bundles found in the textbooks bundles list, consisting of bundles such as *focus on your subject*. Overall, the nature of bundles in the textbooks bundles list have been found to have a balanced functional categorization of *referential expressions*, *stance expressions*, and *discourse organizing functions expressions*, which are represented in the bundles *essay with the title*, *focus on your subject*, and *as a result of*, respectively.

In the next chapter, I will discuss the findings reported in chapter 5 (instructors' materials bundles list) and chapter 6 (textbooks bundles list). In Chapter 7, I will look into the most frequent four-word bundles found within the three lists (instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list), by comparing their types and frequencies, and addressing questions 1 and 3 as it will be explained.

CHAPTER VII

COMPARISON OF THE MOST FREQUENT FOUR-WORD BUNDLES IN THE INSTRUCTORS' MATERIALS BUNDLES LIST, TEXTBOOKS BUNDLES LIST, AND THE WRITTEN AFL SUB-LIST

Research question 3 is focused on ascertaining the extent to which the bundles found in the EAP bundles lists (instructors' materials and textbooks; see Tables 5.1 and 6.1, respectively) are based on data derived from the written AFL sub-list (see Appendix C) in terms of types and frequency counts. To my knowledge, there is a very limited number of existing corpora focusing on the lexical bundles of instructors' materials and on the EAP textbooks used in EAP pre-sessional programmes. Therefore, two corpora were constructed (see Chapter 4); the first corpus was created from instructors' materials and the second one from the textbooks used in the EAP pre-sessional programme.

In accordance with the results presented in Chapters 5 and 6, this chapter will report on the results and discuss the bundles shared across the three lists. I will conduct a comparison of the four-word bundles found in the three bundle lists, in relation to their types and frequency. In addition, I will report on and discuss the types of bundles found in the instructors' materials bundles list and textbooks bundles list. I will also report on the in-context information, focusing on bundles found in reading parts or in instructions of these readings, answering research question 4. Finally, I will examine the bundle types, focusing on bundle types as teachable units, in order to address research question 5 (see Appendix A for research questions.) When needed, qualitative evidence will be provided through checking expanded concordance lines to provide a more in-depth examination. The discussion reported here will add to the pedagogy field related to lexical bundle knowledge.

7.1 Shared bundles in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

A comparison of the four-word bundles in the three lists indicates that only four academic lexical bundles are shared: *on the other hand*, *it is important to*, *it is possible to*, and *can be used to* (see Chapter 5, Table 5.1; Chapter 6, Table 6.1; and Appendix C for full lists). The data from Table 7.1 and Figure 7.1 reveals that the bundle *on the other hand* has higher frequency rates in both the instructors' materials bundles list (173 occurrences per million words; henceforth mw) and the written AFL sub-list (119 occurrences pmw) than in the textbooks bundles list (85 occurrences pmw).

In contrast, the bundle *it is important to* occurs less frequently in the written AFL sub-list (43 occurrences pmw) compared to its equivalent in the instructors' materials bundles list (115 occurrences pmw), and occurs even more frequently in the textbooks bundles list (155 occurrences pmw). The bundles *can be used to* and *it is possible to* both display a similar level of frequency, ranging from 35 to 48 occurrences per mw across the three corpora, with the bundle *it is possible to* having a slightly higher frequency in the textbooks bundles list (66 occurrences pmw), as shown in Table 7.1 and Figure 7.1. Comparing these four academic bundles by checking concordance lines in the three corpora was useful because it allowed me to study each bundle individually and to compare them to the written AFL sub-list and other similar studies. The following is a close examination of the four academic lexical bundles that are shared by the three corpora.

Table 7.1. Shared lexical bundles with frequency counts per million words in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

N	Lexical bundles	Freq. pmw (instructors' materials bundles list)	Freq. pmw (textbooks bundles list)	Freq. pmw (written AFL sub-list)
1	*it is important to*	115	155	43
2	*on the other hand*	173	85	119
3	*it is possible to*	35	66	48
4	*can be used to*	46	42	45

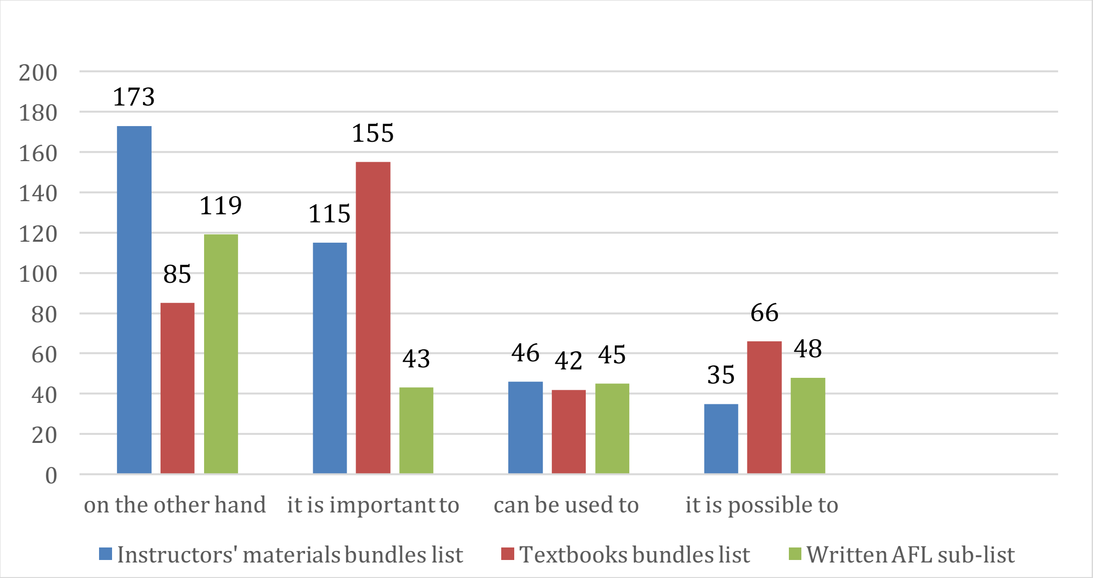

Figure 7.1. Occurrence of the shared bundles in the instructors' materials bundles list, textbooks bundles list and the written AFL sub-list

On the other hand

Having examined numerous studies on lexical bundles, the bundle *on the other hand* stands out as one of the most frequently used expressions. Its significance is derived from its wide prevalence in the literature on lexical bundles in academic registers (e.g. Biber *et al.*, 1999, 2004; Cortes, 2004; Hyland, 2008a, 2008b; Martinez and Schmitt, 2012; Simpson-Vlach and Ellis, 2010; Wood, 2010). It is found across a wide range of academic disciplines such as electrical engineering, biology, history, business, and applied linguistics (e.g. Chen, 2010; Cortes, 2004, 2006; Hyland, 2008b; Jablonkai, 2010; Wood and Appel, 2014). The reason why the textbooks corpus displays a lower occurrence of this bundle compared to the instructors' materials corpus and the written AFL sub-list may be because textbooks seem to focus on classroom-oriented bundles more than academic-based bundles (see Wood, 2010). To illustrate, as shown in Chapter 6, Table 6.1, textbook bundles such as *focus on your subject* and *look at the following* are observed as having higher raw frequencies of 38 and 32, respectively, than the academic bundle *on the other hand*, which occurs 18 times; this will be discussed further in section 7.4.

It is important to, It is possible to, and Can be used to

The results show a degree of variation in the frequency per million words of these three academic shared bundles across the three lists, as shown in Table 7.1 and Figure 7.1. This could be explained by the fact that each of the three corpora incorporated different kinds of texts on various topics that included these types of bundles but at different frequencies, covering different disciplines. This finding is similar to an extent to some results reported in the literature (Chen and Baker, 2010; Cortes, 2004; Wei and Lei, 2011). In those studies, the authors reported that the issue of varied topics might have an impact on the frequency of bundles.

In addition to the written AFL sub-list, the importance of the three academic bundles, *it is important to*, *it is possible to*, and *can be used to*, to the academic register is confirmed by many studies (e.g. Biber *et al.*, 1999) and across different disciplines (e.g. Cortes, 2004; Hyland, 2008b). The presence of these academic bundles in the instructors' materials corpus and the textbooks corpus indicates their usage in the academic register. The authors of the instructors' materials and textbooks make use of these three bundles to indicate their status as subject experts, as well as to engage the learners as participants in performing certain tasks or engage the readers to be a part of an argument (Chen, 2010), as shown in the following examples:

(1) There are no punctuation rules for this group, but ***it is important to*** notice what kinds of words follow these signals. (Instructor 4)

(2) ***It is important to*** consider at this point the extent to which our method genuinely gets at the difficulties that metaphor presented to these students. (Instructor 1)

(3) In citation in academic writing ***it is possible to*** use the present perfect tense with an exact date when citing others. (Instructor 1)

(4) ***It is possible to*** break down each word into four different parts:
de – human – iz(e) – ation
(CAE-Intermediate)

(5) These ***can be used to*** introduce a summary or a paraphrase or a direct quotation. (AEPC Extended Writing Booklet)

The four shared types of bundles in the instructors' materials bundles list and textbooks bundles lists indicate clearly that EAP materials include a relatively small number of academic bundle types compared to those found in the written AFL sub-list at different frequencies. The reasons behind this finding in relation to the instructors' materials and textbooks bundles list will be considered further in sections 7.2 and 7.3.

7.2 Types of bundles in the instructors' materials bundles list compared to the written AFL sub-list

Of the 79 four-word bundles identified in the instructors' materials bundles lists in Chapter 5 (see Table 5.1) and the 57 four-word bundles found in the written AFL sub-list (see Appendix C), and in addition to the four shared bundles noted above, six lexical bundles were shared, which can be categorized into 10 types. Table 7.2 presents the six academic lexical bundles, with types and frequencies (per million words) in both the instructors' materials bundles list and the written AFL sub-list.

Table 7.2. Shared lexical bundles with frequency counts per million words in the instructors' materials bundles list and written AFL sub-list

No.	Lexical bundles	Freq. pmw (instructors' materials bundles list)	Freq. pmw (lexical bundles in the written AFL sub-list)
1	*it is possible that*	58	19
2	*at the time of*	46	32
3	*the purpose of this*	46	13
4	*there are a number*	46	14
5	*to the fact that*	46	23
6	*as part of the*	35	26

As shown in Table 7.2 and Figure 7.2, the six shared bundle types in the instructors' materials bundles list have higher frequencies (per million words) to their counterparts in the written AFL sub-list. These bundles are *it is possible that*, *at the time of*, *the purpose of this*, *there are a number*, *to the fact that*, and *as part of the*. Three reasons might be behind why these academic bundles are more frequent in the instructors' materials corpus than in the written AFL sub-list. First, the six academic bundles might be more relevant to readings or topics (Chen and Baker, 2010; Cortes, 2004; Wei and Lei, 2011) of the instructors' materials provided to EAP learners during their writing classes. Second, it could be due to the fact that the present study focuses on four-word bundles while Simpson-Vlach and Ellis's (2010) written AFL top 200 (see Appendix B) concentrated on different bundle lengths (e.g. three-, four-, and five-word bundles) with different frequencies. Simpson-Vlach and Ellis (2010) reported this range of different bundle lengths in 200 academic bundles, with high and different frequencies (see Appendix B). For example, in the written AFL top 200, the three-word bundles *in terms of* and *the use of* have 282 and 270 occurrences per million words, respectively.

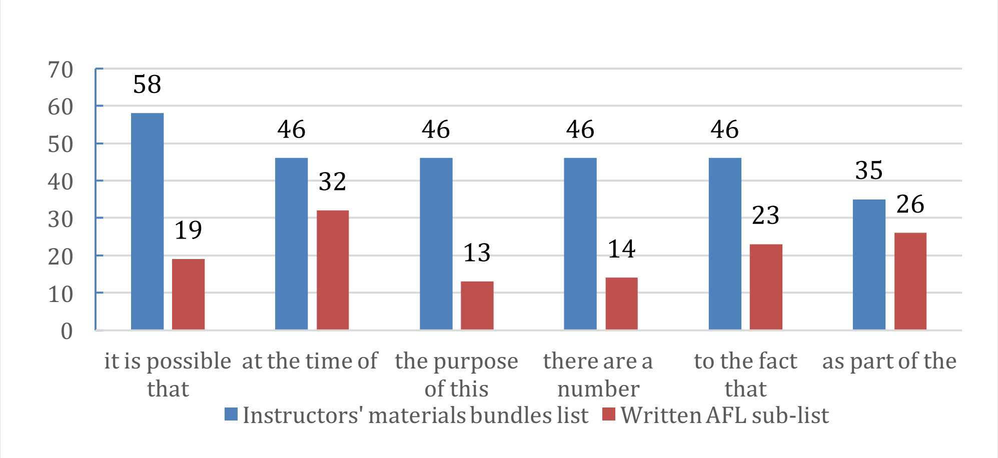

Figure 7.2. Occurrence of the six shared bundles in the instructors' materials bundles list and the written AFL sub-list

A third reason could be that the instructors' materials corpus is relatively small (86,693 words) compared to the much larger corpus on which the written AFL top 200 (2.1 million words) is based. Simpson-Vlach and Ellis's corpus consists of both Hyland's (2004) research articles corpus and selected BNC files sampled across academic disciplines (see Chapter 4), with a wider variety of academic text topics. Therefore, the huge size of the written AFL top 200 corpus and the different bundle lengths across different fields mean there is a vast number of academic bundle types and different bundle lengths. Therefore, although the instructors' materials have higher frequencies of the six shared bundles, the written AFL top 200 (see Appendix B) includes more varied academic bundles with different frequencies, types, and lengths to those found in the instructors' materials corpus. This may also explain the high frequencies of the bundles *on the other hand* and *it is important to* from the four shared bundles in section 7.1 in the instructors' materials corpus and the written AFL sub-list.

The top-ranked bundles (the most frequent, with raw frequency rates from 18 to 7 occurrences) should also be considered (see Chapter 5, section 5.1). These top-ranked bundles have a lower overall distribution of bundle types compared to bundle frequency (with low frequencies of four or three occurrences). The top-ranked bundles have a limited number of types in the instructors' materials bundles list, while the overall distribution of the less frequent bundles (raw frequencies ranging from five to three occurrences) is accompanied by a high level of diversity in terms of types. This is also clearly reflected in the shared bundles (section 7.1), except for the most frequent bundles *on the other hand* and *it is important to*, as previously mentioned. Eight of the ten shared bundles were found to have low raw frequencies within the range of five occurrences and lower (mostly four or three occurrences) in the instructors' materials bundles list, as shown in Table 5.1 in Chapter 5. Simply put, in the instructors' materials bundles list, the bundles with low raw frequencies of four and three occurrences are much more varied in terms of type than high-frequency bundles that range from 18 to 7 occurrences.

Three factors could explain the lower frequency/wide range of types of bundles relationship found in the instructors' materials bundles list: (1) the types of materials selected; (2) the content of the materials used; and (3) the number of documents used in the study. First, the selection of material types by each EAP instructor seems to be one of the three main factors behind the wide range of bundle types reported in the findings. The pre-sessional course, like many EAP programmes, provides a flexible syllabus for teaching staff, as mentioned in Chapter 3. EAP instructors teaching the same writing curriculum may be obligated, to a certain degree, to assign their students with particular, similar academic reading texts or materials (e.g. textbook extracts, reading passages, articles, exercise sheets, writing tasks, etc). At the same time, instructors may be greatly encouraged to bring their own preferred and unique materials to their classes, which would help to enhance learners' academic writing skills, in accordance with the standards required in academia. Therefore, each instructor may select and use different types of materials that they believe are suitable for their EAP learners, including exercise sheets, quiz sheets and reading passages (see Appendix K).

Moreover, depending on the demands and objectives of the writing class, EAP instructors may provide handouts for the purposes of explanation and clarification, and to guide their learners, including work and materials on the conventions of academic writing such as essay conventions (e.g. paraphrasing, summarizing and synthesizing), language support or vocabulary building, (e.g. collocation, metadiscourse/discourse markers and lexical bundles), sentence-level grammar and paragraph writing, reference citation styles and plagiarism, and reading passages. For example, Instructor (1) focused on providing a variety of different writing tasks and handouts (e.g. generating and listing ideas, brainstorming, using linking words in academic writing). The instructor also included grammar handouts (e.g. passive modal verbs, infinitives and gerunds) and academic reading and writing handouts (e.g. reading for key terms and guessing meaning in context). Instructor (1) also included different reading passages and articles. In contrast, Instructor (5) offered grammar handouts (e.g. verbs, cause and effect), information on elements of paragraph writing (e.g. handouts on coherence), and handouts on avoiding plagiarism (see Appendix K).

The second factor, besides the issue of diversity in the types of materials, could be the content of the materials presented by each teacher. For example, Instructor (1) included reading passages and articles (e.g. "A brief history of Sheffield", "Carbon copies", and "Learning the Queen's English... on your mobile phone?"). In contrast, Instructor (5) only provided one document, covering presentation questions on 'Consumers' and physicians' perspectives about high tech wearable health products'. Meanwhile, instructors (2) and (3) presented the same article on plagiarism to their learners, in addition to each presenting unique reading materials.

The third factor concerns the number of materials offered by each instructor, ranging from 1, 6, 8, to 29 documents (see Appendix K). In Chapters 3 and 4, it was observed that instructors submitted an uneven number of learner materials for the purposes of the present study. Thus, these three factors may have played a key role in the lower frequency/wide range of types bundle relationship.

Furthermore, along with the six shared bundles, other low-frequency bundles such *are more likely to* (5 occurrences), *the fact that the*, *the relationship between the* (4 occurrences each), and *as a consequence of*, *in the development of*, *it is difficult to*, and *in relation to the* (3 occurrences each) are identified in the instructors' materials bundles list and are not found in the written AFL sub-list. However, these sequences are considered to be academic bundles found in academic writing and were reported in Biber *et al.'s* (1999) study. Further types of bundles, with low raw frequencies, were also discovered in the instructors' materials bundles list; however, these types of bundles did not seem to appear as common academic bundles found in written prose. These bundles include *form of the verb* (5 occurrences), *to know each other* (4 occurrences), *by no means the*, *in order to avoid*, and *this is not necessarily* (3 occurrences each).

The occurrence of these different types of low-frequency bundles could be related to the nature of the EAP texts. As stated in Chapter 1, lexical bundles are found in texts of all types; however, different kinds of texts include different sets of bundles. In his study on disciplinary variation in lexical bundles, Hyland (2008b: 4 and 20) revealed that his analysis showed that "bundles are not only central to the creation of academic discourse, but that they offer an important means of differentiating written texts by discipline", and "…bundles occur and behave in dissimilar ways in different disciplinary environments". In terms of frequency, Hyland found some interesting disciplinary differences in his corpus when examining lexical bundles found in electrical engineering, biology, applied linguistics, and business studies. In each of these four disciplines, almost completely different sets of items seem to be used. In his conclusion, Hyland (2008b) indicated that electrical engineering incorporated the widest range of types of bundles, which were not found in other disciplines, whereas the narrowest range of types of bundles was found in biology. For example, in biology, bundles such as *in the presence of* and *in the present study* appeared frequently. Moreover, in electrical engineering, Hyland found bundles such as *as shown in figure* and *is shown in fig* to be more common. With regard to engineering, Hyland based his assumption on the consequences of the relatively abstract and graphical nature of technical interaction, stating that technical subjects present their arguments by connecting information or results in routine or formulaic styles.

Drawing from Hyland's (2008b) study, it could thus be the case that the varied content of EAP instructors' materials, addressing a range of topics from different disciplines, could have played a major part in the wide range of bundle types found in the instructors' materials. It is possible that a direct correlation may be made between the lower frequency/wide range of bundle types and the diversified types, content and uneven number of materials.

Moving on to high-frequency lexical bundles, these sequences are considered limited in terms of number of types. There are nine types of bundles with a high number of occurrences. The corresponding nine bundles are found at the top of the instructors' materials bundles list (see Chapter 5). Most of these high-frequency bundles identified in instructors' materials are also common in academic prose, as reported in previous studies (e.g. Biber *et al.*, 1999; Biber *et al.*, 2004). For example, the following bundles: *in the United States* (18 occurrences), *at the end of* (15 occurrences), *the end of the* (12 occurrences), *the extent to which* (9 occurrences), and *as a result of* (8 occurrences) are all apart from the bundle *with the help of* (7 occurrences) recognized in many studies as the most frequent bundles in academic writing (e.g. Biber *et al.*, 1999; Biber *et al.*, 2004; Cortes, 2004, Hyland, 2008b) but are not included in the written AFL sub-list. This finding, therefore, reinforces the fact that instructors' materials include bundles that are mostly found in academic prose.

Three conclusions can be drawn from the bundles shared between the instructors' materials corpus and the written AFL sub-list: (1) instructors' materials contain a relatively small number of academic bundles compared to those found in the written AFL sub-list (57 bundles) (see Appendix C); (2) learners are exposed to a different set of academic bundle types that are not found in the written AFL sub-list but are established as academic bundles in other studies with high and low numbers of occurrences. This does not weaken the significance of the written AFL sub-list. Instead, it reinforces the notion highlighted in Chapters 1 and 2 that a vast number of sequences are used in academic prose (Biber *et al.*, 1999; Cortes, 2004, 2006; Hyland, 2012). This collection of bundle types may make it difficult for EAP material developers and instructors to choose which bundles are most useful for students to learn; (3) instructors' materials include another set of four-word bundles that are not present in the academic prose register: these bundles do not seem to be the most useful for EAP learners.

Overall, I would argue that the types of bundles found and frequencies in the instructors' materials could be because when they create their materials, EAP instructors may not consider lexical bundles as a central factor in material selection. In the present study, I administered a small questionnaire to survey instructors' opinions on lexical bundles in terms of their use in materials. I asked instructors: "How often do you pay attention to lexical bundles when deciding what materials, handouts, and reading passages to use in writing classes?" The responses to this question show the views of instructors on the use of lexical bundles. Instructor (1) reported that he **always** pays attention to lexical bundles, while Instructors (2) and (4) reported that they **sometimes** pay attention to these sequences. In addition, these three instructors consider lexical bundles a useful feature in academic writing. However, there appears to be a conflict between instructors' views and the materials presented in this study.

EAP instructors were also asked to state the reason(s) that made them pay attention to the presence of lexical bundles in the materials, handouts and reading passages that they used during writing classes. Instructors (1) and (4) responded that they were aware of lexical bundles from their previous readings, while Instructor (1) added that he was aware of lexical bundles from previous syllabuses. Instructor (3) stated that she **irregularly** pays attention to lexical bundles, and that she focuses on lexical bundles during teaching but not in her handouts because these sequences are not the focus of the pre-sessional course syllabus. Instructor (5) did not hand in her questionnaire. These responses suggest that some instructors are aware of the usefulness of lexical bundles. However, the limited number of shared bundles (10) suggests that instructors do not heavily rely on bundles from the written AFL sub-list but which are found in other studies. To conclude, instructors' use of academic lexical bundles is rather different compared to the range found in the written AFL sub-list. As stated in the literature review section (see Chapter 2, section 2.3), this conclusion reinforces Wray's notion that the stock of formulaic sequences is always vibrant and changing, meeting the demand of the language user. Thus, the distinctive production of lexical bundles found in the instructors' materials and the written AFL sub-list can be significantly linked to Wray's argument, with authors drawing on different sets of bundles depending on the context and use.

7.3 Types of bundles in the textbooks bundles list compared to the written AFL sub-list

In addition to the four shared bundles in the textbooks bundles list and the written AFL sub-list, there are three additional shared bundles between the two corpora, totalling seven bundles. Table 7.3 presents the three shared lexical bundles identified in both the textbooks bundles list and the written AFL sub-list.

Table 7.3. Shared lexical bundles with frequency counts per million words in the textbooks bundles list and the written AFL sub-list

N	Lexical bundles	Freq. pmw (textbooks bundles list)	Freq. pmw (written AFL sub-list)
1	*a wide range of*	66	31
2	*in the form of*	47	46
3	*a large number of*	42	22

A wide range of and *a large number of* are the two bundles whose frequencies are observed to be higher in the textbooks bundles list than in the written AFL sub-list, as shown in Table 7.3 and Figure 7.3. For example, the bundles *a wide range of* has 66 occurrences pmw compared to 31 occurrences pmw in the written AFL sub-list; and *a large number of* has 42 occurrences per mw compared to 22 occurrences per mw in the written AFL sub-list. The bundle *in the form of* has similar frequencies in the textbooks bundles list and the written AFL sub-list (47 and 46 occurrences pmw, respectively).

Alongside the previously mentioned four bundles, these three bundles have also been established in other studies as academic bundles in addition to occurring in the written AFL sub-list (Cortes, 2004; Hyland, 2008b). It is not clear why these three bundles were more frequent in the textbooks corpus than in the written AFL sub-corpus. Nevertheless, I would argue that four main features were behind the different frequencies and limited number of types of bundles found in the textbooks corpus compared to the written AFL sub-corpus: (1) presentational features, (2) the type of written texts employed in writing EAP textbooks, (3) the practice of instructional language and material selection by textbook authors, and (4) the integrated skills feature. These indicate a different systematic process in the selection of lexical bundles, as will be discussed further below.

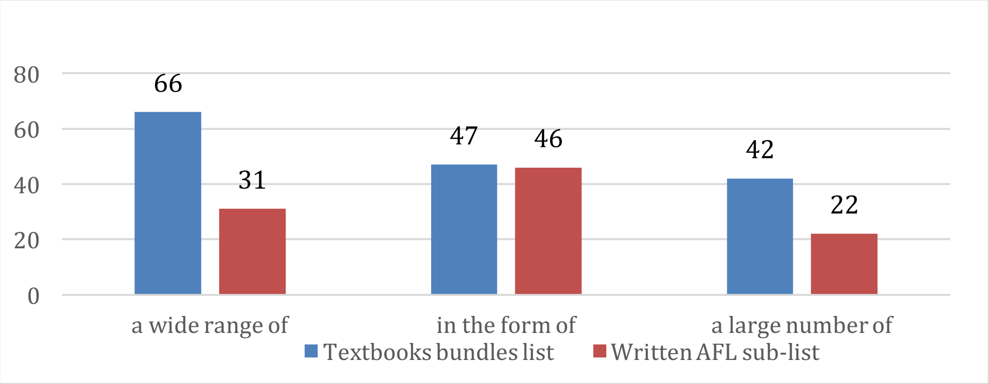

Figure 7.3. Occurrence of the three shared bundles in the textbooks bundles list and the written AFL sub-list

Initially, it may seem that textbook authors employ these bundles without following reliable guidelines on the treatment of lexical bundles and increasingly employing them based on their judgement. This assumption is established on the findings of previous studies. For example, Koprowski (2005) suggests that coursebook designers implemented an unprincipled and careless selection process. The coursebooks he examined included superficial and rare items rather than more useful phrases. He also suggests that coursebook authors may have done poorly at identifying consistently useful lexical phrases. Similarly, Wood (2010), in his study on textbooks, claims that lexical bundles were not treated as a priority by EAP textbook authors. Furthermore, Burton (2012), in his survey of current coursebook authors, suggests that the influence of corpora on coursebooks still remains small due to a large degree of inconsistency among authors when consulting data from corpora. Although these studies examined different textbooks from those examined in the present study (the *Cambridge Academic English* series), the findings are similar, revealing a limited number of types of academic bundles.

However, it is crucial to be cautious in the interpretation of the above findings. Not only do textbook authors follow an organisational framework that helps them structure material for learners, involving a set of instructional language forms, they also consult their own corpora to develop teaching materials. This is indicated on the Cambridge textbooks webpage, which is recommended to serve as supplementary material. This means that textbook authors might not be informed by the written AFL sub-corpus used in this study but are informed by the Cambridge Academic Corpus and the Academic Word List.

In addition, the textbooks included a very small number of bundle types with different frequencies that are not found in the written AFL sub-list but have been identified as lexical bundles associated with academic prose (Biber *et al.*, 1999). The results show evidence of high-frequency to low-frequency bundles such as *as a result of* (34 occurrences), *at the end of* (27 occurrences), *at the beginning of* (18 occurrences), *at the same times* (13 occurrences), and *as part of a* (10 occurrences) in the textbooks list. Such bundle types are considered important and useful to learners because they are associated with the register of academic writing, a notion confirmed by Biber and his colleagues. Other low-frequency bundles found in the textbooks bundles list but which do not seem to be reported by other studies as academic prose bundles are *inferring the meaning of* (11 occurrences) and *a piece of writing* (9 occurrences), indicating that they are non-valuable bundles for EAP leaners. Again, the presence of these bundles may reflect, as similarly indicated in section 7.2, the nature of the EAP context.

Looking into further types of bundles with frequencies in the 20s to 30s in the textbooks bundles list (as shown in Table 6.1, Chapter 6), the findings indicate that the highest-frequency bundles in the textbooks bundles list are primarily bundles which are recognized as instructional language. The bundles are *focus on your subject* (38 occurrences), *look at the following* (32 occurrences), *and answer the questions* (31 occurrences), *the words in the* (29 occurrences), *essay with a title* (28 occurrences), *what do you think* (27 occurrences), *use a dictionary to* (25 occurrences), *you are going to* (23 occurrences), and *the words in bold* (20 occurrences). This finding coincides with the findings in many studies conducted on textbooks (e.g. Chen, 2010; Wood, 2010; Wood and Appel, 2014). The reliance of textbook writers on the use of these types of bundles could be seen as a typical and systematic approach in textbook linguistic production (see Hyland, 2009), used to organize and simplify the information and content for readers/learners, indicating a thematic approach towards producing EAP textbook materials. Some of these bundles are useful for learners to be able to easily complete tasks and exercises, but these types of bundles seem to be insignificant for learners when writing their assignments (Wood and Appel, 2014).

It can be concluded that *Cambridge Academic English: An Integrated Skills Course for EAP*, *Student's Book* uses different four-word bundles than those found in the written AFL sub-list. This leads to the important conclusion that EAP learners encounter a different and limited group of academic bundle types to those found in the written AFL sub-list (see Appendix C). The narrow number of academic lexical bundles in EAP textbooks has been pointed out by many researchers (Wood, 2010; Wood and Appel, 2014). It is important to note that the findings reported here are similar to the results of Wood (2010) and Wood and Appel (2014) on textbooks although the set of textbooks used in the present research is completely different to those used in their studies, as stated in Chapter 2.

From scanning the introductions of the *Cambridge Academic English* series, it was found that the textbook authors claim that they select vocabulary that is important in academic writing. The authors also claim that they make use of the Academic Word List (AWL). They also provide key academic words necessary for the development of academic vocabulary at the back of the Student's Book (*Cambridge Academic English: An Integrated Skills Course for EAP*, 2012). However, there is no evidence that the authors of the *Cambridge Academic English* (2012) series took into consideration formulaic sequences such as lexical bundles used in formula lists, such as the AFL in particular. It is important to note, however, that they claim to have selected the most useful and up-to-date words and phrases used in their own research projects, as mentioned in the information on language and pedagogy research for ELT found at (http://languageresearch.cambridge.org). This could be provided to learners in the form of supplementary materials, which were not examined in this study. The diversity of lexical bundles recognized among the written AFL sub-list and the EAP material corpora in this study supports Nattinger and DeCarrico (1992) and Schmitt and Carter's (2004) argument that any discourse is composed of a large number of sequences, as mentioned in the literature review chapter (see Chapter 2, section 2.4.4).

7.4 In-context information on bundle types

Examining the in-context information of lexical bundles helps to enhance our understanding on the roles they play in academic discourse. For both EAP corpora, once I examined the identified bundles in their textual context, the overall proportions of lexical bundles were determined for the two types of textual environments in which the bundles appear. Figure 7.4 shows the percentages of bundles in the instructional and textual parts in both the instructors' materials bundles list and the textbooks bundles list. From the stacked chart in Figure 7.4, it can be seen that 69% of the bundles in the textbooks corpus occur in instructional parts and only 7% are located in textual parts. However, the significant finding comes from the instructors' materials corpus; only 6% of bundles are found in instructional parts, while 71% of bundles are found in textual parts. Closer proportions of bundles, accounting for 24% and 23% of the total, are found in both parts (instructional and textual context) in the textbooks list and the instructors' materials bundles list.

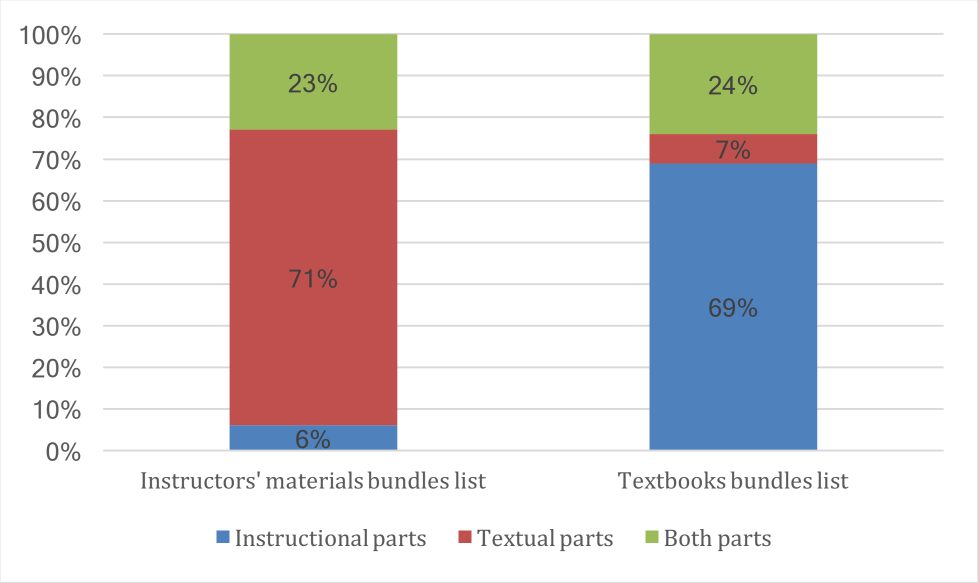

Figure 7.4. In-context percentage distribution of lexical bundles in the instructors' materials bundles list and textbooks bundles list, in instructional, textual, and both parts

Table 7.4 provides a few examples of bundles from both lists with textual environment data, revealing the type of context in which were located (see Appendices Q and R for full lists).

Table 7.4. Examples of bundles in terms of in-context information in the instructors' materials and textbooks bundles lists

Instructors' materials bundles list	Textbooks bundles list
Bundles in textual parts	**Bundles in textual parts**
in the United States	*as a result of*
on the other hand	*on the other hand*
as a result of	*at the same time*
Bundles in instructional parts	**Bundles in instructional parts**
title of the article	*focus on your subject*
form of the verb	*look at the following*
in order to learn	*and answer the questions*
the passive form of	*the words in the*
you may need to	*work in pairs and*
Bundles in both parts	**Bundles in both parts**
at the end of	*it is important to*
the end of the	*at the end of*
it is important to	*in the United States*

In terms of the EAP textbooks, the lexical bundles presented to learners seem to display other academic language, with a pedagogical instructional purpose. The in-context analysis of the lexical bundles in EAP textbooks shows that the largest proportions of bundles are located in instructional parts, including *focus on your subject*, *look at the following*, and *and answer the questions* (see Table 7.4). In fact, from research conducted on textbooks, it has been suggested that lexical bundles in textbooks mostly occur in instructional parts (Wood, 2010; Wood and Appel, 2014). For example, Wood (2010), similar to this study, found that his EAP textbook bundles appeared in the instructional materials but not in texts when he examined four multi-skills (comprehensive intermediate-advance) textbooks (see Table 8.1 in Chapter 8 for samples of Wood's instructional sub-corpus). It should be noted that Wood's (2010) EAP textbook materials were divided into two sub-corpora. The first sub-corpus consisted of reading texts taken from the textbooks, while his second sub-corpus included instructional materials from the textbooks. However, the difference in methodological approach between my study and the other studies did not seem to yield contradictory findings. Wood and Appel (2014: 8) compared multi-word constructions (MWC) in first-year textbooks to those found in EAP textbooks. They concluded that of the MWC found in their corpus of first-year textbooks (engineering and business university textbooks), "less than half appeared in the reading texts in EAP textbooks".

The analysis of the in-context information of the lexical bundles in the instructors' materials revealed that a high percentage of bundles are found in the reading parts. This initial interpretation might lead to the conclusion that instructors' materials have by far the highest proportion of lexical bundles with academic prose characteristics, and a higher proportion than is the case for the textbooks. For example, in Table 5.1 in Chapter 5, the instructors' materials bundles list includes frequent and varied bundles such as *in the United States*, *on the other hand*, *at the end of*, *it is important to*, and *the extent to which*, which are found in textual parts. However, from the analysis in Chapter 5 and in this chapter, it should be clear that although instructors' materials share many bundles found in academic prose, they also rely on bundles not highlighted in academic prose such as *title of the article* and *the best way to*.

7.5 Bundle types as teachable units

A final important aspect to consider is whether the EAP instructors' materials and the textbooks include useful types of academic lexical bundles for learners to practise and use in their written assignments. Thus, from a pedagogical perspective, I examined the texts to see if the identified lexical bundles were explicitly taught to EAP learners, by checking for tasks and exercises for bundle teaching. In the case of the textbooks, of 102 bundles identified, six bundles were explicitly presented and taught to EAP learners. In the instructors' materials, only four of the 79 documented bundles were presented to learners as teachable units. This information is listed in Table 7.5, which includes in-text information.

Table 7.5. Lexical bundles highlighted as teachable units in the EAP bundles lists

Teachable units in textbooks	In-text information	Teachable units in instructors' materials	In-text information
as a result of	one textbook one time	*on the other hand*	Instructor 4 / one time
on the other hand	two textbooks three times	*as a result of*	
at the same time *the way in which* *a great deal of* *a large number of*	one textbook one time	*as a consequence of* *in order to learn*	

Significantly, comparing the bundles highlighted in Table 7.5 to those found in other studies, it would seem that these bundles are considered important academic bundles (see Biber *et al.*, 1999; Cortes, 2004; Hyland, 2008b; Simpson-Vlach and Ellis, 2010) that need to be explicitly taught in EAP writing courses. Two exceptions are the bundles *a great deal of* and *in order to learn*, which were not highlighted as academic bundles in the above-mentioned studies. This suggests that learners are taught unimportant or irrelevant bundles (Koprowski, 2005), as shown in examples (1) and (2) (text source is enclosed in parentheses).

(1) Avoid the use of "a lot of"; use the more formal equivalents of "***a great deal of***", "a large number of", or "many". (AEPC Extended Writing Booklet)

(2) The grammar patterns below are often used for purpose statements.
pattern example
in order to
+ verb
in order to learn (Instructor 4)

Returning to the bundle *on the other hand*, the examination of concordance lines in the present study revealed interesting findings. The following are examples taken from both corpora (see examples 3 to 6):

(3) Transition signals are expressions such as, first, finally, and however, or phrases such as in conclusion, ***on the other hand***, and as a result. (Instructor 4)

(4) Read the extract again and complete it with the best linking phrases from the following list.
1- also, so
2- Furthermore, As a result
3- This can mean that/ ***On the other hand***
(CAE-Intermediate)

(5) Write the following conjunctions and sentence connectors in the box in the correct place in the table. Use a dictionary to check your answers.
as soon as at the same time because***on the other hand***
otherwise
(CAE-Upper Intermediate)

(6) Disagree with this position
However, …
On the other hand, …
(CAE-Upper Intermediate)

From examples 3, 4, 5, and 6, it can be seen that only one instructor (Instructor (4)) out of five, and two textbooks (CAE-Intermediate and CAE-Upper Intermediate), drew learners' attention to the bundles *on the other hand* through explicit and direct teaching. They provided different types of exercises to teach EAP learners the academic bundle *on the other hand*. The tasks in the EAP materials ranged between different practices (e.g. *"read the extract again and complete with the best linking phrases…"*, *"Write the following conjunctions and sentence connecters…",* and "*Disagree with this position"*), as seen in the examples above. This means that not all EAP learners encounter this academic bundle in a similar way. This seems like a plausible argument, considering that each instructor and author are entitled to select the linguistic items they feel are suitable for their learners. However, it is important for instructors and textbook authors to take into account the fact that EAP learners need explicit guidance and more practice in order to understand how to use such an important academic bundle (Cortes, 2004; Jones and Haywood, 2004).

As stated in section 7.1, the bundle *on the other hand* is among the most important lexical bundles. In fact, Hyland (2008b: 13) reported that the bundle *on the other hand* is among the "best candidates for a general EAP course". In addition, in terms of learner corpora, the bundle *on the other hand* is considered to be the most frequently used bundle by professional writers (Ädel and Erman, 2012; Wei and Lei, 2011) and by students (Cortes, 2004). Nekrasova's 2009 study found that second-language speakers of English (L2) found the expression *on the other hand* to be among the most useful bundles. Levy (2008), for example, reported that the bundle *on the other hand* was mostly used by professional writers and upper-level learners. In contrast, Wei and Lei (2011) noted that advanced learners used the bundle *on the other hand* more than professional writers. Cortes (2004) noted that the bundle *on the other hand* was most frequent across three levels of students writing in history. These studies conclude that the bundle *on the other hand* is commonly found in both professional and students' academic writing and is pervasive in the literature.

Even more importantly, the issue of how learners employ the bundle *on the other hand* in their academic writing is a vital area to examine. There is a line of research on learners' corpora which provides insights into how the bundle *on the other hand* is used by learners of different proficiency levels. For example, Jones and Haywood (2004) taught certain formulaic sequences to non-native students in EAP classes. The results showed that learners were able to develop their knowledge of lexical bundles, but most were not able to use these bundles efficiently in their written assignments. Chen and Baker (2016) confirmed that learners have the tendency to overuse certain bundles such as *on the other hand* compared to native student and expert writers. Similarly, but in a different vein, Bychkovska and Lee (2017), in their study of lexical bundles in L1 and L2 university students' argumentative writing, found that the bundle *on the other hand* was among the most misused bundles in the corpus they used. They found that students made structural mistakes, including missing articles (e.g. *on other hand*).

Thus, it can be established from the above discussion that the bundle *on the other hand* is not only an important and favoured academic bundle, but it can also be problematic for EAP learners. The exclusion of such bundles in texts written for pedagogical purposes and the limited amount of exposure and practice to such bundles in instructors' materials and textbooks means that EAP learners may end up lacking the understanding and knowledge needed on how to use these significant academic lexical bundles in their writing (Cortes, 2004). This is also the case with other sequences, where a certain bundle was found in textbooks but not in instructors' materials, as was the case with the next bundle to be reported on.

From scrutinizing the textbooks corpus and instructors' materials corpus, it seems that the textbook authors and instructors did not take advantage of one of most useful bundles in the written AFL sub-list, which is the bundle *a large number of*. There is clearly little evidence of the teaching of this bundle, with both textbooks and instructors' materials missing appropriate exercises or tasks on how EAP learners should use this bundle in their academic writing. The analysis revealed that only one textbook provided the following advice on the usage of the bundle *a large number of*, while the instructors' materials did not include such advice at all:

> Avoid the use of "a lot of"; use the more formal equivalents of "a great deal of", "***a large number of***", or "many". (AEPC Extended Writing Booklet: 14 and 15)

In the above example, it is not difficult to notice the explicit recommendation on the usage of the bundle *a large number of*. Scanning the textbook for further details showed that the author provided one task for readers to practise this academic bundle. The exercise stated: "change the following sentences into a more appropriate style for written academic English".

Regarding the remaining bundles *at the same time*, *as a consequence of*, and *a large number of*, I reviewed the literature on the use of these three bundles in academia (see Table 7.4). The bundle *a large number of* and *as a consequence of* mostly appeared in biology and academic prose (Cortes, 2004; Simpson-Vlach and Ellis, 2010), whereas the bundle *at the same time* was found in biology writing (Cortes, 2004; Hyland, 2008b) and applied linguistics, electrical engineering, and business writing (Hyland, 2008b). These studies confirm the importance of these bundles and endorse their usefulness not only in EAP materials (Hyland, 2008b; Simpson-Vlach and Ellis, 2010) but also in discipline-specific materials (Cortes, 2004; Hyland, 2008b). Examples (7) and (8) highlight their usage:

(7) Note that owing to, ***as a consequence of***, and an account of can be all used in the same way. (Instructor 4)

(8) Sequencing phrases
first of all…
after that …
next …
then …
at the same time … (CAE-Upper Intermediate)

Finally, scanning concordance lines for the academic bundle *the way in which* (Biber *et al.*, 1999) showed that one textbook treats this bundle as an unnecessary item, guiding the learners to a certain grammatical feature (see Example 9).

(9) Complex noun phrases. Academic texts contain many examples of complex noun phrases. Often these take the form of a noun followed by a prepositional phrase beginning with *of*. Compare the following two sentences, where the second one uses a complex noun phrase with *of* to express the idea more efficiently.
… focusses on ***the way in which*** second-language pronunciation is acquired.
… focuses on the acquisition of second-language pronunciation.
Notice how the structure of the sentence changes.
The verb becomes a noun:
is acquired = the acquisition of
Some words are not necessary: ***the way in which*** (CAE-Upper Intermediate)

Overall, in terms of the handling of lexical bundles in the textbooks and by the instructors and from the above in-context analysis, it can be seen that EAP materials examined here provide little information and coverage on the teaching of limited types of academic lexical bundles to EAP learners. In order for learners to learn how to use academic lexical bundles, they need considerable exposure to academic language in general through substantial readings, and to academic lexical bundles in particular. Li and Schmitt (2009: 93) stated that "it is not surprising that academic reading materials play such an important role in lexical phrase acquisition and use, since international postgraduates are required to understand their subject topics and build up their knowledge through large amounts of reading". In the present study, because instructors' contribution of reference materials such as reading articles was relatively small, the academic lexical bundles generated from these academic readings may also be considered small. This interpretation is based purely on the materials examined in the present study; however, instructors and the EAP pre-sessional course may have provided further online reading articles that were not included in the study. In addition to exposure to reading materials, learners need awareness-raising activities and ample practice devoted to learning useful academic bundles (Cortes, 2006; Jones and Haywood, 2004; Lewis, 1997) (see Chapter 11, section 11.2.1).

Furthermore, as part of looking into lexical bundle teaching during writing classes, in the final part of the questionnaire my aim was to find out what type of academic lists instructors were directly offering to their EAP learners. Two instructors, (2) and (3), reported that they used the Academic Word List (AWL), while Instructors (1) and (4) indicated that they were using lists found at the end of chapters or textbooks. Instructor (1) added that he also offered the AFL to his learners. However, when scanning the instructors' materials handouts and from the analysis offered in this section, there was no evidence of the AWL or AFL among the handouts. In addition, as shown in Table 7.4, all four bundles featured as teachable units were presented in the material of one instructor (Instructor (4)). This leads to two interpretations: (1) Only one instructor's learners could be aware of the importance of these limited bundles; (2) instructors seem to pay little or no attention to academic lexical bundles as teachable units in writing classes.

The limited treatment of lexical bundles suggests that instructors and textbook designers pay little attention to the teaching of academic bundles. Therefore, the limited tasks and exercises presented in EAP materials related to the teaching of useful bundles may not lead readers to gaining an understanding of their use (Cortes, 2006). This is considered a drawback for textbook users as well as instructors since studies on teaching lexical bundles (Jones and Haywood, 2004; Cortes, 2006) have suggested that learners need more exposure to the use of lexical bundles to help them to express technically complicated thoughts in an economical way.

7.6 Concluding remarks

This chapter has provided a discussion of the quantitative findings, reporting on and comparing the types and frequency of the bundles found in the instructors' materials bundles list, textbooks bundles list, and written AFL sub-list. It also provided an overall discussion on the types of bundles found in both the instructors' materials corpus and the textbooks corpus compared to the written AFL sub-list. Finally, the types of bundles found in particular contexts and as teachable units were addressed. Only four bundles were shared among the three lists, revealing that EAP materials draw on small numbers of academic bundles found in the written AFL sub-list. In addition, EAP instructors and textbooks do not provide ample related exercises or tasks to teach students the important types of lexical bundles found in the written AFL sub-list and academic prose. However, this conclusion is based on the EAP materials examined in this study, whereas it could be the case that the Cambridge textbooks webpage and the pre-sessional course compensate for this problem by supplementing further online readings and worksheets on lexical bundles that were not examined in this study.

In the following chapter, I will provide structural comparisons of the shared bundles and an overall structural comparison of bundles in the three lists.

CHAPTER VIII

STRUCTURAL COMPARISON OF LEXICAL BUNDLES IN THE INSTRUCTORS' MATERIALS BUNDLES LIST, TEXTBOOKS BUNDLES LIST, AND THE WRITTEN AFL SUB-LIST

Few studies have reported on the structural classifications aspect of EAP textbooks and even fewer have focused on the unpublished materials of EAP instructors. Therefore, the purpose of the present study has been to profile the most frequent four-word bundles in an EAP pre-sessional programme in the UK, focusing on materials relating to teaching academic writing.

Structural classifications are needed to address research question 3, which is concerned with understanding the extent to which the bundles identified in the EAP corpora are based on data retrieved from corpus-driven lists such as the AFL (Simpson-Vlach and Ellis, 2010). In the following sections, the results and discussion of the structural categorizations of the bundles will be presented, comparing the overall and the shared structural patterns in the three lists: instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list.

8.1 Overall structural comparison between lexical bundles in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

A structural comparison between the four-word bundles in the EAP lists and the written AFL sub-list revealed that the lexical bundles identified appear to account for the highest proportion of VP-based bundles across the three lists, and are most frequent in the textbooks bundles list (44%) and the written AFL sub-list (42%) (see Figure 8.1). In terms of PP-based structures, the proportion of lexical bundles with this structure is the lowest in the textbooks bundles list (20%), compared to the higher and very close percentages in the instructors' materials bundles list (33%) and the written AFL sub-list (34%). There are varied proportions associated with NP-based bundles, with the majority of these lexical bundles found in the textbooks bundles list (29%), while the lowest proportion is found in the written AFL sub-list (19%) (see Figure 8.1). Structures with other expressions display the lowest proportions of bundles across the three lists. For example, as shown in Figure 8.1, the written AFL sub-list (5%) has the lowest proportion of this type of bundle compared to the instructors' materials bundles list (10%).

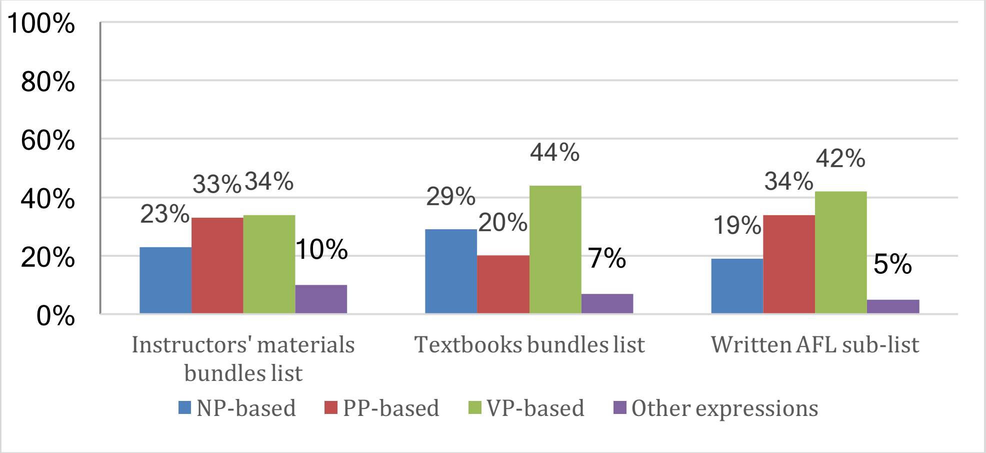

Figure 8.1. Overall structural distribution of lexical bundle types in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

Based on the structural features of lexical bundles found in the spoken and written registers reported in Biber *et al.* (1999) and Biber (2006, 2009) (see Chapter 2), it appears that the three lists mostly contain VP-based bundles. This is a characteristic of the spoken register, despite the fact that the written AFL sub-list should be closest to the norms of the written academic register. This means that EAP materials and the written AFL sub-list mostly make use of VP-based structures often found in conversation and classroom teaching registers (Biber *et al.*, 2004). At the same time, an initial finding is that EAP materials make use of NP-based and PP-based structures, matching the structural characteristics of academic prose. However, at this primary stage, it is rather early to make firm conclusions without looking in more depth at the sub-categorization of the lexical bundles. By looking further into the sub-categories of each structural group, more detailed information will emerge. An analysis of the overall structural comparison (similarities and differences) of the three lists will be reported; by doing so, this will reveal the structures of bundles found in the EAP lists (see sections 8.1.1, 8.1.2 and 8.1.3).

8.1.1 VP-based structures

As reported in section 8.1, although VP-based bundles appear to account for the highest proportion of lexical bundles across the three lists, there are significant differences in their sub-categories, bearing in mind the fact that the written AFL sub-list is considered a good representative of bundles found in academic prose (see Chapter 4 for the reasons behind this). The results show that the instructors' materials bundles list includes 11 sub-categories in the VP-based structure (see Chapter 5, Table 5.3), the textbooks bundles list includes 15 (see Chapter 6, Table 6.3) while the written AFL sub-list includes 7 sub-categories (see Appendix F and G). Figure 8.2 shows the proportions of the VP-based sub-categories in the three lists, highlighting significant structural differences and similarities.

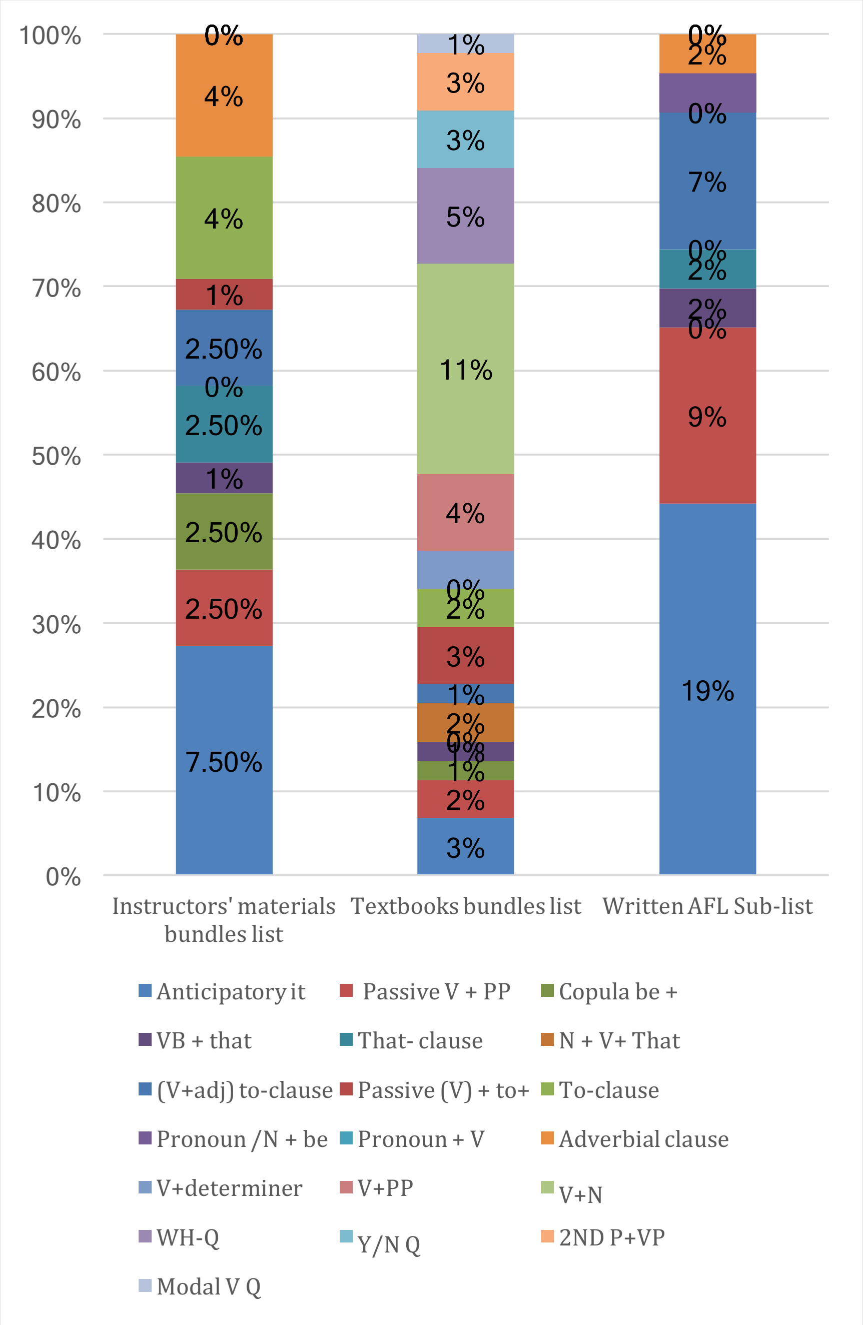

Figure 8.2. Distribution of lexical bundle types of VP-based sub-categories in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

The most significant difference is that *anticipatory it + phrase* is a prominent VP-based sub-category in the written AFL sub-list, signifying a common and important sub-structure in the written register, as demonstrated in many research studies (Biber *et al.*, 1999; Biber *et al.*, 2004; Cortes, 2004; Hyland 2008b). The *anticipatory it + phrase* pattern accounts for the largest proportion of bundles in the written AFL sub-list (19%) and instructors' materials bundles list (7.5%) (see Figure 8.2). In contrast, the textbooks bundles list includes only a limited number of bundles in this sub-category, accounting for only 3% of the total. For example, the *anticipatory it + phrase* bundles *it is important to* and *it is possible to* are two shared bundles in this sub-category in the three lists, as first noted in Chapter 7. In this sub-category, the instructors' materials bundles list also includes bundles (e.g. *it is difficult to*; see Chapter 5, Table 5.1) that are not found in the written AFL sub-list but which are reported in other studies (Biber *et al.*, 1999; Cortes, 2004).

With regard to the instructors' materials corpus, four VP-based sub-structures are found across the instructors' materials bundles list and the written AFL sub-list, including *passive verb + prepositional phrase*, (*verb phrase* +) *that-clause*, (*verb/adjective* +) *to-clause*, and *pronoun/noun phrase + be* (+...). However, the instructors' materials bundles list includes different bundles with these VP-based sub-structures compared to those in the written AFL sub-list (see Chapter 5, Table 5.2 and Appendix F for the bundles). For example, in the *passive verb + prepositional phrase* sub-category, the bundle *can be used in* was found in the instructors' materials bundles list but not the written AFL sub-list. Under the sub-category (*verb/adjective* +) *to-clause* and *pronoun/noun phrase + be* (+...), the bundles *can be used to* and *there are a number*, respectively, are both found in both lists.

In addition, instructors make use of sub-categories that are not included in the written AFL sub-list classifications (e.g. *pronoun phrase + verb* and *adverbial clause fragment by the subordinator phrase (in order to*), accounting for 4% of the total; see Figure 8.2). This VP-based structure includes bundles such as *in order to avoid*, *in order to learn*, and *in order to test*. The *pronoun phrase + verb* structure seems to include bundles such as *he was unable to* and *you may need to*, which are likely to be used in classroom teaching and conversation but not in academic prose (Biber *et al.*, 2004).

Taking into consideration the VP-based grammatical structures, it is rather surprising that the instructors' materials make extensive use of *VP-based* bundles.

An obvious interpretation of the presence of these verb-based patterns may be that the instructors' materials contain some sections and excerpts from other coursebook chapters, providing grammar exercises and writing skills practice, which may have influenced the structural form. For example, the bundle *can be used in* can be found in the instructors' materials to explain a certain grammar point to learners, as in "*pronouns can be used in the…*". A second example is the VP-based structure *adverbial clause fragment by the subordinator phrase (in order to*); one instructor's materials included the following grammatical example: "The grammar patterns below are often used for purposes statements. Pattern example: in order to + verb = *in order to learn*." Therefore, the reason for instructors' materials bundles having VP-based structures may be that the instructors' materials include similar types of resources, exercises and tasks to those found and used in grammar textbooks, explaining the high use of these verb phrase structures in the instructors' materials. However, in Chapter 7, it was established that the bundles in the instructors' materials are not all similar to those found in most ELT grammar textbooks although they seem to share similar educational and pedagogical features.

Furthermore, concordance lines show that the different types of bundles with *VP-based* structures in the instructors' materials are from parts which are intended for learners to read, such as sentences or academic passages. Repeatedly across the corpus, I found that the instructors' materials rely heavily on *VP-based* patterns that are not related to grammar exercises or tasks. For example, the bundle *it is important to* occurs 10 times, mostly in academic reading articles or paragraphs. This means that the instructors' materials incorporate bundles with verb structures, as part of the construction of the written EAP materials such as reading passages.

In a similar vein, Jablonkai (2010) confirmed that the bundles in the study, which were extracted from texts in the European Union (EU) corpus, contained a large number of verb phrase structures. However, Jablonkai's (2010) bundles were found to consist of a *verb phrase with passive verbs*, similar to the written AFL sub-list (accounting for 9% of the total); in contrast, in the instructors' materials in the present study, this structure accounts for only 2.5% of the total. The majority of the instructors' materials bundles list consist of bundles composed of *anticipatory it phrases. Anticipatory it phrases* include bundles such as *it is important to*, which is a common structure in academic prose, as reported in Biber *et al.* (1999) and Hyland (2008b) and as repeatedly stated within the present study. It is important to note that *anticipatory it phrases* and *passive verbs* are two structures that are commonly associated with the language of science (Sánchez, 2014). Therefore, such structures commonly occur in the written AFL sub-list, resembling structures commonly used in research articles across different disciplines (e.g. humanities and arts, social sciences, natural sciences/medicine, and technology and engineering).

It is somewhat difficult to pinpoint the real reason behind the high number of bundles containing verb structures (which account for 34% of the total) (see Figure 8.1) in the instructors' materials. Examining the instructors' materials shows that a number of instructors provided materials focusing on different topics. For example, EAP instructors included topics on climate change, effects of colour, gold, dolphins, Ancient Egypt, and wearable health products. The technical or scientific aspect of the above themes may have had an influence on the presence of the verb structures. This speculation is supported by Hyland's (2008b: 11) claim, where he states:

> The Science and Engineering text, on the other hand, employed significantly more passive bundles, normally followed by a prepositional phrase fragment typically marking a locative or logical relation. … Interestingly, the science writers also tended to employ more examples of the anticipatory –it pattern, which is another means of disguising authorial interpretations.

In Hyland's (2008b) study, Engineering writers relied heavily on bundles with different structures to those in any other fields such as biology, due to the unique nature of academic practice in the engineering department. It is important to bear in mind that the instructors in the present study provided materials that included articles and activities written by different authors across different fields. In line with Hyland's finding, the different verb structures used across EAP instructors' materials are perhaps related to the scientific topics or the distinctive writing styles found in the current study. For example, the bundle *to know each other* has a *to-clause fragment*, which occurs five times in the corpus and in the materials of two instructors. By examining this bundle closely, it was found that this bundle is used in academic readings on the topic of friendship and in a task involving "match the sentences". Furthermore, the bundle *it is important to* is found in a topic related to plagiarism, and the bundle *have shown that the* (*verb phrase+) that-clause*) is found in a topic related to the effect of colour.

Furthermore, perhaps this finding results from the fact that instructors may want or need to draw on different resources to develop and maintain their materials and lesson plans, covering a wide range of writing conventions, styles and fields. For example, in EAP programmes, instructors who teach writing classes are required to incorporate a wide range of focused lessons, covering linguistic and grammatical items, and to develop their learners' writing skills by providing information and presenting handouts on areas such as paraphrasing, referencing, writing styles, and grammatical and lexical knowledge. This may eventually result in the frequent production of these *VP-based* structures, as mentioned above. For example, an instructor used a bundle with the form *pronoun/noun phrase + be (+…)* to present a point to EAP learners related to writing styles (e.g. *There are a number of different connectors, which are covered in this section*). Overall, the prevalence of *VP-based* sub-categories in the instructors' materials is an indication that EAP learners encounter large numbers of bundles with VP-based structures. Instructors' materials include bundles with the *anticipatory it* form as a main VP-based structure, similar to that found in the written AFL sub-list and in academic prose (see Biber *et al.*, 1999). However, the instructors' materials bundles list also incorporates VP-based structures that are not used in the written AFL sub-list and are not relevant to academic prose such as *adverbial clause fragment by the subordinator phrase (in order to*).

Regarding EAP textbooks, in Chapter 6, I categorized the lexical bundles drawn from the textbooks corpus, and established that textbooks exhibit the highest number of lexical bundles, mostly incorporating a *verb phrase* structure (Figure 8.1). Comparing the sub-categories in the textbooks list and the written AFL sub-list, it was found that textbook authors not only rely on varied VP-based structures, but also make great use of the following verb sub-structures that are not found in the written AFL sub-list. Textbook authors used bundles with *copula be + noun or adjective*, *noun + verb phrase + that-clause*, *(passive) verb phrase + to-clause*, *to-clause fragment*, *verb phrase + determiner phrase*, *verb phrase + prepositional phrase*, *verb phrase + noun phrase*, *wh-questions*, *Yes-no questions*, *2nd person pronoun you +VP fragment*, and *modal verb question* when constructing their materials (see Chapter 7, Table 7.2). In addition, five of the VP-based structures were found in Biber *et al.* (2004), including *wh-questions*, *Yes-no questions*, and *2nd person pronoun you +VP.* However, in their study, Biber *et al.* (2004) did not explain whether these VP-based structures were part of the patterns used in textbooks; in fact, they encouraged a fuller analysis to be conducted of the structures found in textbooks.

The findings on VP-based structures confirm one of the main features emphasized in EAP textbooks, that is, that verb-centred structures are widely used in English grammar textbooks and in ESL and EFL textbooks and curricula (Coxhead and Byrd, 2007). Sánchez (2014) also pointed out that, in terms of verb phrase structures, the highest percentage of bundles incorporated a verb phrase (*with passive verb*), accounting for nearly 52% of occurrences in a study of biology textbooks. In the present study, the most frequently occurring bundles in the textbooks list (accounting for 11% of the total) consist of a *verb phrase + noun phrase fragment* (e.g. *look at the following*), followed by the sub-category *wh-questions* (5%), and *verb phrase + prepositional phrase* (4%) (Figure 8.2).

Similarly, but with a different VP-based categorization, Jablonkai's (2010) study of written English within European Union texts shows use of a somewhat high number of lexical bundles incorporating *verb phrases*. With regard to *VP-based* bundles, these findings seem to contradict the findings from earlier studies of written registers. For example, Biber *et al.* (2004) and Biber (2006, 2009) reported that textbooks only include a small number of bundles containing verb phrases. Furthermore, Sánchez (2014) reported that the 2nd person pronoun and *wh-questions* were not found in the university-level textbooks under investigation. In addition, Biber *et al.* (1999) included bundles with verb-based structures in academic prose.

Although the EAP textbooks used in the present study are not grammar-focused texts but EAP integrated skills coursebooks, the results nevertheless have revealed bundles incorporating a high number of *verb phrases*. It should be noted that, unlike the present study, Biber (2006) examined textbooks that were not targeted at teaching EAP learners. Instead, the study examined university-level textbooks aimed at a specific discipline such as business, education, engineering, humanities, natural science, and social science. I would argue that the reason behind the structural disparity is dependent on the differences between the kinds of textbooks considered for analysis. This claim is supported by studies conducted by Wood and Appel (2014) and Chen (2010), in which they found that the majority of sequences in EAP textbooks either did not appear at all or did not produce similar bundles to those found in an introductory first-year business or engineering university textbooks corpus.

Jones and Haywood (2004) reviewed four academic writing textbooks commonly used in EAP courses to examine their use of lexical bundles. Unlike the integrated skills nature of the *Cambridge Academic English* and EWB textbooks (see Chapter 3) used in the present study, Jones and Haywood (2004) examined academic writing focused textbooks: S*kills in Action* (Sellen, 1982), *Academic Writing Course* (Jordan, 1990), *Writing* (White and McGovern, 1994), and *Writing Academic English* (3^{rd} ed) (Oshima and Hogue, 1999). Jones and Haywood (2004: 270) found that the "Structure and Vocabulary aid" at the end of each chapter with academic phrases was not very useful for learners and concluded that students may end up using expressions that are rare in academic prose. Similarly, Koprowski (2005) in his study comparing lexical phrases in ELT coursebooks to data taken from the COBUILD corpus revealed that many useful lexical bundles were missing from the textbooks. This means that textbooks not only display different or other types of bundles (Chapter 7) but they also present different and other VP-based structures to EAP learners.

The dense use of instructional language and grammatical tasks may perhaps be another reason why EAP textbooks rely extremely heavily on verb phrase patterns. In the present study, it could be argued that the strong emphasis on only focusing on and analysing writing materials may have influenced the redundant use of *verb-based* structures. During the clean-up process (see Chapter 4, section 4.2.1.1), I only included and examined texts and tasks targeted at teaching writing skills, which included grammar sections. The clearly established distinction between spoken and written registers in previous studies (e.g. Biber *et al.*, 2004, 2009) gave support to the idea to focus completely on the EAP written register (see Chapter 2). Hence, I included all tasks, rules and exercises with vocabulary and grammar guidelines relating to the teaching of academic writing. For example, I included all explanations and exercises on different grammatical features (e.g. using the simple present tense in generalization and statements), which is a typical pedagogical feature or learning strategy when teaching academic writing in EAP coursebooks. Even if this is taken as a drawback or a negative element behind the existence of *verb phrase* structures in the instructors' materials, it is an even more important reason to explore only the materials aimed at teaching academic writing, to show the reality of the bundle types learners encounter, in terms of frequency, structure and function.

As noted, the repetitive and explicit use of instructional language found in bundles (e.g. *look at the following* [32 occurrences], *what do you think* [27 occurrences], and *work in pairs and* [29 occurrences]) seems to have boosted the production of *verb phrase* structures. On the one hand, such bundles, which incorporate verb phrase patterns, may be critiqued for being inappropriate, irrelevant or insignificant for expanding learners' academic vocabulary. On the other hand, for systematic and organizational purposes, textbook authors seem to prefer to use these types of structures to offer clear and systematic instructions. In the case of textbooks, the instructional language consists of a variety of lexical bundles all working together to express different instructions for learners to follow (e.g. *use a dictionary to* [25 occurrences] and *complete the following sentences* [11 occurrences]). Through the use of these repeated instructional bundles, EAP textbook authors seem to find this type of verb phrase structure very useful, providing learners with language that offers easy-to-follow directions and guidelines. Even in the case of *anticipatory it + phrase*, textbook authors make a limited use of this sub-category structure. And when they do use it, it is found either within reading passages, as demonstrated in Example 1, or it is mostly used to supply learners with information relating to academic writing conventions, as shown in Example 2:

(1) ***It is important to*** consider why children experience difficulties with writing poetry, in spite of reading and hearing it. (CAE-Advanced)

(2) ***It is important to*** make sure that all your information relates to the topic of the essay and to check that you have included all the information that you need to. (CAE-Intermediate)

Overall, the initial analysis showed that textbook authors and instructors' materials make use of relatively large numbers of VP-based bundles, which contrasts with earlier findings on written registers (e.g. Biber *et al.*, 2004; Biber 2006, 2009), where some verb-based structures were detected in academic prose. To conclude this section, EAP textbooks widely use VP-based structures that are not commonly found in the written AFL sub-list.

8.1.2 PP-based structures

In terms of PP-based structures, the instructors' materials bundles list includes the sub-categories *prepositional phrase with embedded of-phrase* and *other prepositional phrase*, which account for high percentages of bundles; a similar finding applies to the written AFL sub-list (see Figure 8.3). *Prepositional phrase with embedded of-phrase* is clearly among the most frequently occurring sub-categories, accounting for 20% and 19% of bundles in the instructors' materials bundles list and the written AFL sub-list, respectively. However, in the textbooks bundles list, the lexical bundle percentage is the lowest in the sub-category *prepositional phrase with embedded of-phrase* (5%), but the sub-category *with other prepositional phrase (fragment)* (15%) has a similar percentage total to that in the written AFL sub-list (see Figure 8.3).

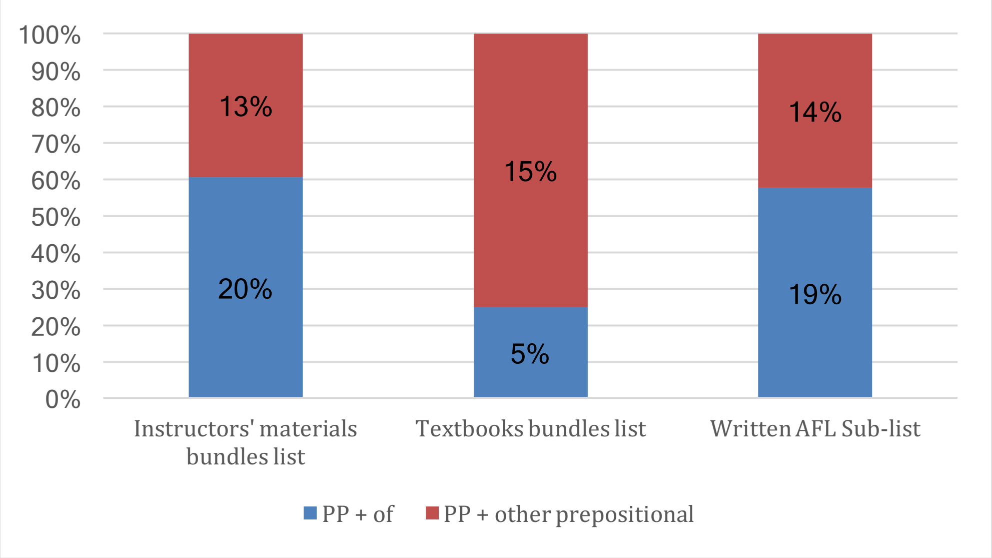

Figure 8.3. Distribution of lexical bundles types of PP-based sub-categories in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

The results of the sub-category *prepositional phrase with embedded of-phrase* and *other prepositional phrase* show that the instructors' materials bundles list includes the academic bundles and PP-based structures found in academic writing. The PP-based bundles *in the United States* (18 occurrences), *at the end of* (15 occurrences), *on the other hand* (15 occurrences), *as a result of* (8 occurrences), are among the most frequent in the instructors' materials while *at the beginning of* (6 occurrences), *at the time of* (4 occurrences), *at the university of* (4 occurrences), *to the fact that* (4 occurrences), and *in relation to the* (3 occurrences) have frequencies of from six to four (see Chapter 6 for the full list of bundles found in the instructors' materials), and are confirmed in other studies (Biber *et al.*, 1999; Cortes 2004).

Unlike the verb-based structure, the PP-based structural form is widely recognized in academic writing (see Biber *et al.*, 1999; Biber *et al.*, 2004; Cortes 2004). The structural characteristics of the PP-based bundles in the instructors' materials align with the forms found in previous studies of academic writing, supporting the notion that academic writing is marked by prepositional phrases. In other words, since the EAP materials gathered for this study (see Appendix K) were either created by the EAP teachers, adopted from other texts, or included passages and reading articles written by professional writers, it would be expected for the nature of the bundles found in EAP instructors' materials to exemplify the inherent quality of the academic written register. In summary, the instructors' materials seem to provide learners with PP-based structures, echoing those found in the written AFL sub-list and in academic prose.

In the textbooks bundles list, in the case of PP-based sub-structures, the form *other prepositional phrase* incorporates a wide range of bundle types with different frequencies (e.g. *from the text in* [25 occurrences], *in the text in* [23 occurrences], *in the United States* [23 occurrences], *on the other hand* [18 occurrences], *in bold in the* [14 occurrences], and *in the correct order* [9 occurrences]). In addition, in the structure *prepositional phrase with embedded of-phrase*, the textbooks include a small range of bundle types in this sub-category, such as in *as a result of* (34 occurrences), *at the end of* (27 occurrences), *at the beginning of* (16 occurrences), and *as part of a* (10 occurrences).

The research by Biber *et al.* (2004: 382), Biber (2006), Coxhead and Byrd (2007), and Hyland (2008a, 2008b) strongly confirms that lexical bundles in academic prose are phrasal rather than clausal. Although the textbooks in the present study have similar bundle percentages to those in the written AFL sub-list in the sub-category *other prepositional phrase*, a closer examination revealed that most types of bundles found in this sub-category are less likely to be identified as the most common in academic prose, such as *from the text in* and *in the text in*. The structural characteristics of the PP-based bundles in the textbooks may appear similar to those in academic prose, but the bundles found in these sub-categories appear to support the findings of previous studies (Jones and Haywood, 2004; Koprowski, 2005), demonstrating that textbooks not only present other and different bundles to EAP learners, as mentioned above, but they use a relatively small number of academic bundles. This mismatch in the bundles found in these structures could be related to genre differences between textbooks texts and the texts included in the written AFL corpus (see Chapters 3 and 4). The former is considered to comprise instructional-based texts while the latter is research and article-based. Furthermore, as suggested previously, textbook authors usually refer their learners to supplementary materials provided on the Cambridge Academic

English webpage, to complement their teaching materials, in particular offering information for learners on academic phrases.

8.1.3 NP-based structures

In the instructors' materials bundles list, in regard to NP-based sub-structures, *noun phrase with of-phrase* accounts for the highest percentage of bundles (18%), which is greater than the percentage of bundles in the written AFL sub-list (10%) (Figure 9.4). The *noun phrase with other post-modifier* structure, in contrast, has a higher proportion of bundles in the written AFL sub-list (9%) than in the instructors' materials bundles list (4%) (Figure 8.4).

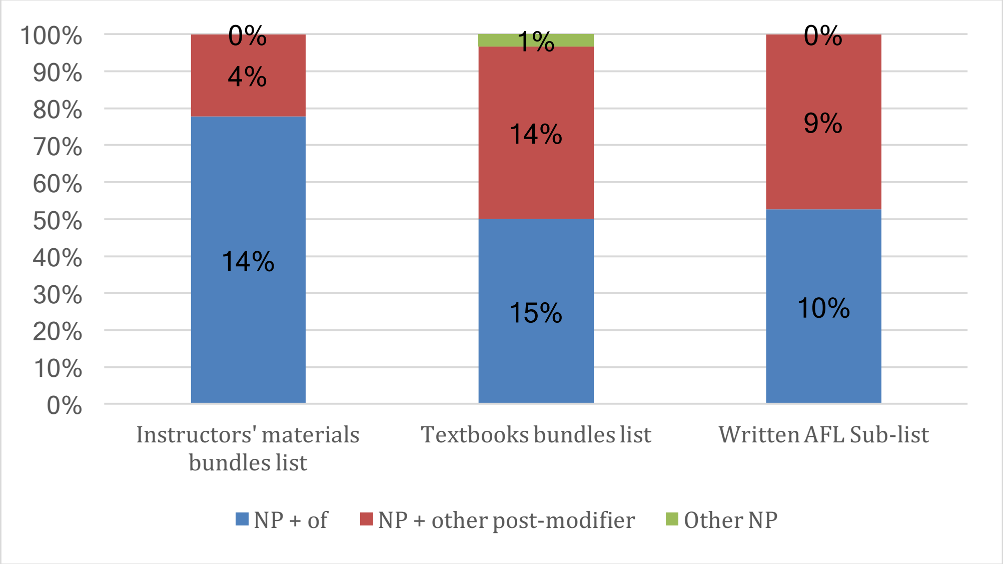

Figure 8.4. Distribution of lexical bundle types of NP-based sub-categories in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

According to research by Biber *et al.* (2004), Biber (2006), Coxhead and Byrd (2007), and Hyland (2008a, 2008b), academic prose is more phrasal than clausal, which is an important finding repeatedly mentioned in this chapter; this includes NP-based structures. The instructors' materials bundles list includes notable bundles within the sub-category *noun phrase with embedded of-phrase* structures (see Table 5.2 in Chapter 5). As with PP-based structures, the presence of these *noun-based* bundles in the instructors' materials may be due to the inclusion of academic reading text passages provided by instructors as supplementary materials in their writing classes. According to Hyland (2008b), these NP-based patterns occur reasonably often and widely in research writing. This argument is supported in the current study by the fact that some of the instructors' materials include academic reading articles created by professional authors that made use of these patterns. For example, the bundle *the end of the* with (*noun phrase with embedded of-phrase*) was found in the articles "On self-promotional I and we in academic writing" and "Brief History on Sheffield".

In addition, from Appendix F and Table 5.2 in Chapter 5, it can be observed that there no shared bundles in these sub-categories between the written AFL sub-list and the instructors' materials bundles list. This structure includes bundles such as *the end of the* (12 occurrences), *the extent to which* (9 occurrences), *a great deal of* (5 occurrences), and *the fact that the* (4 occurrences) under the structures *noun phrase with of-phrase* and *noun phrase with other post-modifier.* Although these bundles are not found in the written AFL sub-list, they are commonly used in academic writing (Biber *et al.*, 1999).

In terms of the textbooks bundles list, the authors make use of *noun phrase with of-phrase* and *noun phrase with other post-modifier*, which account for 15% and 14% (see Figure 8.4) of the total, respectively, which is slightly higher than the percentages of these bundles in the written AFL sub-list (accounting for 10% and 9%, respectively). In addition to the verb-centred structures found in the textbooks bundles list, this initial finding may indicate that textbooks make use of NP-based structures. According to Biber *et al.* (2004) and Coxhead and Byrd (2007: 134), "academic prose is noun-centric rather than verb-centric, such writing is not just made up simply of nouns but particular kinds of nouns combined with particular kinds of verbs and used with a range of other grammatical features expected by members of that discourse community".

Through reading concordance lines, a similar finding to that associated with *VP-based* structures emerged, that is, these two *NP-based* structures are widely used in textbooks to direct learners' attention and to call on learners to complete certain writing tasks. The types of bundles under the structures of *NP-based* bundles (e.g. *the subject of the*, *the words in the*, and *essay with a title*) are different from those found in the written AFL sub-list and in academic writing. This is confirmed in studies such as Wood (2010), Chen (2010), and Wood and Appel (2014). Even when certain bundles are cross-referenced between studies (e.g. *the end of the*), textbook authors seem to employ them differently. Most of these bundles are used to provide instructions to learners. For example, Table 8.1 provides a list of bundles that contain instructional language offered by Wood (2010), which are different from the academic bundles used in Biber *et al.* (2004), Cortes (2004), and Hyland (2008b).

Table 8.1. Lexical bundles in Wood's (2010: 97) instructional sub-corpus

at the end of	***do you think the***	*this part of the*
the meaning of the	***the end of the***	*scan the reading to*
the words in the	***in the United States***	*in the box below*
of the reading resources	*of the words in*	*you listen to the*
the rest of the	***what do you think***	*to answer the following*
with the rest of	*in the following sentences*	*of the underlined words*
answer the following questions	*guess the meaning of*	*is the main idea*
at the beginning of	*answer the questions that*	*and the numbers of*

In Table 8.1, the ten bundles highlighted in bold, identified by Wood (2010), also appear in the EAP textbooks materials bundles list in the present study (e.g. *the end of the*, *the meaning of the*). As previously stated, these bundles are used by authors to provide added information to guide learners in different tasks. In addition, PP-based bundles such as *at the beginning of* and *in the United States* and VP-based bundles such as *answer the following questions* in the textbooks bundles list are used in the same manner. As stated in the literature review chapter (see Chapter 2, section 2.1.1), according to the *idiom principle,* language users rely on ready-made sequences that they have used or produced before. Thus, the similar choices of sequences in Wood (2010) and the EAP textbooks suggest that authors may employ constructions that were used formerly in the process of forming instructional language for use in EAP coursebooks. In addition, from an entrenchment process perspective, the repeated usage of these given linguistic structures could result in these instructional bundles employed by textbook authors being processed as a holistic unit (see Chapter 2, section 2.1.2). This indicates that constructions that are used regularly eventually become more entrenched than those constructions that are used rarely (see Chapter 2, section 2.4.3).

In agreement with similar results described in previous studies (e.g. Biber *et al.*, 2004; Sánchez, 2014), NP-based and PP-based structures together make up almost half of the bundles in the textbooks corpus. This means that, in the case of *NP-based* structures, *noun phrase with of-phrase* (e.g. *the subject of the* [17 occurrences] and *the end of the* [16 occurrences]) and the *noun phrase with other post-modifier fragments* (e.g. *the words in the* [29 occurrences] and *essay with a title* [28 occurrences]) not only include varied bundles but also contain high-frequency bundles. These two structures are commonly found and reported in other studies such as Biber (2006), Jablonkai (2010), and Sánchez (2014). However, textbook authors employ PP-based structures differently to those found in the written AFL sub-list. Textbooks rely heavily on PP-based structures, including bundles mostly relevant to EAP language such as *from the text in*.

To conclude this section, textbooks rely heavily on *VP-based* structures. Textbooks, in a similar vein to academic prose, incorporate some NP-based and PP-based forms. However, even when textbook writers use NP-based and PP-based structural bundle classifications, they do not use or represent these structures in accordance with those used in the written AFL sub-list and academic prose. Hence, EAP learners encounter bundles with structural forms that are different from those found in academic writing.

8.1.4 Other expressions

The structure *other expressions* accounts for a lower percentage of bundle types than VP-, NP-, and PP-based structures because it includes expressions that "do not fit neatly into any other categories" (Biber *et al.*, 1999: 1024). The classification *other expressions* accounts for the highest proportion of bundles in the instructors' bundles list (15%) compared to lower proportions in the written AFL sub-list (10%) and the textbooks bundles list (5%). In each of the three lists used in this research, the sets of four-word bundles under each corpus differ from one another (see Table 5.2 in Chapter 5, Table 6.2 in Chapter 6, and Appendix F).

There are two possible interpretations relating to these differences in terms of proportions and structural categories. The issue of the topics of the different texts used and the authors' preferences in using bundles (in each corpus) may be key determiners behind this varied structure. To illustrate, textbook authors used the bundle *common in academic writing*, while instructors' materials include the bundle *as opposed to only*, and the written AFL sub-list includes the bundle *whether or not the.* For example, in the instructors' corpus, I further studied the bundle *as opposed to only,* which is used only four times. This bundle is found in the materials of two instructors who provided the same academic article to their learners, discussing plagiarism in Japanese universities. Textbook authors seem to prefer using the bundle *common in academic writing* to convince the learner/reader of a study tip linked to academic writing. This bundle appears in all three *Cambridge Academic English* series but not the EWB.

The usefulness and importance of such types of bundles (with other structures) in academic prose greatly varies from discipline to discipline, depending on the practice and usage of these bundles across different fields. Hyland's (2008b: 20) study showed that "writers in different fields draw on different resources to develop their arguments, establish their credibility and persuade their readers, with less than half of the top 50 bundles in each list occurring in any other list". In Hyland's 50 four-word bundle lists, across the four disciplines (e.g. electrical engineering, microbiology, and applied linguistics), over half the items in each list were not found at all in any other discipline. This means that there is considerable variation not only in the frequency but also in the structures of bundles across types of academic writing and across disciplines. In terms of the EAP materials examined here, the structures found in the *other expressions* category are often rare structures. Therefore, I would conclude that the bundles found in this classification have different and infrequent structures, which also may be viewed as different or unhelpful for learners to use in their writing classes.

8.2 Shared structures in instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

As reported in Chapter 7, four bundles are shared between the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list. From observing their structural classifications, it was found that *it is important to* and *it is possible to* has the structure *anticipatory it + verb phrase* classification. As for the bundle *on the other hand*, it has the pattern *other prepositional phrase* while the structure *passive verb phrase + to-clause* is the structure found in the bundle *can be used to*. The four shared bundles found and the types of bundles recognized within these sub-structures are a clear indication that EAP materials share three structures with the written AFL sub-list. However, it was observed above that EAP materials make use of NP- and PP-based structures commonly found in the AFL sub-list and academic writing but with different bundle types. This means that EAP materials employ different set of bundles when it comes to academic prose structures. Thus, the structural classification revealed that EAP materials (mostly textbooks) tend to misrepresent some important structures used in the written AFL sub-list and academic writing, and could implicitly be misleading learners and providing them with meaningless four-word bundles. However, as stated repeatedly, this interpretation is completely based on materials provided to this study; the EAP pre-sessional course and textbook authors could have accommodated these structures in bundles found in online supplementary materials.

8.3 Concluding remarks

To conclude, although four-word bundles in the instructors' materials bundles list and textbooks bundles list resemble the structures of bundles found in academic writing and make use of PP/NP-based patterns that are commonly identified in the written AFL sub-list, EAP materials rely heavily on VP-based patterns, which is not considered a defining characteristic of academic writing. The prevalence of VP-based bundles over NP- and PP-based bundles is an indication that EAP learners encounter sub-structures that are insignificant to EAP learners' academic writing. It is important to note that learners' exposure to NP- and PP-based structural patterns is very useful and important, allowing them to follow academic writing norms, helping them to build their academic lexical knowledge and to be able to express their thoughts by moving successfully from everyday language to more specialized and academic language. The structural practice of sequences in textbooks reinforces the fact that textbooks may not be effective in the treatment of lexical bundles (Chen, 2010; Wood, 2010; Wood and Appel, 2014). However, instructors' materials include PP-NP-based structures and bundles that may seem more relevant for EAP learners to encounter, particularly when constructing learners' composition, compared to those found in textbooks.

In the next chapter, I will provide functional comparisons of the overall and shared types of bundles in the instructors' materials bundles list, textbooks bundles list, and written AFL bundles list, reporting on their proportions and concordance lines when needed.

CHAPTER IX

FUNCTIONAL COMPARISON OF LEXICAL BUNDLES IN THE INSTRUCTORS' MATERIALS BUNDLES LIST, TEXTBOOKS BUNDLES LIST, AND THE WRITTEN AFL SUB-LIST

The functional classifications aspect of lexical bundles is a focal interest in formulaic sequences studies and an innovative strategic teaching approach in EAP materials development and language teaching. Hence, the aim of the present research was to profile the most frequent four-word bundles in an EAP pre-sessional programme in the UK, focusing on materials relating to teaching academic writing. This was accomplished by exploring textbooks and the unpublished materials of EAP instructors. Two EAP corpora were built to extract the lexical bundles, creating two EAP lists (see Chapter 4). In Chapters 5 and 6, I presented the results of the functional classifications. For comparison purposes, I conducted a functional classification (see Chapter 4 and Appendices D and E) on the modified version of the written AFL sub-list (Appendix C) used for the present study. The functional classifications are also important for addressing research question 3. I am interested in establishing the extent to which the bundles identified in the EAP materials are retrieved from corpus-driven lists such as the AFL (Simpson-Vlach and Ellis, 2010). Consequently, this chapter discusses the functional categorizations, comparing the overall and shared patterns of functions identified across the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list. I also discuss the treatment of overall functions of lexical bundles in EAP materials.

9.1 Overall functional comparison between lexical bundles in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

As can be seen from the results presented in Figure 9.1, *referential expressions* are the most frequently occurring of the three main discourse functions found in the instructors' materials bundles list and the written AFL sub-list, accounting for 57% and 54% of the total, respectively. In contrast, *stance expressions* occurs more frequently in the textbooks bundles list, accounting for 37% of the total, compared to close percentages in both the instructors' materials bundles list and written AFL sub-list. Moreover, *discourse organizing* bundles occur less frequently in the instructors' materials bundles list and the written AFL sub-list (accounting for 13% and 14% of the total, respectively) compared to the textbooks bundles list, where this function accounts for a higher proportion of bundles (29%).

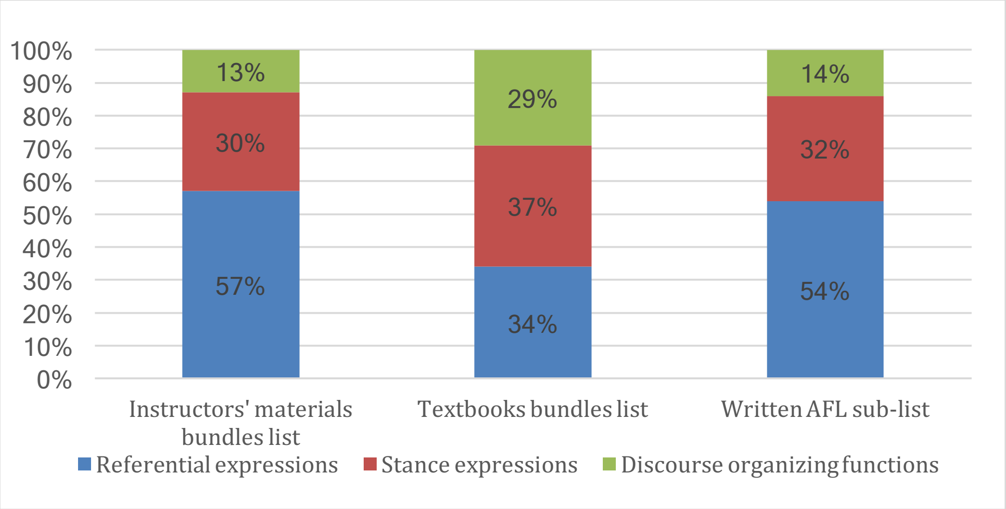

Figure 9.1. Overall functional distribution of lexical bundle types in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

The initial findings show that the instructors' materials bundles list has a similar distribution of functions to the written AFL sub-list while the functional distribution of the textbooks bundles differs from the other two lists. In addition, in contrast to the textbooks bundles, the instructors' materials bundles rely on functional types found in academic prose (see Biber *et al.*, 2004; Chen and Baker, 2010; Jablonkai, 2010). For example, Chen and Baker (2010) reported that native expert writings showed substantial use of *referential expressions* and less use of *discourse organizers*. On the whole, the main functions identified in the instructors' materials are closer to those found in academic prose than the functions found in the textbooks bundles.

In Simpson-Vlach and Ellis's (2010) research, *referential expressions* were marked as the largest functional category in their academic written list. In contrast, in Biber *et al.*'s (2004) study, textbooks and academic prose were found to have the smallest number of *referential bundles* compared to the extremely frequent number of *referential bundles* found in classroom teaching. However, Biber *et al.* (2004) also reported that *referential bundles* were the most common functional classification in academic prose, and more common than *stance* and *discourse* *bundles*. They indicated that classroom teaching combines the functional and communicative requirements of spoken discourse with the requirements of informational written discourse to organize and structure discourse. This means that classroom teaching (marked as a spoken register) heavily relies on two major functional patterns, *stance* and *referential bundles*, to determine and construct discourse. Moreover, the expressions used in academic prose from Biber *et al.*'s study and in the instructors' materials in the present study (both marked as written registers) are mostly structured with bundles functioning as *referential expressions*.

Similar to the structural analysis (Chapter 8), three important considerations relating to the types, content and number of texts provided by each EAP instructor (see Chapter 7) could have greatly influenced the results. Furthermore, conducting concordance checks helped to verify interpretations. The findings in this study suggest that the findings on the three features could have been influenced primarily by the differences in the disciplinary distribution of the resources that instructors provided in their handout materials. This refers to the different styles and linguistic choices that writers in different disciplines adhere to in order to convey meaning to their readers. For example, Instructor (1) provided 29 different texts, ranging from grammar exercises to varied topics from articles and academic journals. In terms of disciplinary variation within classes, Instructor (1) sought to cover topics such as "What is materials science?" and "Navigating our way through computer files uses the same brain structures as a dog finding its way". Instructor 2 provided eight texts, focusing on paraphrasing, quotations and referencing handouts, reading exercises, and two articles – one on plagiarism and the other on climate change (see Appendix K). Instructor 4 provided many handouts, including a passage on "Generic Engineering". As previously concluded, I would argue that this diversity in terms of the types, themes and number of texts used in the instructors' materials may have greatly influenced the functional classification of the use of *referential bundles*.

To achieve in-depth analysis, I will provide a comparative and detailed examination of the sub-functions. In doing so, I will reveal the percentages and types of bundles, together with in-context information derived through checking expanded concordance lines when needed, in each of the two corpora compared to the written AFL list.

9.1.1 A comparison of *referential expressions* in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

Figure 9.2 presents the percentage distribution of *referential expressions* sub-functions identified in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list. A comparison of the sub-functions of *referential bundles* across the three lists yielded valuable results.

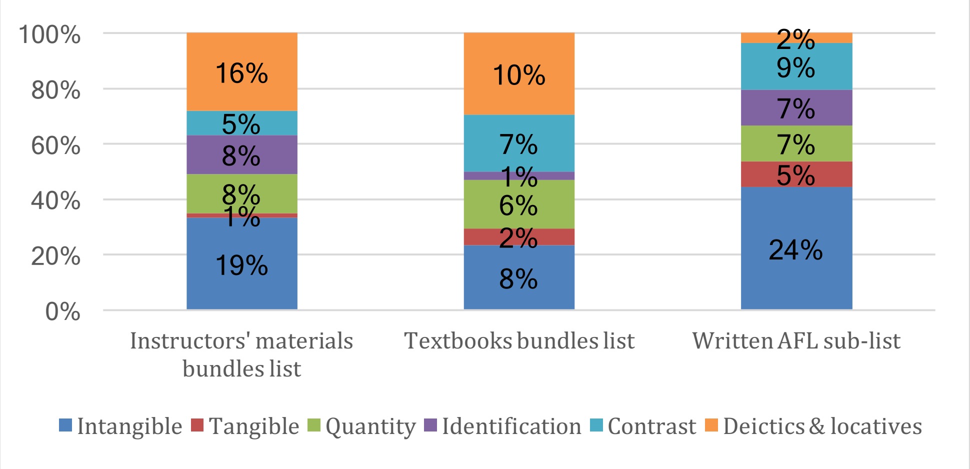

Figure 9.2. Percentage distribution of lexical bundle types in the sub-categories of *referential expressions* across the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

Intangible framing bundles

Intangible framing bundles: Within the two lists (the written AFL sub-list and the instructors' materials bundles list), the vast majority of *referential expressions* concern the *specification of attributes* (*intangible framing attribute*). However, the textbook authors use a small number of bundles with this type of sub-function (8%), compared with the instructors' materials bundles list (19%), while the proportion of bundles with this type of sub-function is the highest in the written AFL sub-list (24%). Comparing the bundles in this sub-category across the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list revealed differences in the types of bundles identified (see sections 5.3 and 6.3 and Appendices D and E for a full list of the functional bundles). Table 9.1 lists a few examples of lexical bundles found in the *referential expressions* sub-category across the three lists.

Table 9.1. Some examples of bundles in the *referential expressions* sub-category in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

Number	Instructors' materials bundles list	Textbooks bundles list	Written AFL sub-list
1	Intangible framing		
	the extent to which *in the development of* *with the help of*	*in a way that* *the subject of the* *the way in which*	*in the course of* *in accordance with the* *such a way that*
2	Tangible framing		
	as part of the *the rest of the*	*the rest of the* *as part of a*	*an increase in the* *over a period of* *as part of the*
3	Deictics and locatives		
	in the United States *at the end of* *the end of the*	*essay with a title* *at the end of* *in the United States*	*at the time of*
4	Contrast and comparison		
	on the other hand *as opposed to only* *the relationship between the*	*on the other hand* *with a similar meaning* *the relationship between the* *from the same family* *the advantages and disadvantages of*	*be related to the* *the other hand the* *the differences between the*
5	Quantity		
	a great deal of *as one of the* *there are a number* *a third of the*	*a wide range of* *one of the most* *be more than one* *each of the following*	*a large number of* *in a number of* *there are a number*
6	Identification and focus		
	to be one of *of different types of* *of his or her* *that he or she*	*used to refer to*	*as can be seen* *can be seen in* *that there is no*

As can be seen in Table 9.1, *intangible framing* bundles include a variety of bundles across the three lists. For example, the written AFL sub-list includes bundles such as *in the course of*, *in accordance with the*, and *such a way that*, which appear to be different to the bundles found in the instructors' materials bundles list and the textbooks bundles list. Depending on the context in which the bundles appear, each of the two lists (the instructors' materials bundles list and the textbooks bundles list) has its own distinctive set of bundles, as shown in Table 9.1. Regarding *intangible framing* bundles, for example, the instructors' materials bundles list includes bundles such as *the extent to which*, *with the help of*, and *in the development of.* These *intangible framing* bundles serve as devices that frame concrete entities and abstract concepts or categories. The *intangible framing attributes* sub-category found in the EAP instructors' materials are similar to those found in Simpson-Vlach and Ellis's (2010) research, where this sub-function was regarded as the most frequent pragmatic sub-function among all the sub-categories. It is important to note that Simpson-Vlach and Ellis's (2010) study indicated that the bundles within the *intangible framing devices* sub-category are clearly classified as important academic phrases. Similarly, most bundles in the *referential expressions* have *intangible framing* functions in the English European Union Documents (EEUD) Corpus (Jablonkai, 2010).

Checking concordance lines showed that the instructors' materials bundles consist of a similar set of bundles to those found in academic writing, serving *intangible framing* functions but are also different from those found in academic writing (see Biber *et al.*, 1999). The following are two examples, one which is used as an academic bundle, *the extent to which*, while the other, *with the help of*, is not used in academic prose and was found in a passage regarding "Drugs and Olympic Games".

(a) The purpose of this paper is to discuss ***the extent to which*** higher education does promote these abilities which can be summarised as independent learning. (Instructor 3)

(b) …of athletes who have set new records have done so ***with the help of*** performance-enhancing drugs. (Instructor 4)

The textbooks bundles list also has a different set of bundle types, as shown in Table 9.1. Examining the corpus in more detail showed that textbook authors use *intangible framing expressions* such as the academic bundle *in a way that*, which is used within a reading passage on "Communication" or to provide teaching information to learners, while the bundle *the subject of the* is found in only instructional parts, as shown in the following examples:

(a) Are the ideas grouped ***in a way that*** will correspond to paragraphs? (AEPC Extended Writing Booklet)

(b) Information is presented ***in a way that*** will make potential customers want to come back to the destination. (CAE-Intermediate)

(c) Teachers use advertising texts, so 'teachers' is ***the subject of the*** –ing clause. (CAE-Advanced)

From a metadiscourse point of view, according to Hyland (2010: 128), in his *interactive model* (see Chapter 2), instructors' materials and textbook authors use these bundles as devices to "anticipate readers' knowledge and reflect the writers' assessment of what needs to be made explicit to constrain and guide what can be recovered from the texts".

From the data presented, it can be seen that EAP learners are exposed to *intangible framing* bundles, which are mostly found in the instructors' materials (e.g. *the extent to which*) (see Chapter 5, section 5.3). As previously mentioned, according to Simpson-Vlach and Ellis (2010), the majority of *intangible framing attributes* appear in the AFL list, indicating that these bundles are important academic phrases. However, the bundle *the way in which* is found among the most commonly used bundles in academic prose (Biber *et al.*,1999). Thus, explicit instructions by EAP authors on the teaching of lexical bundles should be dealt with pedagogically, and through corpus-based materials (Burton, 2012).

It seems that EAP instructors are unfamiliar with research on lexical bundles and are "unsure of how to address word-learning needs" (Zimmerman and Schmitt, 2005: 1). In addition, learners' corpora provide useful insights into students' treatment of bundles, which could help us understand and make a connection with the present study. In their learner corpus, Chen and Baker (2010) found that university students did not use the bundle *the extent to which*, but native speakers did. Additionally, Cortes (2006) reported that prior to teaching lexical bundles to history students, these students had never used expressions such as *the extent to which* in their writing. However, after formally introducing lexical bundles to students, Cortes (in her micro-lessons) found that one student started using this bundle.

Deictics and locatives bundles

Deictics and locatives bundles: The written sub-AFL consists of a small proportion of bundles in the *deictics and locatives* sub-category (2%) compared to the larger proportions found in the instructors' materials bundles list (16%) and textbooks bundles lists (10%) (Figure 9.2). Unlike in Simpson-Vlach and Ellis's (2010) study, where the number of *deictics and locatives* bundles was found to be small, the *deictics and locatives* devices in the present study constitute the second most frequent pragmatic sub-function among all the sub-categories in the instructors' materials bundles list and the most frequent in the textbooks bundles list (see Figure 9.2). However, Simpson-Vlach and Ellis (2010) noted that *deictics and locatives* bundles are an important functional sub-category. These *referential expressions* refer to a particular place or location in the text itself or to a physical location in the environment. Their importance for EAP learners derives from their communicational function, showing interaction with readers and engaging readers in *textual acts* by referring to a certain time, place, or text (Hyland, 2002; Chen, 2010). In addition, from a metadiscourse perspective, the importance and usefulness of such expressions lies in the notion that these bundles "help relate a text to its context, taking the readers' needs, understandings, existing knowledge, prior experiences with texts, and relative status into account" (Hyland and Tse, 2004: 161).

For example, Table 9.1 shows that in the *deictics and locatives* sub-function in the written AFL sub-list, only one bundle type (*at the time of*) is found. In contrast, in the instructors' materials bundles list and textbooks bundles list, two highly frequent bundles such as *in the United States* and *at the end of* are found within this sub-category (see Chapters 5 and 6). In terms of the bundle *in the United States* in the instructors' materials corpus, this bundle is mostly found in reading passages. In the textbooks corpus, the bundle *in the United States* is found in both reading passages and to provide instructions to learners, as can be seen in the following examples:

(a) Soccer is also popular ***in the United States*** now, although football is still more popular. (Instructor 4)

(b) Before you read, look at the results from a survey of the working population ***in the United States***. In pairs…(CAE-Upper Intermediate)

(c) ***In the United States***, the deflation of employment, output, and prices continued until March 1933, the lowest point of depression. (CAE-Intermediate)

In the case of the bundle *in the United States* (18 occurrences), which is the most frequent bundle, it may have appeared in the instructors' materials bundles list because of the inclusion of reading articles and passages discussing the United States regarding a certain matter. As stated above, the bundles in the instructors' corpus are drawn from articles and reading passages, which could be taken from a variety of original sources, exploring a wide range of themes. This explains the presence of the bundle *in the United States* in studies such as Biber *et al.* (2004) and Biber *et al.* (1999). Thus, I would argue that the inclusion of these articles and reading passages by instructors has strongly impacted the bundles having similar functions to those found in academic prose in general.

As for the bundle *at the end of*, concordance line checks indicated that the *deictics and locative* bundle appears in both reading passages and to instruct learners, in both the instructors' materials and textbooks corpora. Such bundles serve as connectors that link information presented in different parts of the texts. In addition, this bundle is considered a multifunctional bundle, which makes time and place references. Unlike in the present study, multifunctional references bundles were missing from the English for Specific Purposes textbooks corpus but were found in the English Engineering Introductory Textbooks corpus examined by Chen (2010). However, it is important to note that this bundle is mostly found in instructional sections; authors tend to use it to refer the reader/learner to a particular time or to a particular point in the environment or textual location (Simpson-Vlach and Ellis, 2010).

(a) A series of important discoveries were made ***at the end of*** the century. (Instructor 4)

(b) The rise in student fees begins ***at the end of*** the year. (CAE-Advanced)

(c) ***At the end of*** writing you will have a list of references. Record the source as follows. (Instructor 2)

(d) Answer the following questions about your level and study goals. ***At the end of*** each question (in brackets) there are some suggestions to help you write your answers. (CAE-Intermediate)

Given that the *deictics and locatives* sub-category includes the highest percentage of *referential expressions* (e.g. *at the end of*), it is not unexpected to see that textbooks include this bundle type. The results obtained in my corpus for *referential expressions* in general and *deictics and locatives* in particular coincide with the results found by Biber *et al.* (2004), where text *deixis and locative* bundles were commonly found in written registers such as in textbooks. Although bundles such as *at the end of* and *in the United States* are included in Biber *et al.'*s (2004) corpus, there are more bundles in the present corpus that can be interpreted as textbook instructional language such as *essay with the title*, *the written academic corpus*, *scientists and their work*, but are not mentioned in Biber *et al.*'s research (see Chapter 6, section 6.3). Simpson-Vlach and Ellis (2010) listed the bundle (*a and b*), which functioned as part of the *deictics and locatives* sub-category in their corpora.

Similarly, the bundle *essay with a title* is among the *deictics and locative* expressions found in textbooks, mostly appearing in instructional parts. This bundle is used by authors to refer to a temporal reference point in the discourse (see example).

(a) After a seminar, you have been given an ***essay with a title*** "Discuss the influence of advertising on consumer behaviour". You decide… (CAE-Advanced)

It seems that this type of sub-function includes bundles associated with EAP language. This means that in EAP writing materials, it is common to find bundles with words and vocabulary associated with academic writing (e.g. *essay*, *dissertation*, and so on). This may be due to the nature of EAP textbook language, whereas EAP learners need to encounter additional and diverse types of academic bundles which are presented and taught according to the written academic prose or the written register (Coxhead and Byrd, 2007; Hyland, 2008b).

In addition, Chen (2010) found that an ESP textbooks corpus lacked *text deictics and locatives* bundles, which was not the case for the electrical engineering introductory textbooks corpus. Looking into learners' corpora, Wei and Lei (2011) found that advanced learners and professional authors used the same amounts of research-oriented bundles, including bundles such as *at the end of*. In contrast, Cortes (2006) stated that students rarely used the bundle *at the end of* in their writing production in the history class she taught. Students, most of the time, used another strategy to introduce dates or events in their writing which was different from history. Instead, these students used the exact date, repeatedly, every time they wanted to refer to it. This shows that it is important to highlight bundles in this sub-functional category to learners. The following is an example from a student's production, according to Cortes (2006: 400):

> In fact, *by March of 1947*, it appeared that, from an American perspective, they were only two ways to live in the world. *In March of 1947*, President Harry S. Truman announced to Congress the "Truman Doctrine".

Tangible framing bundles and contrast and comparison bundles

Tangible framing bundles: In comparison to *intangible framing bundles*, *tangible framing expressions* account for the smallest proportion of bundle types in the instructors' materials bundles list (1%), and likewise include a small number of bundle types in both the textbooks bundles list (2%) and the written AFL sub-list (5%) (Figure 9.2). Table 9.1 presents different types of bundles among the three lists, in relation to the pattern *tangible framing expressions*. EAP learners need these *tangible framing* devices to refer to physical or measurable attributes of the subsequent noun. The written AFL sub-list includes bundles such as *an increase in the*, *over a period of* and *as part of the*, while the textbooks bundles list includes bundles such as *the rest of the* and *as part of a*, both occurring in textual and instructional parts as shown in the following examples:

(a) Report your ideas back to ***the rest of the*** class. (CAE-Upper Intermediate)

(b) The socioeconomic status of the region is much lower in many categories than ***the rest of the*** province. (CAE-Advanced)

(c) They are usually treated ***as part of a*** tradition of pictorial art. (CAE-Upper Intermediate)

(d) ***As part of a*** Business Studies course, you attended a lecture with the title Innovation and invention. You found it hard to follow, but you noted down as much as you could. Work in pairs… (CAE-Advanced)

In the present study, the bundle *the rest of the* functions as a *tangible framer*, to refer to physical or measurable entities (see example (a) above). However, Biber *et al.* (2004) reported that the bundle *the rest of the* is categorized as a *referential expression*, with a *quantity* sub-function. The various sub-functions of an individual bundle (e.g. *the rest of the*) serves different functions in different contexts. As mentioned in the literature, this issue is often referred to as indicating the multifunctionality of bundles (Biber *et al.*, 2004; Ädel and Erman, 2012; Salazar, 2014). It is an important issue that needs to be addressed from a pedagogical perspective. This issue is fully explained in the literature review (Chapter 2) as well as in the limitations in the conclusion chapter (Chapter 11). However, here I will highlight the bundle *the rest of the* and how this bundle should be presented to learners, in terms of its multiple functional characteristics. Biber *et al.* (2004) and Chen and Baker (2016) emphasized that when classifying a bundle, it is important to examine its extended context to establish what the main function is, as stated in the literature review. In the EAP materials, examining the extended context of the bundle *the rest of the* showed that it serves two different referential expression sub-functions: (1) *quantity* (similar to Biber *et al.*, 2004) and (2) *tangible framing bundles*, as shown in the following sentences:

(a) *As tangible framing bundles*: Report your ideas back to ***the rest of the*** class. (CAE-Upper Intermediate)

(b) *As quantity*: The socioeconomic status of the region is much lower in many categories than ***the rest of the*** province. (CAE-Advanced)

The multifunctionality of bundles such as *the rest of the* reveals an important point. Depending on the textbook or instructor's materials that each EAP learner is given during their writing class, some EAP learners may come across the bundle as a *tangible framing bundle* while other students will encounter it as a *quantity* bundle. For example, upper intermediate learners may come across the bundle *the rest of the* as a *tangible framing bundle* in the CAE-Upper Intermediate textbook. Other EAP students could be exposed to the same bundle, but as a *quantifier expression* in the CAE-Advanced textbook. As a result, not all learners will be exposed to the same context and sub-functions. This is not to argue whether this useful or not, but to address the difficulties and complications that we need to consider in teaching or developing materials on formulaic sequences (see Chapter 11 for implications).

Moreover, in the context of learners' corpora, Cortes (2004) reported that the bundle *the rest of the* was used in history students' writing, serving a variety of functions. Cortes (2004) also found that students at different academic levels used bundles such as *the rest of the*, which had different functions to those found in the writing of published history writers. She argued that when using such bundles, students were perhaps being more creative in their types of writing. Through the use of different functions they conveyed more creative styles than was typical of academic writing, as explained in the literature review.

Contrast and comparison bundles: In terms of *contrast and comparison* expressions, Figure 9.2 indicates that the instructors' materials bundles list includes a slightly lower proportion (5%) of this type of expression compared to the textbooks bundles list (7%) and the written AFL sub-list (9%). The information presented in Tables 5.4, 6.4, and Appendix D also shows that there are some differences in bundle types across the three lists. For example, the written AFL sub-list includes the following bundles: *be related to the*, *the other hand the*, and *the differences between the*, which are clearly different to the bundles found in the textbooks bundles list (e.g. *with a similar meaning*, *what is the difference,* and *on the other hand*), and also the instructors' materials bundles list (e.g. *as opposed to only*, *the relationship between the*, and *on the one hand*) (see Table 9.1). Several of these *contrastive and comparative expressions* consist of explicit markers such as *same*, *different*, or *relationship*, to provide a comparison between items. Unlike the present study, Biber *et al.* (2004) classified bundles such as *on the other hand* as explicit markers, under *discourse organizers* instead of *referential expressions*.

The *contrastive* bundle *the relationship between the* is found in both the textbooks bundles and instructors' materials bundles lists. The bundle *the relationship between the*, in the textbooks corpus, however, is used only in instructional parts while in the instructors' materials corpus, it is used in both textual and instructional (see the following examples):

(a) Explain ***the relationship between the*** sentences and what comes before? (CAE-Advanced)

(b) One technique which can be used is to change ***the relationship between the*** ideas in the sentence. (Instructor 2)

(c) The price is dependent on ***the relationship between the*** supply and demand for buns. (Instructor 3)

Based on the functional classification of metadiscourse described in Chapter 2, this sub-function is called *transitions*, which aids the reader to make connections between thoughts in the academic discourse, projecting contrastive or comparative connections between ideas. (Hyland 2005, 2010). In addition, Khedri *et al.*'s (2013) study found *transition markers* were used with significant differences in applied linguistics and economics, regarding the way in which writers organized their argument through the use of these expressions to guide readers to their proposition.

Quantity bundles

Quantity bundles: Figure 9.2 shows that the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list contain comparable proportions of *quantity expressions*, which account for between 6% and 8% of the total. However, across the three lists, *quantity expressions* include different types of lexical bundles (see Table 9.1). These bundles specify quantity, by including either cataphoric expressions such as counting or specifying quantities of a following noun phrase. Other bundles within this sub-functional category describe the size, form and abstract characteristics of the following head noun (Biber *et al.*, 2004). For example, bundles such as *a large number of* and *in a number of* appear in the written AFL sub-list (see Appendix D) while the instructors' materials bundles list includes various types of bundles: *as one of the* and *a third of the* (see the following examples).

(a) Although plagiarism is considered among western academic circles ***as one of the*** worst "crimes" student can commit, many scholars suggest that these attitudes do not apply to students from areas outside this sphere. (Instructor 3)

(b) The survey showed 725 people, about ***a third of the*** population were 'not able to live without the charity of their neighbours.' (Instructor 1)

Among the bundles found in the textbooks bundles list are *one of the most*, found in instructional and textual parts, and the bundle *be more than one* appears in instructional parts, as shown in the following examples.

(a) ***One of the most*** important aspects is making use of ideas of other people. (AEPC Extended Writing Booklet)

(b) CNN set up in 1995 is ***one of the most*** popular news websites in the United States. (CAE-Intermediate)

(c) There may ***be more than one*** possible answer. (CAE-Upper Intermediate)

In addition, *a great deal of* is an academic *quantity* expression that is found in the instructors' materials bundles list and textbooks bundles lists. In the instructors' materials corpus, the bundle *a great deal of* occurs in textual and instructional parts while in the textbooks corpus, this *quantity* bundle only appears in textual parts (see the following examples).

(a) This person understands you and you have ***a great deal*** of respect for them. (Instructor 3)

(b) (92) The relationship between cause and effect is central to ***a great deal of*** academic writing, and may be expressed in a number of different ways. (Instructor 4)

(c) ***A great deal of*** research has been conducted on how birds fly. (CAE-Upper Intermediate)

The EAP materials and the written AFL sub-list contain a similar number of *quantity bundles* but different types of bundles (see Chapter 5 and 6). In addition, the EAP materials include bundles commonly found in academic prose (e.g. *a wide range of* and *a great deal of*; Biber *et al.*, 1999). Chen (2010) reported that in the electrical engineering introductory textbooks examined, high percentages of *quantity* specification bundles were included. However, unlike the present study, the ESP textbooks examined seem to lack or underrepresent this sub-category. In a learner corpus study, Chen and Baker (2016) confirmed the misuse of conversational and academic writing phrases. The study found that students at lower levels overused and misused *quantifying* bundles. For example, they reported that learners at lower levels included bundles that incorporated speech markers such as *a lot of* and *many* and tended to use these markers with "existential *there* constructions", forming bundles such as *there are a lot of* and *there are too many*. Chen and Baker (2016: 871) interpreted this as a sign of a "colloquial tone" in the writings of lower-level learners. It can be concluded that in some instances, textbook authors provide useful information on lexical bundles, when dealing with writing conventions.

Identification and focus bundles

Identification and focus bundles: Figure 9.2 shows that the instructors' materials bundles list and written AFL sub-list contain similar proportions of *identification and focus bundles* (8% and 7%, respectively), while the equivalent proportion is 1% in the textbooks bundles list. It should be noted that the written AFL sub-list (modified) has relatively small proportions of four-word bundles compared to the written AFL top 200 list of three-, four- and five-word bundles. Simpson-Vlach and Ellis (2010) stated that *identification and focus* is considered the second most common functional category in *referential expressions.* This means that the AFL makes a large use of this sub-function; however, it seems that 7% of the total comprise four-word bundles. In addition, the types of bundles in the written AFL sub-list (e.g. *as can be seen* and *can be seen in*) are different compared to those found in the EAP materials (e.g. *to be one of*, *used to refer to*, and *of his or her*; see Table 9.1).

Exposing EAP learners to *identification and focus* bundles helps them to understand how to provide exemplification and identification, which is considered a basic pragmatic sub-function in academic writing (Simpson-Vlach and Ellis, 2010). However, the EAP materials include different types of *identification and focus* bundles; not only are they different from those found in the written AFL sub-list but also from those found in academic writing in general (see Biber *et al.*, 1999, 2004). For example, the EAP materials (namely instructors' materials) use a bundle such as *to be one of* to make additional information salient and available to the reader to focus on. As for locational results, the bundle *to be one of* is located in the textual parts while the bundle *of his or her* is found in the textual and instructional parts of the instructors' materials corpus (see examples).

(a) Japan is often believed ***to be one of*** these countries in which plagiarism is not considered a moral transgression. (Instructor 3)

(b) A learner who is aware ***of his or her*** learning style will be more successful at learning independently. Discuss. (Instructor 1)

(c) It is difficult to measure a leader's success without taking into account the state of the business at the time ***of his or her*** appointment. (Instructor 4)

In comparison, the textbooks bundles list includes one bundle type in this sub-category, *used to refer to*, which is a typical expository phrase. This bundle occurs in both textual and instructional parts (see examples):

(a) It is also ***used to refer to*** how a picture 'frames' a scene, and how a newspaper 'frames' a story. (CAE-Intermediate)

(b) The words from the texts in 2.2 in the box are ***used to refer to*** similarity and difference. (CAE-Upper Intermediate)

Biber *et al.'s* (2004: 394) study reported that *identification and focus* bundles are often found in classroom teaching. For example, the bundle *those of you who* identifies a certain group of students who are in focus: "For *those of you who* came late I have the, uh, the quiz". Similarly, the above sentences include *identification and focus* bundles and are used in the manner described by Biber *et al.* (2004). Unlike the present study, Simpson-Vlach and Ellis (2010: 504) identified different *identification and focus* bundles, including *such as the*, which is found in both academic speech and writing (e.g. "so many religions, *such as the* religion of Ancient Egypt, for instance…"). From a metadiscourse perspective (see Chapter 2), these types of *interactive resources* are called *code glosses*, functioning to help readers understand the meaning of ideational information (Hyland and Tse, 2004; Hyland, 2010).

9.1.2 A comparison of *stance expressions* in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

Obligation and directives bundles and Ability and possibility bundles

Obligation and directives bundles: A comparison of lexical bundles in the *stance expressions* sub-category reveals that the proportion of bundles in the sub-function *obligation and directives* in the textbooks bundles list (20%) is much greater than the proportion of bundles in the written AFL sub-list, and is ten times greater than that found in the instructors' materials bundles list as shown in Figure 9.3. Figure 9.3 illustrates the sub-functions of *stance expressions* identified in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list, showing the proportional relationship of the overall functions. *Obligation and directives* bundles usually have verb forms directing the readers to perform an action, or recall or consider observations, facts or conclusions (Simpson-Vlach and Ellis, 2010). Table 9.2 lists some of the bundles found in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list (see Chapter 5, Table 5.1; Chapter 6, Table 6.1; and Appendix C for full lists).

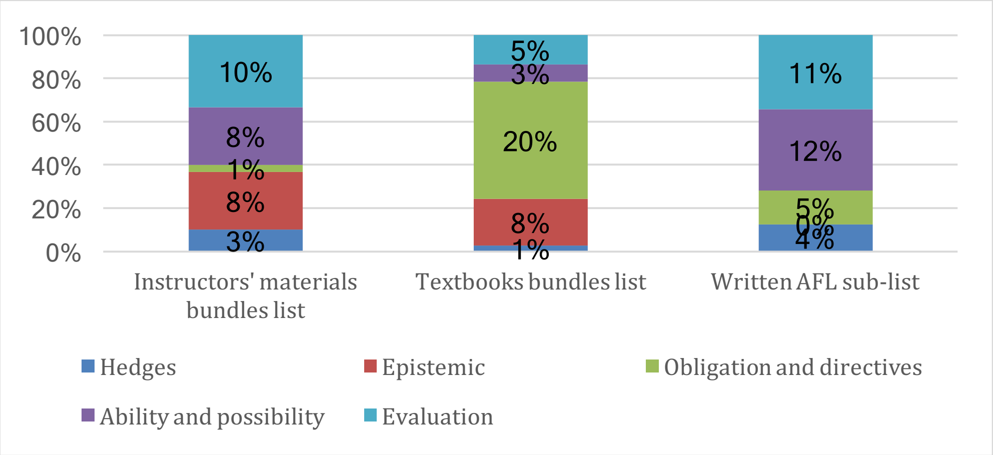

Figure 9.3. Percentage distribution of lexical bundles types in the sub-categories of *stance expressions* across the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

Table 9.2. Some examples of bundles in the *stance expressions* sub-category in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

Number	Instructors' materials bundles list	Textbooks bundles list	Written AFL sub-list
1	Hedges		
	are more likely *is more of a* *may or may not*	*are more likely to*	*is likely to be* *it is likely that*
2	Epistemic stance		
	is not considered a *have shown that the* *be aware of the* *be done by a*	*what do you think* *do you think the* *why do you think* *decide which of the*	
3	Obligation and directives		
	you may need to	*focus on your subject and answer the questions* *work in pairs and* *use a dictionary to*	*it should be noted* *take into account the* *to ensure that the*
4	Express of ability and possibility		
	it is possible that *it is quite possible* *can be used to* *can be used in*	*can you think of* *it is possible to* *can be used to*	*be used as a* *it is possible that* *can be found in* *can be used to*
5	Evaluation		
	it is important to *the best way to* *is important to note* *it is interesting that*	*it is important to* *common in academic writing* *is the most important*	*it is important to* *it is necessary to* *it is clear that* *it is obvious that*

In terms of the textbooks corpus, as reported in Chapter 6 (section 6.3.1), not only do most of these bundles have the highest frequency rates in the textbooks bundles list, *obligation and directive are* also the most used function in *stance expressions* and in the overall bundles in the textbooks (see Figure 9.1). Table 9.3 presents a list of *obligation and directives* bundles taken from the textbooks bundles list;

these bundles were all found in the instructional parts after examining the bundles in context through concordance line checks.

Table 9.3. *Obligation and directives* bundles in textbooks

focus on your subject	*to check your answers*
and answer the questions	*complete the following sentences*
work in pairs and	*make notes on the*
use a dictionary to	*read the text again*
you are going to	*complete the sentences using*
answer the following questions	*inferring the meaning of*
answer the questions about	*check your answers in*
have been asked to	*discuss the following questions*
in your own words	*in the correct order*
make a note of	*to write an essay*

In terms of the textbooks bundles list, considering the nature of the data, it is not surprising to find that the bundles identified and classified in *stance expressions*, particularly under the sub-category *obligation and directives*, account for the highest proportion of bundle types. The sub-category includes bundles (e.g. *focus on your subject* and *work in pairs and*). Furthermore, it is to be expected that *obligation and directive* bundles appear at the top of the textbooks bundles list with high frequencies, given the fact that the linguistic features associated with instructional or teaching language are uniquely characteristic of EAP textbooks (Wood, 2010; Wood and Appel 2014). This means that the most distinctive features of EAP textbooks, as repeatedly mentioned, are a reliance on specific grammatical features associated with a set of bundles functioning as *obligation and directive* bundles. Such bundles are used by textbook authors to ease and accelerate the process of performance in different tasks for EAP learners. The presence of these types of bundles support the interpretation made by previous studies that textbooks rely heavily on functions that are considered classroom-based and instructional rather than the language found in academic prose (Wood, 2010). The following are two examples of *obligation and directives* bundles taken from the textbooks corpus:

(a) ***Focus on your subject***

You may see variations from the layout and punctuation conventions given here. (CAE-Advanced)

(b) Read the text ***and answer the questions*** which follow. (AEPC Extended Writing Booklet)

In addition, *obligation and directives* bundles in Biber *et al.* (2004: 390) seem to be associated with classroom teaching (e.g. *take a look at*). However, unlike the present study, many of the bundles in Biber *et al.'s* (2004) study in the *obligation and directives* function include a second person pronoun (*you*) (e.g. *you have to do*) (see example).

> "All *you have to do* is work on it." (classroom teaching).

Similar to the present study, Wood (2010) reported that the comprehensive intermediate-advanced English for Academic Purposes (EAP) textbooks and reading and writing skills-focused textbooks included similar *stance* bundles (e.g. *guess the meaning of* and *answer the following questions*). Although the present research examined EAP textbooks (such as the Cambridge Academic English series), which to my knowledge have not been analysed in previous research, the findings are similar to those obtained by Wood (2010). The results add insights to the treatment of lexical bundles in varied EAP textbooks.

From the data presented, it should be noted that most of the textbooks bundles in this sub-function are geared towards directing EAP learners/readers to perform different acts (e.g. *work in pairs and* and *use a dictionary to*). Chen (2010) established that one technique textbook authors employ in their use of bundles is to directly instruct readers to perform the next step. Furthermore, clearly, there seems to be a great deal of misrepresentation of the *obligation and directives* sub-function in the textbook materials. Previous studies on textbooks, such as Biber (2006) and Wood (2010), also revealed that these types of lexical bundles share the same characteristics used in classroom talk – conversation – to those found in written academic texts.

A major contrast can be observed, in terms of *obligation and directives bundles*, in the instructors' materials bundles list, which has an extremely low proportion of this type of bundle. To illustrate, the instructors' materials bundles list includes only one type of bundle in this sub-category, which is *you may need to* (see example). This includes a pronoun with a modal construction, indicating a regularly used fixed phrase.

(a) ***You may need to*** change the form of the verb. (Instructor 4)

(b) ***You may need to*** repeat your summary to the group. (Instructor 3)

Furthermore, unlike the high-frequency bundles *focus on your subject* and *and answer the questions*, the bundle *you may need to* is a low-frequency bundle, occurring three times per million words. This type of bundle is similar to those found in Biber *et al.'s* (2004) study, reported above. In the written AFL sub-list, different sets of bundle types such as *it should be noted* and *take into account* are found, serving as *obligation and directives*, which are missing from the instructors' materials bundles list and textbooks bundles list (see Table 9.2). From the perspective of metadiscourse, these bundles (e.g. *it should be noted*) explicitly engage the reader by referring to the reader directly or by building a relationship with him or her (Hyland, 2010).

Ability and possibility bundles: The data in Figure 9.3 also reveals that bundles of the sub-functions *ability and possibility* are more common in the written AFL sub-list (12%) than in the instructors' materials bundles list (8%). As seen in Table 9.2, the written AFL sub-list includes bundles such as *be used as a*, *can be found in*, and *it is possible that* (see Appendix D for the full functional list). These are considered *interactional resources*, involving the reader in the argument by withholding the writer's full commitment to the proposition, according to Hyland's 2010 model of metadiscourse. Instructors' materials use the *ability and possibility* bundles to introduce some suggestions or possible actions. Examples include the bundle *it is quite possible*, which is found in the textual parts, and *can be used in*, located in both textual and instructional parts (see examples).

(a) ***It is quite possible*** that he or she would receive a hundred different answers. (Instructor 2)

(b) …pronouns ***can be used in*** the act of citing oneself,… (Instructor 2)

(c) Because of (followed by a noun or the –ing form of a verb) ***can be used in*** the middle of a sentence,… (Instructor 4)

On the other hand, *ability and possibility* expressions in the textbooks bundles list include very few bundle types, accounting for (3%) (Figure 9.3) compared to those found in the instructors' materials bundles list and the written AFL sub-list. For example, the bundle *can you think of* suggests a possible action for learners to perform (see example).

(a) ***Can you think of*** better synonyms of some words? (AEPC Extended Writing Booklet)

The *ability and possibility* bundles such as *can be used to* and *it is possible to* have been identified in the EAP materials (see Table 9.2). It needs to be acknowledged, however, that there is a limited number of bundles with this function. This is an indication that EAP materials seem to draw on few phrases with the *ability and possibility* function, and when this function is used, the bundles appear predominantly in classroom procedural language. This means that *ability and possibility* bundles are typically not found in reading passages.

Evaluation bundles, Epistemic bundles and Hedges bundles

Evaluation bundles: The next sub-category function to be examined is the sub-category of *evaluation expressions*. Although the instructors' materials bundles list and the written AFL sub-list contain similar proportions of bundle types, as shown in Figure 9.3, there are clear differences in the types of bundles they each contain. For example, the written AFL sub-list includes bundles such as *it is necessary to*, *it is clear that*, *it is impossible to*, and *it is obvious that* (see Table 9.2 and Appendix D for a full list of functions). Furthermore, the instructors' materials bundles list, for example, contains bundles such as *the best way to* and *it is interesting that*, expressing directive and impersonal obligation and maintaining *evaluative* functions, as mentioned in Chapter 6. In terms of in-context considerations, the bundle *the best way to* is located in textual parts and instructional parts. The evaluative expression *it is interesting that* is only found in the textual parts (see examples).

(a) Recycling is not ***the best way to*** solve environmental problems because it uses so much energy itself. (Instructor 2)

(b) In your opinion, what is ***the best way to*** find out what type and standard of writing will be expected in your school or department? (Instructor 1)

(c) In this respect ***it is interesting that*** the subjects seem to find it easier to revise (mean 2.94) than proof-read (mean 2.78) their written assignments. (Instructor 3)

The textbook list shows limited use of *evaluative* expressions (5%) (see Figure 9.3) compared to the *evaluative* bundles found in the other two lists. Besides the *evaluative* bundle *it is important to*, identified within the three lists, other distinctive *evaluation* bundles, such as *common in academic writing* and *is the most important*, are used by the textbook authors. The following are some patterns of *evaluation bundles* found in the textbooks bundles list:

(a) After a complex subject (***common in academic writing***) the verb must agree with the main noun in the subject. (CAE-Advanced)

(b) Pre-treatment ***is the most important*** stage in the case of RO plants, as untreated water would clog the membranes and affect their performance. (AEPC Extended Writing Booklet)

(c) 4.2 For each element in 4.1, decide which part of speech ***is the most important***. (CAE-Intermediate)

Epistemic bundles: In the instructors' materials bundles list, *epistemic expressions* bundles account for a similar proportion of bundles to that found in the textbooks bundles list (8%), as shown in Figure 9.3. *Epistemic bundles* such as *is not considered a* and *have shown that the* express thoughts or demonstrate a degree of certainty, concerning knowledge claims. The concordance line results also show that all *epistemic bundles* in the instructors' materials bundles list are found in textual parts. In contrast, most of the *epistemic bundles* in the textbooks bundles list (e.g. *what do you think* and *why do you think*) are located in the instructional parts and are used to retrieve knowledge claims, beliefs, and thoughts, or to detail the claims of others (see examples).

(a) More importantly, if plagiarism ***is not considered a*** major issue, as Dryden asserts, why would they believe they were expected to condemn the practice? (Instructor 3)

(b) Other studies ***have shown that the*** color green is calming. (Instructor 4)

(c) Work in pairs. ***What do you think*** are the key terms in the essay title? (CAE-Advanced)

(d) ***Why do you think*** the passive voice was chosen in the text? (CAE-Upper Intermediate)

However, Figure 9.3 shows that *epistemic expressions* do not occur in the written AFL sub-list. This is due to the nature of the present study, which only focuses on four-word bundles from the written AFL top 200. As mentioned previously (this issue is addressed in the limitations section), the original AFL included 200 bundles, ranging between three-, four-, and five-word bundles; *epistemic expressions* can be found among those bundles not included in the sub-AFL.

Hedges bundles: The final sub-category in the *stance expressions* category is *hedges*. According to Figure 9.3, the three lists have low percentages of *hedge expressions*, being the lowest in the textbooks bundles list (1%). The *hedge expressions* in the written AFL sub-list are different to those identified in the textbooks bundles list and instructors' materials bundles lists. The written AFL sub-list consists of bundles such as *is likely to be* (see Table 9.3). The *hedge* bundle *are more likely to* is found in both the instructors' materials bundles list and the textbooks bundles list, used to indicate some degree of qualification or signalling the writer's claim with a degree of confidence in a connected assertion. In the textbooks bundles list, the *hedge* bundle *are more likely* is the only expression of this type used by textbook authors, and is located in textual and instructional parts. In contrast, in the instructors' materials bundles list there are other *hedge* devices (e.g. *may or may not*) besides the *hedge* device *are more likely* to, which is found in the textual parts of texts (see examples).

(a) The tutors ***are more likely*** to be highly qualified and up-to-date with advanced research. (AEPC Extended Writing Booklet)

(b) New or less common compounds ***are more likely*** to use a hyphen,… (CAE-Intermediate)

(c) The utility of weak ties in one's network revolves around the access to novel information and resources, as weak ties ***are more likely to*** serve as bridges between different network clusters. (Instructor 3)

A conclusion that can be drawn from the data presented is that EAP learners encounter three other sub-categories in the *stance* function (*evaluation*, *epistemic*, and *hedge*), mostly from the instructors' bundles materials but also at a low level in textbooks. Moreover, it appears that EAP materials display different types of bundles under each of these sub-functions than those found in the written sub-AFL. In the EAP materials, the sub-functions include *evaluative expressions* (e.g. *it is interesting that*) and *hedge expressions* (e.g. *are more likely to*), which are presented in academic prose (see Biber *et al*., 1999, 2004). According to Hyland (2010), these functions are important to help refer to or build a relationship with the reader and take responsibility for the proposition.

It can be seen (see Chapter 5 and 6), however, that EAP materials display other (*evaluation*, *epistemic*, and *hedge*) bundles not regularly found in academic writing (e.g. *the best way to*, *common to academic writing*, and *what do you think*). There are two possible interpretations for this noticeable use of *evaluation*, *epistemic*, and *hedge* bundles by EAP authors.

1. As previously mentioned, Wood (2010) reported that there are bundles with functions that are consistent with classroom language such as questions, demands, directives, and so on. For example, the bundle *use a dictionary to* is employed to function as *obligation and directives* in EAP materials. With the same concept in mind, it appears that EAP authors use sub-functions not only to convey instructional language through *demands and directives* bundles, but also to convey personal opinions through a set of bundles related to language and mostly linked to EAP language. For example, the *evaluative* bundles *the best way to* and *common to academic writing* are employed by EAP authors, guiding readers to important facts or arguments.

2. EAP materials include *epistemic* sub-function bundles found in the spoken register (e.g. *what do you think*; see Biber *et al.*, 2004) to signal their intentions. This is a recurring issue in EAP materials, as argued above. EAP materials include bundles in the five sub-category functions of *stance expressions*. However, the overall conclusion is that *stance expressions* seem to be misrepresented to EAP learners. This is because EAP students are exposed to limited types of useful bundles in the *stance expressions* category. In Chen's (2010) research, it was found that the ESP textbooks

> corpus had only focused on one sub-category of *stance expressions*: *impersonal epistemic stance bundles*. She argued that ESP texts did not provide their learners with the possibility to explore the interpersonal relationship between readers and writers conveyed by *stance expressions* found in the electrical engineering introductory textbooks corpus. She continued her argument by stating that students may not know how to use *stance bundles* since they have not been given the opportunity to observe their use.

In addition, some bundles in the instructors' materials bundles list were found to function as *evaluation expressions* (e.g. *it is important to*, *the best way to*, *it is interesting that*). In the literature, the bundle *it is important to* is one of the most typical bundles that is frequently found across studies (such as Biber *et al.*, 1999; Cortes, 2004; Hyland, 2008b, Simpson-Vlach and Ellis, 2010). The bundle *it is interesting to* was included in Biber *et al.*'s (1999) research, typically being found in academic prose, while in the instructors' materials bundles list, the bundle *it is interesting that* was used. However, in the case of the bundle *the best way to*, this sequence is not found in the three previously reported studies. Nonetheless, in the present study, this bundle is found in both reading passages, as well as in instructional language used for providing rules or information to learners. This leads to the conclusion that although the instructors' materials seem to rely on sequences functioning as *stance expressions*, they incorporate several different types of bundles to those found in academic prose.

9.1.3 A comparison of *discourse organizing expressions* in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

Figure 9.4 provides a clear representation of the sub-functions of *discourse organizing expressions* identified in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list, presenting the percentage distribution of the overall functions. Insightful results were obtained from the comparison of the sub-functions in the *discourse organizing expressions* category. In addition, Table 9.4 provides some examples of bundles found in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list (see Chapter 5, Table 5.1; Chapter 6, Table 6.1; and Appendix C for the full lists).

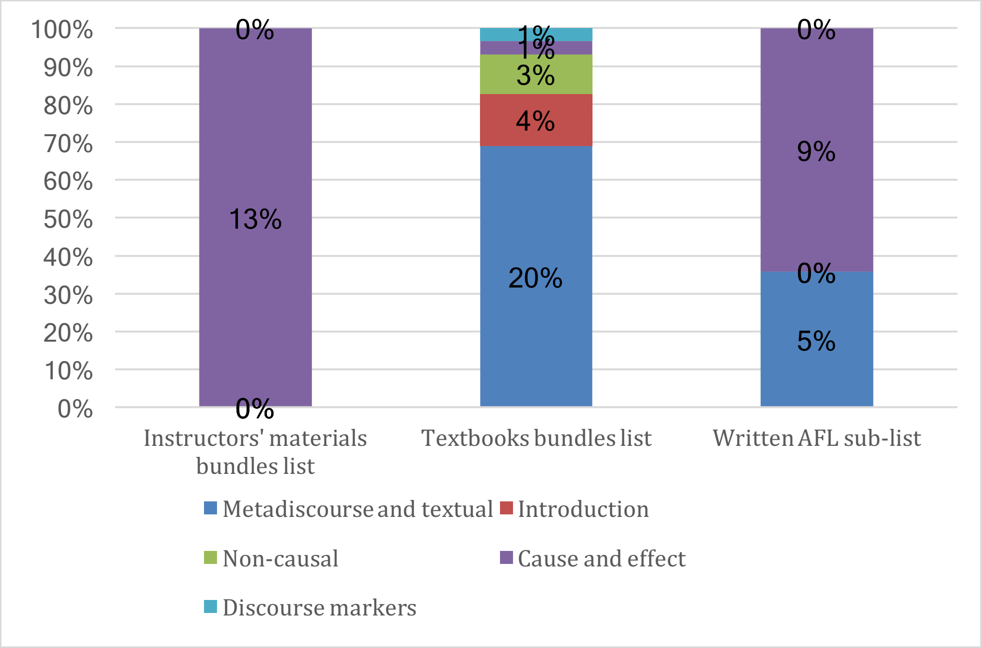

Figure 9.4. Percentage distribution of lexical bundles types in the *discourse organizing expressions* sub-category across the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

Table 9.4. Some examples of bundles in the *discourse organizing expressions* sub-category in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

Number	Instructors' materials bundles list	Textbooks bundles list	Written AFL sub-list
1	Metadiscourse/Textual reference		
		the words in the *from the text in the* *in the text in* *the information in the*	*in the next section* *in this paper* *in the present study*
2	Topic introduction and focus		
		look at the following *look again at the* *you have been given*	
3	Topic elaboration Non-causal		
		research shows that	

		in *research shows that the*	
4	Topic elaboration: cause and effect		
	as a result of *the results of the* *to know each other* *the purpose of this*	*as a result of*	*for the proposes of* *due to the fact* *whether or not the*
5	Discourse markers		
		at the same time	

Metadiscourse and textual reference bundles

Metadiscourse and textual reference bundles: The *discourse organizing bundles* in the textbooks bundles list mostly function as *metadiscourse and textual reference*, and account for four times the proportion of bundles (20%) found in the written AFL sub-list (5%). The instructors' materials bundles list lacks this sub-functional category, as shown in Figure 9.4, which may be due to the issue of the multifunctionality of lexical bundles (see Chapter 2, section 2.2.3.2 and Chapter 11). Bundles in this sub-function are used by authors to either signal or refer to prior or upcoming discourse. As shown in Table 9.4, the written AFL sub-list includes bundles such as *in the next section* and *in this paper we* (see Appendix D for the full functional classification). The textbooks bundles list includes bundles such as *the words in the* and *from the text in*. Examining the mentioned bundles in context, it can be seen that they all appear in the instructional parts (see examples):

(a) First look at the list of vocabulary and locate the ***words in the text***. (AEPC Extended Writing Booklet)

(b) Read the following extracts ***from the text in*** 8.2 and underline the key verbs that describe the trend in youth unemployment. (CAE-Intermediate)

Not only do the proportions of *metadiscourse and textual reference* bundles (e.g. *the words in the* and *from the text in)* in the *discourse organizing functions* have the same proportions as *obligation and directives* in the *stance expressions*, they also both incorporate the largest number of bundles across the textbooks corpus. Such bundles are used in many cases to either signal or refer to prior or upcoming discourse. A similar overall finding was reported in Biber (2006) and Wood (2010: 99), who noted that "in some respect, these types of bundles are similar to those used in classroom talk, in that they tend to be more *stance-oriented* and *discourse focused* than those in written academic text".

It should be noted that EAP learners are exposed to the sub-function of *metadiscourse and textual reference* mostly from the textbooks corpus since this sub-function is missing from the instructors' materials bundles list. The *metadiscourse and textual reference* bundles in the textbooks corpus are clearly different to those bundles found in the written AFL sub-list (see Chapter 6, section 6.3.3, and Appendix D). In addition, the *metadiscourse and textual reference* bundles found in textbooks are different from those in Biber *et al.* (2004), where one bundle (*in this chapter we*) was identified. It needs to be acknowledged, however, that Biber *et al.*'s (2004) taxonomy did not include this sub-functional category. Unlike Simpson-Vlach and Ellis (2010) and the present study, most of Biber *et al.*'s (2004) bundles were grouped under the *topic introduction/focus* category. This is another example of the multifunctional aspect of bundles. EAP materials (mostly textbooks) seem to rely on *metadiscourse and textual reference* bundles to link prior and coming discourse in text (e.g. *the words in the* and *from the text in*). Two possible interpretations, previously mentioned (see section 9.1 of this chapter), seem to be relevant here and are consistent with the textbook findings. For example, the bundles *the words in the* and *from the text in* appear to convey language commonly used in EAP textbooks. It appears that textbook authors use the bundles *the words in the* and *from the text in* to direct readers around the text to specify certain elements, while the bundle *what you have read* appears to convey language mostly found in spoken registers (Biber *et al.*, 2004, 2009).

Topic introduction and focus bundles, Topic elaboration*: *non-causal bundles*, and *Discourse markers bundles

Topic introduction and focus bundles: From the stacked bar chart in Figure 9.4, it is evident that the bundles in the sub-categories *introduction*, *non-causal*, and *discourse markers* are only present in the textbooks bundles list and are not found in either the instructors' materials bundles list or the written AFL sub-list. In addition, Figure 9.4 shows the proportions of bundles in the sub-function *topic* *introduction*, which is only found in the textbooks bundles list with a *discourse organizing function* (4%). As can be seen in Table 9.4, bundles such as *look at the following* (see examples) are found within this sub-function, signaling or directing to prior or upcoming discourse (see Chapter 6, section 6.3 for the full list). This bundle is found in the instructional parts of the textbook texts.

(a) ***Look at the following*** steps which describe the process of reading an essay title. (CAE-Intermediate)

(b) After a seminar, ***you have been given*** an essay with the title "Discuss the influence of advertising on consumer behaviour." (CAE-Advanced)

Similar bundles were found in Biber *et al.*'s (2004) study (e.g. *if we look at*, *take a look at*, and *to look at the*) in classroom teaching but not in the textbooks they examined. The *topic introduction and focus* sub-function overlaps with the *obligation and directives* sub-function in *stance expressions* but to a certain degree. The main difference between them is that *topic introduction and focus bundles* are used for a special function – *introducing a new topic*. This sub-category, in Simpson-Vlach and Ellis (2010), was also found to overlap with *identification and functions* in *referential expressions*, adding to the multifunctional complexity of lexical bundles (see Chapters 2 and 11).

Topic elaboration*: *non-causal bundles: In *topic elaboration: non-causal expressions*, the textbooks bundles list includes bundles such as *research shows that in*, where this phrase is used to elaborate and summarize, without explicitly implying a causative relationship. This bundle is mostly found in the instructional parts in textbooks (see example).

(a) ***Research shows that in*** the written academic corpus, the most frequent adverbs that come before less/more common are much and far. (CAE-Advanced)

Discourse markers bundles: Within the sub-function *discourse markers*, the bundle *at the same time* is used to connect and signal transitions between clauses or elements in the textbooks corpus. The contextual information shows that this bundle is found in textual parts (see examples).

(a) Unfortunately, the drop in the price of butter and the increase of milk production happened ***at the same time***. (CAE-Intermediate)

(b) We cannot deal with all global problems ***at the same time*** so we have to find ways of deciding the order in which they are dealt with. (CAE-Upper Intermediate)

As stated above, the three mentioned sub-functions: *topic introduction and focus*, *topic elaboration*: *non-causal*, and *discourse markers* are largely found in the textbooks bundles list, indicating that EAP learners are exposed to these three sub-functions via textbooks. It is clearly evident that *topic introduction and focus* functions include particular bundles (e.g. *look at the following*) which are similar to those found in spoken registers (Biber *et al*., 2004, 2009). The function *topic elaboration*: *non-causal* is found in limited and unique types of bundles (e.g. *research shows that in*). This bundle seems to be commonly related to EAP language. In terms of the *discourse markers* function, as can be seen, only one bundle is found: *at the same time.* In fact, this *discourse marker* bundle seems to be one of the useful bundle, considering it is found in academic prose (Biber *et al*., 2004) as well as in biology (Cortes, 2004), electrical engineering, applied linguistics, and business (Hyland, 2008b) texts.

Learners' corpus studies such as Chen and Baker (2016) reported that as students progress and advance to higher levels, the transition in their writing from an informal manner to a more academic style of writing can be detected in the use of *at the same time* instead of "a lot of". It is important, however, to report that in regard to the bundle *at the same time*, there is a categorization difference between Chen and Baker's (2010) and Biber *et al.*'s (2004) studies and the present study. Both of these studies classified the bundle *at the same time* as a *deictic* bundle. However, in the present study this bundle is classified, using Simpson-Vlach and Ellis's (2010) framework, as a discourse marker. As explained in Chapters 2 and 11, this is due to the multifunctional aspect of bundles. To conclude, it should be clear from this data that the sub-functions: *topic introduction and focus*, *topic elaboration*: *non-causal*, and *discourse markers* are not richly represented with academic bundles for EAP learners to encounter and learn, offering very limited instances of explicit instruction.

Topic elaboration: cause and effect: Apart from the previously analysed sub-functions of the *discourse organizing function*, the stacked bar chart in Figure 9.4

shows the proportions of the final sub-category: *cause and effect*. Across the three lists, *cause and effect* expressions have noticeable different proportions. In the instructors' materials bundles list, relative to the *discourse organizing function*, most bundle types have the sub-function *cause and effect* expressions, accounting for 13% of the total. In the written AFL sub-list, the percentage of bundles in this sub-function (9%), is also greater than the other sub-functions in this category. In the textbooks list, unlike the *metadiscourse and textual reference* category, a very restricted number of bundles (one bundle) comes from the *cause and effect* expressions, as shown in Figure 9.4 and Table 9.4. The written AFL sub-list includes bundles such as *for the purposes of* and *due to the fact*. The *cause and effect* expressions signal a cause and effect relationship, as in the bundle *as a result of*, which is identified in both the instructors' materials bundles list and the textbooks bundles list. This pattern is usually located in the textual parts (see examples).

(a) Nobody knows what's likely to happen in the future ***as a result of*** what's going on in the Middle East at the moment. (AEPC Extended Writing Booklet)

(b) People migrated to towns ***as a result of*** industry. (CAE-Advanced)

With regard to the instructors' materials bundles list, *as a result of*, *as a consequence of*, *in order to learn*, and *to know each other* are *cause and effect* expressions, signaling a reason or an effect relationship. These three bundles were checked through examining concordance lines, according to their location in context. It was found that the bundles *as a result of* and *as a consequence of* are located in textual parts while the bundle *in order to learn* is found in instructional parts (see examples).

(a) The future is unpredictable ***as a result of*** current development in the Middle East. (Instructor 1)

(b) …seem to have their every action 'interpreted with stunning regularity ***as a consequence of*** their 'Asianness,' their …' (Instructor 3)

(c) Using social media ***in order to learn*** English. (Instructor 1)

(d) …in many cases linguistic knowledge inhibits individual from getting ***to know each other*** (Instructor 3)

(e) ***the purpose of this*** study is to explore attitudes students at… (Instructor 2)

EAP learners seem to encounter *topic elaboration: cause and effect* bundles mostly from instructors' materials, and these are different from those found in the written AFL sub-list (see Chapters 5, 6 and Appendix D). For example, *as a result of* is the only bundle found in the textbooks bundles list. However, the instructors' materials seem to provide different bundles in this sub-function for EAP learners to encounter, including *the purpose of this* and *to know each other* (see Table 9.4). The bundle *as a result of* is found in academic writing (Biber *et al.*, 2004, 2009) and in biology (Cortes, 2004), electrical engineering, applied linguistics, and business (Hyland, 2008b) texts. The bundle *as a consequence of* was found in academic writing (Biber *et al.*, 1999). To conclude, although the *topic elaboration: cause and effect* sub-function includes useful bundles, it seems to be misrepresented and underrepresented (Chen, 2010) to EAP learners.

In terms of the instructors' materials list, most bundles in this sub-category have the sub-function *cause and effect* (e.g. *as a result of*, *to know each other*, *the purpose of this*). These three bundles are found in the reading articles in the instructors' materials corpus. Again, the bundles *as a result of* and *the purpose of this* have been established as important bundles in academic prose (Biber *et al.*, 2004; Simpson-Vlach and Ellis, 2010). However, the bundle *to know each other* is not found in previous studies, so it is not recognized in the literature as a typical academic bundle. It seems that the content of materials provided by the teachers has influenced the findings associated with this bundle because the bundle *to know each other* is found in reading passages, discussing either friendship or student life. It is important to see the whole picture that emerges from the use of these types of bundles. *Metadiscourse and textual reference* bundles are useful and effective for learners to encounter while reading, and authors need them to construct and guide learners to complete tasks. However, the real question that needs to be explored is whether students benefit from the treatment of such bundles in EAP materials. This brings us to the overall discussion of the treatment of the pragmatic functions of lexical bundles in EAP materials, which will be addressed after briefly discussing the functions of shared bundles in the three lists.

9.2 Shared functions in the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list

In the instructors' materials bundles list, textbooks bundles list and the written AFL sub-list, four bundles are shared (see Chapter 7), revealing two functional classifications. The bundles *it is important to*, *it is possible to*, and *can be used to* serve as *stance expression* bundles while the bundle *on the other hand* functions as a *referential* bundle. These results indicate that EAP authors and instructors do not pay attention to essential academic bundles or their functions that are found in the written AFL sub-list when constructing their materials. This leads to the issue whereby EAP learners are likely to encounter other varied useful bundles found in academic prose, because, as revealed in this section, instructors' materials make use of other academic bundles found in other studies, such as Biber *et al.* (1999) and Biber *et al.* (2004).

In the literature on lexical bundles, many studies stress the value of these *stance* lexical bundles (e.g. *it is important to*, *it is possible to*, and *can be used to)*, and highlight them as the most common and useful academic bundles for learners. For example, Hyland (2008b) found the *stance* bundle *it is important to* in applied linguistics and business texts, while Cortes (2004) asserted that biology professional writers make use of this bundle. Regarding the *stance* bundle *it is possible to*, Hyland (2008b) found this bundle to be used by electrical engineering authors. According to Hyland (2008b), the *stance* bundle *can be used to* is used across a wide range of disciplines (e.g. biology, electrical engineering, and business). In addition, Biber *et al.* (1999, 2004) confirmed that these three *stance* bundles are common in academic discourse and academic prose.

9.3 Treatment of overall functions of lexical bundles in EAP materials

The comparison of the pragmatic functions revealed in this chapter relating to EAP materials has indicated two important existing gaps:

(1) in the representations of functions in EAP materials
(2) in the pedagogical treatment of the functions in EAP materials

Representations of functions in EAP materials: The functional comparison of four-word lexical bundles indicates a language gap in EAP materials in terms of *referential*, *stance*, and *discourse organizing bundles*. In terms of *referential* bundles, on one level, instructors' materials seem to be representative of the sub-functions (e.g. *intangible framing*, *quantity*, *identification*, *contrast and comparison*, and *deictics and locatives*), including different numbers of bundles. However, *tangible bundles* in the *referential* function are underrepresented. In terms of textbooks, most of the sub-functions in *referential bundles* are misrepresented, while the *tangible* and *identification and focus* sub-functions are underrepresented. Overall, it seems that *tangible framing bundles* are underrepresented in EAP materials, resulting in EAP learners having a limited and distinctive exposure to these forms. This result supports the findings of Chen (2010), who found that three sub-categories of lexical bundles were underrepresented in the ESP textbooks that the study examined.

With regard to the sub-function *stance*, the instructors' materials are representative of *epistemic*, *expressions of ability and possibility*, and *evaluation* bundles with varied numbers of bundles, while *hedges* and *obligation* and directives are misrepresented. On the contrary, the *obligation* and *directives*, and *epistemic* sub-functions are largely misrepresented in textbooks while the sub-functions *hedges*, *ability and possibility* are underrepresented. It needs to be acknowledged, however, that *hedges* are largely misrepresented to learners in the EAP materials with limited exposure to bundles such as *are more likely*. Again, similar to Wood (2010), this means that EAP learners may not be benefitting from this exposure. Overall, EAP materials misrepresent and underrepresent *discourse organizing functions*. Even with these findings, it seems that instructors' materials are less effective in handling lexical bundles with their functions than textbooks and use bundles differently. According to Hyland (1999), textbook authors in his study employed metadiscourse language in a different way to research article writers. This leads to the conclusion that the misrepresentation and underrepresentation of these functions by textbook authors, as previously stated in Chapter 8, is simply due to the textbook authors writing for different genres, including different types of bundles with different functions. This contrasts with the bundles used in research article writing found in the written AFL sub-corpus.

Pedagogical treatment of main functions in EAP materials: At the level of the organizational and structural purposes of EAP textbooks, the authors' employment of bundles in general may seem to be efficiently and logically managed. However, at the level of pedagogy, the treatment of bundles needs to be readdressed (Wood, 2010). Thus, the limited focus on the functional classification of the most frequent four-word bundles found in textbooks gives a clear reflection of the reality of functions found in EAP textbooks in the examined pre-sessional course. Coxhead and Nation (2001) highly recommended the idea of teaching specialized academic vocabulary to learners, including the teaching of lexical bundles (Coxhead and Byrd, 2007). In addition, Zimmerman and Schmitt (2005) and Cortes (2006) argued that it is recommended for instructors to create opportunities for their learners to encounter vocabulary in a variety of natural contexts, which will contribute to gradual lexical learning.

Explicit instructions by EAP authors on the teaching of lexical bundles and their functions should be dealt with pedagogically, and through corpus-based materials (Burton, 2012). It may be that EAP teachers/authors are less familiar with lexical bundle research and are unsure of how to address these bundles. This means that some EAP instructors and authors do not pay attention to such essential academic bundles when constructing their materials, leading to EAP learners not encountering these useful bundles. This could be a drawback in lexical bundle treatment. Also, the academic bundles *it is important to*, *it is possible to*, and *can be used to* seem not to be subject to any pedagogical treatment in both the textbooks and instructors' materials corpora, suggesting another major weakness in EAP materials. These drawbacks were similarly detected in Wood (2010) and Wood and Appel's (2014) work. For example, Wood and Appel (2014) stated that the formulaic sequences found in EAP textbooks appeared mostly in reading parts or passages, but were not covered in any pedagogical or educational form to familiarize EAP learners with these expressions. As repeatedly stated, this conclusion is based only on the material provided, collected and analysed in the present study. It is important not to make general conclusions on the teaching of lexical bundles in EAP textbooks and instructors' materials because online supplementary materials, word/phrase lists, CD-ROMS, websites, extended activities etc. on lexical bundle handling could have been available to EAP learners to support their writing needs, as stated in the introduction of the Cambridge Academic English textbooks and the course syllabus.

It is important to highlight these sub-functional categories to learners. Many scholars and researchers have concluded that learners are encouraged to *notice* lexical bundles and learn how to use them properly (Nattinger and DeCarrico, 1992; Lewis 2000; Hyland 2008b), including *deictics and locatives* (Chen, 2010). This is because, as stated in Chapter 2, section 2.1.2, the entrenchment process suggests that language users' (learners') linguistic knowledge is continuously renewed and reorganized, and is encouraged by social and repeated interactions. In addition, Chen (2010) indicates a gap between the linguistic items found in her electrical engineering introductory textbooks and an ESP textbook. I would suggest that the fluctuation in bundle functions reinforces the previously mentioned argument that the lexical bundle choices of authors and in instructors' materials are based on instructional materials, variation and stylistic preferences (Hyland, 2008a), reflecting the influence of other corpora-based materials rather than the corpus behind the AFL that was used in this study.

9.4 Concluding remarks

This chapter has profiled the functional categorizations of the most frequent bundles in the instructors' materials bundles list and textbooks bundles list compared to the written AFL sub-list, adding to the research on EAP materials and materials development. While the instructors' materials corpus includes bundles with *referential expressions* and the textbooks corpus has bundles with *stance expressions*, they both include bundles with distinctive characteristics from those explored in the written AFL sub-list and academic prose.

At first glance, these bundles might be taken as appropriate for learners to encounter as examples of language-in-use, marking a good starting point for examining bundles in academic writing classes. However, the above discussion has shown the reality and the nature of these bundles, which may not benefit learners, as strongly claimed by Wood (2010) and Wood and Appel (2014). There are marked differences in the *discourse organizing bundles* among the three corpora. It can be concluded that the *referential*, *stance*, and *discourse organizing* functions seem to be either underrepresented, misrepresented or ignored in the EAP materials examined.

The following chapter will provide general conclusions, implications, and recommendations for future research.

CHAPTER X

GENERAL DISCUSSION AND SUMMARY OF FINDINGS

In the field of academic discourse, relatively few studies have examined lexical bundles, a significant kind of formulaic sequence, in EAP textbooks (see Jones and Haywood, 2004; Wood, 2010; Wood and Appel, 2014). The present research is a corpus-based investigation into the content of EAP materials, focusing on examining lexical bundles in the teaching of academic writing in textbooks and instructors' materials. I was interested in exploring a new area of research interest regarding formulaic sequences, by examining unexplored materials (e.g. handouts, worksheets, research articles and quizzes) selected by the instructors of an EAP programme. By doing so, I was able to enhance the knowledge and contribute to the understanding of the preferred uses of lexical bundles in instructors' materials.

The study set out to identify the most frequent four-word bundles found in EAP materials and compare the bundles to those in the ready-made Academic Formulas List (AFL) constructed by Simpson-Vlach and Ellis (2010). The decision to use the ready-made AFL is an innovative approach taken by this study. The decision was taken because the sequences offered in the AFL are considered to be representative of lexis in the academic register, and as such the AFL is a valuable resource. The application of the MI statistical and psychological measure of utility called 'formula teaching worth' (FTW) to complement frequency measures in the identification of lexical bundles was also a key motivating factor for selecting the AFL, enabling a reliable data evidence base for use in the present study (see Chapter 4). Additionally, for comparison purposes, from the original AFL (Simpson-Vlach and Ellis, 2010), four-word bundles were selected and listed, and structural and functional classifications were conducted, creating the written AFL sub-list (see Chapter 4 and Appendix C).

In the literature review chapter, I sought to draw a clear picture of the extent of the literature on formulaic sequences, particularly in relation to providing a full account of lexical bundles in terms of their identification criteria and their structural and functional classification, and regarding lexical bundles and metadiscourse. I also sought to show the role of corpora in analysing sequences and bundles by discussing key issues such as the issue of the size and representativeness of corpora, the influence of corpus studies on the intersection between lexis and grammar, contributions of corpus studies to the analysis of sequences and bundles, methodological approaches to investigating sequences and bundles, and corpus-based research on sequence and bundles. In addition, I discussed studies that examined sequences and bundles in academic discourse and EAP, focusing on the treatment of sequences and bundles in EAP textbooks. In doing this, I set out to discover how previous studies investigated lexical bundles in textbooks when researching how material designers and writers present sequences and bundles in their work. At the end of the literature review chapter, five research questions were established to achieve the goal of the investigation (see Chapter 2, section 2.4.6.3 and Appendix A).

10.1 Summary of findings

The following section presents an overall summary of the findings together with overall conclusions to the five research questions (see Appendix A):

10.1.1 Research question one and three: frequency and type of bundles

Answering the first research question entailed the identification of the most frequently occurring lexical bundles in EAP materials in an EAP pre-sessional programme in the UK particularly aimed at teaching academic writing. This was achieved by creating two EAP corpora (textbooks and instructors' materials). Two EAP lists were formed (see Chapters 5, 6, and 7). This was conducted via a computer software program using bundle identification criteria, as the first step in the analysis process (see Chapter 4). Answering the third research question entailed ascertaining whether the lexical bundles in the EAP materials were selected on the basis of the written AFL sub-list. This was achieved by comparing the most frequent bundles found in the EAP materials to the written AFL sub-list.

In total, 79 different four-word bundles were generated from the frequency analysis in the instructors' materials corpus. The most frequent bundles found at the top of the instructors' materials bundles list (Chapter 5) were *in the United States*, *at the end of*, *the end of the*, *it is important to*, and *the extent to which*. The textbooks corpus yielded 102 four-word lexical bundles that comprised the textbooks bundles list (Chapter 6). At the top of the textbooks bundles list, the most frequent four-word lexical bundles were *focus on your subject*, *as a result of*, *it is important to*, *look at the following*, *and answer the questions*, and *the words in the*. The comparison revealed that only four four-word bundles were shared between the instructors' materials bundles list, textbooks bundles list, and the written AFL sub-list, including *it is important to*, *it is possible to*, *on the other hand*, and *can be used to*. To conclude this section, although there are four bundles that are shared with the written AFL sub-list, EAP materials include varied types with different bundle frequencies compared to the types found in the written AFL sub-list. The differences in bundle types are related to representational and organizational features, the use of instructional language to guide learners, and genre differences (between textbooks, instructors' materials and the written AFL sub-list, which is taken from research articles). This means that instructors and textbook authors are required to use certain linguistic bundles more frequently than others to organize and facilitate tasks for learners. At the same time, textbook authors and instructors may offer other means of teaching opportunities on lexical bundles to their EAP learners, which were not directly related to the aim of this study.

10.1.2 Research question two and three: structures of bundles

The second research question involved surveying the structural features of the four-word lexical bundles from EAP textbooks and instructors' materials through concordance analysis, organizing them using Biber *et al.*'s (1999) structural framework (Chapters 5, 6, and 8). The third research question sought to uncover whether EAP materials are using the same bundle structures found in ready-made academic lists such as the AFL, drawing on the similarities and differences between the EAP list and the written AFL list. The structural analysis showed how lexical bundles are used by EAP authors, contributing to the academic nature of the content of EAP materials.

In the following, first, I will list the overall main structures of the identified bundles in the EAP materials compared to the written AFL sub-list:

1. VP-based structures are the most frequently occurring across the three lists, accounting for the highest proportion of bundles in the textbooks and in the written AFL sub-list.

2. PP-based structures are the second most frequently occurring with very close percentages of bundles in the instructors' materials and the written AFL sub-list compared to a low occurrence in the textbooks.
3. There are varied proportions of NP-based structures; most bundles of this type are found in the textbooks, with slightly lower proportions in the instructors' materials and the written AFL sub-list.
4. Other expressions structures are the least frequently occurring structure across the three lists; the highest proportion of bundles of this type are found in the instructors' materials with lower proportions in the textbooks and written AFL sub-list.

Second, I will list the overall patterns of VP-based structures in the EAP materials compared to the written AFL list:

1. The *anticipatory it + phrase* pattern accounts for the largest proportion of bundles in the written AFL sub-list compared to the instructors' materials, with an even lower proportion in the textbooks.
2. Four VP-based sub-structures are found across the instructors' materials bundles list and the written AFL sub-list, including *passive verb + prepositional phrase*, (*verb phrase* +) *that-clause*, (*verb/adjective* +) *to-clause*, and *pronoun/noun phrase + be* (+*…*).
3. The instructors' materials include the categories *pronoun phrase + verb* and *adverbial clause fragment by the subordinator phrase (in order to*), which are not included in the written AFL sub-list.
4. 11 VP-based sub-structures are found in the textbooks but not in the written AFL sub-list, including *copula be + noun or adjective*, *noun + verb phrase + that-clause*, *(passive) verb phrase + to-clause*, *to-clause fragment*, *verb phrase + determiner phrase*, *verb phrase + prepositional phrase*, *verb phrase + noun phrase*, *wh-questions*, *Yes-no questions*, *2nd person pronoun you +VP fragment*, and *modal verb question*.

Third, I will list the overall patterns of PP-based structures in the EAP materials compared to the written AFL list:

1. The *prepositional phrase with embedded of-phrase* structure is the most frequent structure in the instructors' materials and the written AFL sub-list but occurs the least frequently in the textbooks.

2. The *other prepositional phrase* structure accounts for a similar percentage of bundles across the three lists.

Fourth, the following list presents the overall patterns of NP-based structures in the EAP materials compared to the written AFL sub-list:

1. The *noun phrase with of-phrase* structure accounts for the highest percentage of bundles across the three lists, with the highest percentages in the instructors' materials and textbooks compared to the written AFL sub-list.
2. The *noun phrase with other post-modifier* structure accounts for the highest proportion of bundles in the textbooks compared to the written AFL sub-list, and occurs even more frequently than in the instructors' materials.

To conclude this section, overall, the structures of the most frequent four-word bundles identified in the EAP materials are considered to be verb-centred echoing ESL and EFL grammar and grammar writing textbooks (Coxhead and Byrd, 2007). Although EAP materials include few VP-based structures that are commonly found in the written AFL sub-list and academic prose, it seems that EAP textbooks use structures that are not common to academic writing such as *adverbial clause fragment by the subordinator phrase (in order to*). Textbooks rely greatly on PP-based structures, including bundles mostly relevant to EAP language such as *from the text in*. However, instructors' materials seem to provide learners with PP-based structures that are usually found in the written AFL sub-list and in academic prose. Furthermore, when textbook authors employ NP-based and PP-based structures, they use different types of bundles, incorporating NP-based and PP-based patterns, to those found in the AFL sub-list and academic prose.

10.1.3 Research question two and three: functions of bundles

The second research question also involves the exploration of the functional features of the four-word lexical bundles found in EAP textbooks and instructors' materials through concordance analysis, and classification using Simpson-Vlach and Ellis's (2010) functional taxonomy (Chapters 5, 6 and 9). The third research question involves the comparison of the functional features of bundles in the instructors' materials bundles list, textbooks bundles list, and written AFL sub-list, to ascertain whether the EAP materials are based on data taken from corpus- driven lists.

Here, first, I will list the overall patterns of the main functions of bundles in the EAP materials compared to the written AFL sub-list:

1. The instructors' materials and the written AFL sub-list have higher percentages of bundles in the *referential expressions* category than in the textbooks.
2. The EAP materials and written AFL sub-list have close percentages of bundles in *stance expressions*.
3. *Discourse organizers function expressions* have higher percentages in textbooks than in the instructors' materials and the written AFL sub-list.

Second, I will list the overall patterns of *referential expressions* in the EAP materials compared to the written AFL list:

1. Instructors' materials and the written AFL sub-list contain similar proportions of *intangible framing bundles* but textbooks contain fewer such bundles.
2. The written AFL sub-list contains a lower proportion of bundles in the sub-function *ability and possibility* expressions than is the case in the instructors' materials and textbooks.
3. Textbooks and the written AFL sub-list contain similar proportions of *tangible framing* bundles but the instructors' materials contain fewer such bundles.
4. Instructors' materials, textbooks, and the written AFL sub-list contain similar proportions of quantity *expressions* bundles.
5. Instructors' materials contain slightly lower proportions of *contrast and comparison* bundles than those found in textbooks and the written AFL sub-list.
6. The written AFL sub-list and instructors' materials contain similar proportions of *identification and focus* bundles to those found in textbooks.

Third, I will list the overall patterns of *stance expressions* in the EAP materials compared to the written AFL sub-list:

1. Textbooks contain a higher proportion of *obligation and directives expressions* than in the instructors' materials and the written AFL sub-list.
2. The written AFL sub-list contains a higher proportion of bundles of *ability and possibility expressions* than the instructors' materials and textbooks.
3. Instructors' materials and the written AFL sub-list contain similar proportions of *evaluation expressions* but these occur less frequently in the textbooks.
4. Instructors' materials contain a slightly higher proportion of *epistemic expressions* than the textbooks, while this sub-function is absent from the written AFL sub-list.
5. Textbooks contain a lower proportion of *hedges expressions* than the written AFL sub-list and instructors' materials.

Fourth, I will present an overall comparison of *discourse organizers* in the EAP materials and the written AFL sub-list. In the following points, I will list the overall patterns of *discourse organizing expessions* in the EAP materials compared to the written AFL list.

1. Textbooks have a higher proportion of *metadiscourse and textual reference expressions* than the written AFL sub-list; the instructors' materials lack this bundle type.
2. Only the textbooks make use of the following sub-functions: *topic introduction and focus*, *topic elaboration*: *non-causal*, and *discourse markers*.
3. The written AFL sub-list and instructors' materials make use of *topic elaboration: cause and effect expressions*; bundles of these type occur more frequently than in the textbooks.

In conclusion, overall, the close percentages of bundles in the instructors' materials and the written AFL sub-list is an indication that instructors' materials make use of similar functions to those found in the written AFL sub-list, in which *referential expressions* is the most common feature. In contrast, textbooks seem to have different percentages of bundles, in terms of functions, compared to the AFL sub-list, with *stance expressions* being the most common in the textbooks. Additionally, the overall comparison with the written AFL sub-list revealed that *referential*, *stance*, and *discourse organizing* functions in the EAP bundles are underrepresented, misrepresented, or ignored. This means that EAP materials make use of the structures found in the written AFL sub-list; however, the types of bundles that EAP learners are exposed to under these sub-functions could be viewed as not very useful in the process of composition writing. At the level of the organizational and structural purposes of EAP textbooks, the authors' employment of bundles in general may seem to be efficiently and logically managed due to genre differences. However, at the level of pedagogy, the treatment of bundles needs to be readdressed (Wood, 2010). Thus, the focus on the structural and functional classification of the most frequent four-word bundles found in textbooks provides a clear reflection of the functions found in the textbooks examined in this study.

10.1.4 Research question four: in-context information

An in-context analysis showed that 69% of the bundles found in the textbooks corpus appear in instructional parts and only 7% are found in textual parts. However, in the instructors' materials corpus, only 6% of bundles are found in instructional parts, while 71% of bundles are found in textual parts. It seems that the majority of bundle types in the instructors' materials are found in reading passages, while in textbooks the majority of bundles (academic and non-academic) come from classroom-based language parts. It appears that most of the bundles found in the instructors' corpus resemble the academic bundles found in academic prose (Biber *et al.*, 1999; Biber *et al.*, 2004; Cortes, 2004; Hyland 2008b). On the contrary, however, the lexical bundles used by the textbook authors comprise academic, non-academic and classroom-based language, which is commonly found in EAP textbooks (e.g. *focus on your subject*) (Wood, 2010).

10.1.5 Research question five: teachable units

It seems that most of the four-word lexical bundles studied in the present research have not been sufficiently treated in a pedagogical manner by EAP instructors and textbook authors, with only infrequent instances of tasks focusing on the teaching of some academic bundles (see Chapter 7, section 7.5). This could mean that lexical bundles are not well treated from a pedagogical perspective in the EAP materials examined due to the nature of the texts or genre differences in the course material and textbook writing in regard to lexical bundles.

10.2 Concluding remarks

This chapter has summarized how this study addressed the five research questions and provided general conclusions. I will now move on to the next and final chapter, which discusses some limitations encountered in this study, addresses implications for pedagogical and materials development, and suggests the opportunities this research offers for further studies.

CHAPTER XI

LIMITATIONS, IMPLICATIONS, AND RECOMMENDATIONS

The present thesis continues the line of research on formulaic sequences, making an original contribution to the study of lexical bundles by focusing on an EAP pre-sessional programme, examining textbooks and the unexplored area of instructors' materials directed at teaching academic writing. In my research, I have not encountered studies exploring instructors' materials for lexical bundles focusing on academic writing. Therefore, it seems that my research is novel and contributes to the field of formulaic sequences, particularly lexical bundles. However, though there are related empirical studies that analyse lexical bundles in EAP textbooks; these studies mostly examine the types and functional aspects of lexical bundles, but rarely examine their structural features. In addition, it seems that *Cambridge Academic English* (advanced) (Hewings and Thaine, 2012) was not analysed by other studies. Based on these two points, it can be noted that my study contributes empirical evidence on this contemporary set of textbooks, including their functional and structural classifications, enabling readers to gain further insights into the field of enquiry on lexical bundles. Furthermore, this study analysed the retrieved four-word bundles by looking at their location in texts in terms of being in reading sections and/or instructional sections and if there were any exercises or tasks that handled lexical bundles. This analysis will aid EAP practitioners in gaining a general understanding and perspective in relation to the lexical bundles used in such EAP materials.

Overall, this study constitutes a valuable and significant contribution for EAP practitioners, pre-sessional programmes, and material developments or designers. In addition, my research may introduce different kinds of data for inclusion in teacher corpora, which will be discussed below. In this chapter, I will start by identifying potential limitations associated with this research and touch upon its implications and present recommendations for future research.

11.1 Limitations of this study

Although for functional purposes, I managed to address the limitations of the present research as they occurred, there were still issues that needed to be revisited or explored. Issues of frequency (Chapter 2), size and representativeness (Chapter 2), and exclusion criteria (Chapter 4) have been fully recognized and considered. These were important elements in building the corpus and so addressing them as early as possible helped to resolve any impending misunderstandings. At the same time, certain limitations need further explanation, including the generalizability of the results and its limitations, the multifunctionality of lexical bundles, and the written AFL sub-list as a comparative tool.

11.1.1 Generalisability of the results and its limitations

A very important matter needs to be clarified; that is, the aim of the present study is not to seek to generalize the findings for application to a wider domain, but rather to provide a rich and contextualized understanding of the types of bundles, frequencies, structures and functions provided to EAP learners through building a specialized corpus (see Chapter 2, section 2.4.2) from materials obtained from a pre-sessional course. However, in an academic environment where improving the lexical repertoire of EAP learners is sought, generalization of the results in relation to knowledge acquired merits careful attention, both quantitatively and qualitatively. The quantitative results presented here profile the nature of bundles and their frequencies and are compared to the AFL, seeking evidence for pedagogical interpretation and demand. For example, the frequency results identify bundles such as *focus on your subject* (in the textbooks corpus) and *in the United States* (in the instructors' materials corpus) compared to the bundle *on the other hand* (in the written AFL sub-corpus), highlighting the kind of lexical bundles that learners encounter.

Moreover, the qualitative results obtained from the samples examined (see Chapters 7, 8 and 9) address these findings by looking into issues such as the nature of readings, worksheets and handouts provided by instructors and the textbooks' presentational and organizational features. By doing so, the study incorporates unique features such as functional and structural categorizations, in-depth analysis of bundles in the materials, and a small questionnaire. These findings might be helpful to explain the relevance of the results in a more general manner. The number of instructors (five) participating and textbooks examined (4) might seem relatively small to make a complete generalisation of the EAP material. However, the data examined seem to be important, sufficient and representative (see the methodology section) when reporting on the practices and lexical bundle treatment of EAP materials on a specific course.

At the same time, as stated in Chapter 2, section 2.4.2, results obtained from small corpora may be limited. For example, in the analysis, it was vital not to draw broad inferences from particular instances. This was especially difficult in the comparison between the EAP materials and the written AFL sub-list, where the modified AFL, in the present study, only examined four-word bundles. This limitation is fully addressed in section 11.1.2. of this chapter. In addition, the issue of examining the textbooks which were produced by the same publisher and written by two authors could hinder the overview of the results. However, according to Dörnyei (2007), purposive sampling is considered a crucial consideration in qualitative research. Therefore, together, instructors' materials and textbooks, in the present research, provided frequency information, structural and functional classifications, and findings that offered insightful considerations into the pre-sessional programme, which in turn could serve the wider EAP community.

11.1.2 Multifunctionality of lexical bundles

A significant part of this research study is devoted to performing functional classifications of the four-word bundles generated from the EAP corpora of instructors' materials, textbooks, and the written AFL sub-list (see Chapters 5, 6 and 7). One critical feature that should be carefully considered when assigning functional categories and sub-categories to lexical bundles is the multifunctional aspect of lexical bundles, as discussed in Chapter 2 (see section 2.2.3.2). The multifunctionality of lexical bundles is a procedural problem noted in previous studies (Ädel and Erman, 2012; Biber *et al.*, 2004: 384; Chen and Baker, 2016; Salazar, 2014; Simpson-Vlach and Ellis, 2010) and classifying lexical bundles in this way could be driven by subjectivity.

Despite the fact that it was straightforward to classify most four-word bundles during the process of functional categorization (see Chapter 4), there was some uncertainty regarding the functional classifications of a few four-word bundles. The first multifunctionality limitation is related to selecting the most appropriate sub-function for a four-word bundle since the bundle could be allocated to two functional classifications. This issue was resolved by conducting concordance line checks to allocate the bundle under "the most common use" (Ädel and Erman, 2012; Biber *et al.*, 2004: 384; Chen and Baker, 2016; Salazar, 2014; Simpson-Vlach and Ellis, 2010).

The second limitation stems from the idea that every single occurrence of a four-word bundle is noted as important, being considered as an "encounter" for a group of EAP learners. Thus, basing my decision only on "the most common use" and neglecting its other functions could be misleading. At the same time, the decision to allocate the bundle under two functional categories could be confusing and complicated. Therefore, to overcome this limitation, I decided to keep the solution of functional classification to "the most common use", and to report on the other functional classification when needed in the analysis, by consulting concordance lines.

11.1.3 The written AFL sub-list as a comparison tool

In research question 3 (see Chapter 2), I was interested in understanding whether an EAP programme (instructors' materials and textbooks) bases its teaching writing materials (lexical bundles) on data retrieved from corpus-driven lists such as the AFL (Simpson-Vlach and Ellis, 2010). The written AFL was used as a practical and comparative tool in the present study (see Chapter 4 for general information, selection criteria and information on modifications to the AFL). Originally, the written AFL corpus was composed of 200 bundles, including three-, four-, and five-word bundles, focusing on bundles that are used in academic writing (Appendix B). However, the present study concentrates on four-word bundles (see Chapter 4, section 4.3.2.1) and used a much smaller corpus with different types of materials (see Appendix K). For comparative purposes, four-word bundles were selected from the AFL. This means that the original list of 200 bundles was significantly reduced, totalling 57 bundle types.

This reduction, however, was inevitable and important, enabling the completion of the comparisons needed for the present study. I tried to maintain a cautious and thorough approach, particularly when eliciting final and overall conclusions, relating to frequency, form, and functional examination of the written AFL sub-list. Offering general conclusions on certain aspects of a large corpus could be a potential limitation; however, focusing on relevant aspects such as four-word bundles may give an in-depth examination, yielding insightful findings that contribute to the whole analysis. In spite of the limitations mentioned above, I have enough confidence to assert that I took the necessary steps to overcome any obstacles to ensure that my research is reliable and trustworthy. In the next section, I will present some of the implications that resulted from this study.

11.2 Implications of this study

The present study has shown that the lexical bundles in the instructors' materials corpus not only were drawn from parts related to reading passages, but also that the variation in the bundles is related to the different topics and fields that instructors provided in these different reading passages to EAP learners during writing classes. This issue is clear from the different articles and reading passages instructors submitted in the present study (see Appendix K).

Regarding EAP textbooks, the present research indicated that authors used most of the lexical bundles to express commands, communicating instructions and orders to learners/readers to perform certain tasks (e.g. *focus on your topic* and *use a dictionary to*). In addition, the issue of a limited number of shared bundles in the EAP lists compared to the written AFL sub-list seems to show that EAP materials are less likely to be based on the written AFL sub-corpus. This is also seen in the differences between the EAP materials and the written AFL sub-list in terms of types, some structures, and the pragmatic functions of lexical bundles. Furthermore, the absence of some teaching tasks on bundle types show a gap in language use. Thus, the EAP materials in this study, especially the textbooks, appear to reinforce the calls by Chen (2010), Wood (2010), Wood and Appel (2014) and others that textbooks are not effective in the treatment of formulaic sequences – lexical bundles. The possibility of misrepresentation and underrepresentation of the types, structures and functions of lexical bundles in the present study demonstrates that EAP learners could be misguided in the use of bundles that are required for academic writing. I would contend that the EAP materials provided are ineffective in the pedagogical treatment of lexical bundles found in the examined pre-sessional writing course materials. It is important to note, however, that EAP learners could have gained knowledge on lexical bundles from being exposed to online resources and other supplementary materials provided by textbook publishers or instructors that were not provided in this study.

These findings have clear implications for EAP practitioners, pre-sessional courses, and materials development. I strongly hold the view that lexical bundles need to be reviewed, redefined, clearly introduced, and developed in EAP materials because they are a significant part of the academic writing register.

11.2.1 Implications for instructors teaching lexical bundles, possible EAP pre-sessional courses, and for material designers or development

In the light of studies confirming the importance of formulaic sequences in academic written registers (Biber *et al.*, 2004; Cortes, 2006; Hyland, 2008b), teaching lexical bundles to EAP learners has been widely encouraged and stressed in the literature. According to Hyland (2012: 165) and many others (Nattinger and DeCarrico, 1992; Lewis, 1997; Willis, 2003), “the study of high-frequency strings [lexical bundles] and their possible variation may thus have great pedagogic value to teachers of English for academic purposes”. In addition, the studies conducted by Cortes (2006), Chen (2010), Hyland (2008b), Wood (2010) and Wood and Appel (2014) highlight the importance of raising students’ awareness through explicit instructions regarding lexical bundles. These long-held assumptions are a focal point in the teaching of lexical bundles, and have been widely addressed and recommended. However, in the light of the findings of the present study, lexical bundles in the EAP pre-sessional programme appear to still be neglected and overlooked in some recent EAP materials. EAP learners encounter lexical bundles that are less appropriate for EAP learners wanting to improve their lexical knowledge for academic writing purposes. The main issue that needs to be taken into account, then, is the pedagogical aspect concerning the handling of lexical bundles.

The literature holds two main views when approaching the teaching of the phrasal lexicon: a pedagogical approach focusing on a common core of academic formulas (Simpson-Vlach and Ellis, 2010) versus the discipline-specific pedagogical approach to lexical bundles (Hyland and Tse, 2007; Hyland, 2008b). These two approaches relate to the types of lexical bundles students should be taught. For example, should we provide learners with bundles that are common in multiple fields (e.g. *on the other hand* and *at the same time*)? Or should we present learners with bundles that are needed for their special area of interest (e.g. for applied linguistics learners, we teach learners the bundle *the starting point of* and so on)? This matter could be a main challenge or drawback facing many EAP instructors, EAP pre-sessional programmes, and material designers when constructing materials for writing classes. I would contend that there are many ways to address this difficulty.

In an effort to help students gain the language knowledge relating to lexical bundles, as a starting point, I suggest merging the two above-mentioned pedagogical approaches towards the teaching of lexical bundles. We should take advantage of corpus-informed lists such as the AFL (Appendix B) or the written AFL sub-list (Appendix C), and Hyland (2008b) or Cortes's (2004) ready-made supplementary materials for inclusion in EAP pre-sessional courses, textbooks, and instructors' materials. These and other lists serve as very useful reference tools for EAP learners (Ädel, 2010b).

Regarding EAP instructors and EAP programmes, first, I would recommend that all EAP learners be provided with a *core academic bundles list* (see Appendices B and C) as a required list. Second, depending on the EAP learners' specific disciplines, I propose that the EAP pre-sessional programme or instructors offer a lexical bundle list selection process to their students. This process would be based on EAP learners' specific target contexts, on the basis of which this second list could be determined (see Hyland and Hamp-Lyons, 2002). Therefore, for EAP mixed classes, engineering-based students could work with a list of the most used bundles in engineering (see list in Hyland (2008b)) while a biology list of the most used bundles could be given to biology major students (see list in Cortes (2004)). This way learners could recognize bundles in both their own fields and those common across academic writing. These provided lists should not be treated as finalized lists but as a reference guide to possible lexical bundles, which EAP learners could choose from and further examine in their readings and use in their writings. More importantly, in order to familiarize learners with the lexical bundles that are relevant to their target field of study and which are mostly used by advanced writers in their field, they need repeated exposure by being provided with enough reading passages and ample exercises tasks that include the required bundles. For example, depending on the disciplines of the learners, the teacher may wish to select a number of useful phrases from the designated lists, introducing their functions and structures. Teachers, then, can provide exercises such as fill-in-the-blank and reading articles, focusing on the selected phrases and providing multiple opportunities for learners to encounter them (see Lewis, 1993, 1997, 2000 and Nattinger and Decarrico, 1992) for more exercises on the teaching of phrases).

At a final stage, learners can start writing sentences and/or paragraphs, practising these phrases in writing classes, and teachers may wish to give immediate feedback.

Regarding textbooks and materials designers, I believe that authors need to pay attention to formulaic sequences and provide plenty of exercises and tasks on lexical bundles by taking advantage of these ready-made lists (see Appendices B and C) and including them in their textbook materials. Repeated exposure to a large number of lexical bundles and consciousness-raising tasks (Lewis, 2000), and conducting concordance line checks for information on in-context examples and use (Hyland 2008a) are among the important strategies which could be used to enhance learners' use of academic bundles. I am aware that I have not discussed other aspects of teaching academic writing and did not observe (in class) how instructors handle lexical bundles during teaching academic writing; thus, these and other factors associated with teaching writing are beyond the scope of this study. This is because this study only focused on lexical bundles in EAP materials.

11.2.2 Implications for teacher corpora

The present study also contributes to teacher development through "exploring teacher corpora" (O'Keeffe *et al*., 2007). Teacher corpora are usually built with a reflective purpose, where it is seen as an effective practice often intended to inform teachers about the way they teach as part of their professional development. The data in the teacher corpora is often concerned with classroom interactions compared to media interviews to demonstrate the distinctiveness of this register (O'Keeffe *et al.*, 2007). For example, teacher corpora have included data on teacher talk such as questioning strategies taken from language used in actual classrooms (O'Keeffe and Farr, 2003). Unlike previous data found in teacher corpora, the present research explored a different set of data, including written materials. In future research, instructors' materials could be analysed in a similar way, contributing to the ongoing career development of teachers. Although it is not possible to use the data from the current study in teacher corpora for reasons of participant confidentiality, it is hoped that this study will influence other studies, enabling their collected lexical bundle data to be included in teacher corpora. In so doing, the present study will contribute to the future development of teacher corpora, where teachers' understanding and close reflection on materials related to teaching lexical bundles will be taken into consideration.

11.3 Recommendations for future research

The present study focused on examining lexical bundles in the teaching of academic writing materials in an EAP pre-sessional programme. A number of suggestions have emerged from my present research that call for further studies to be conducted. To begin with, it appears that more work needs to be done to replicate this study on a larger scale and in other contexts. Researchers need to look into more than one EAP programme within and outside the UK, to enhance knowledge of lexical bundles in more varied contexts.

Second, during my study, some key issues arose, exploration of which would add to the understanding on lexical bundles in the academic written register. I make this recommendation because my study was limited to closely examining lexical bundles in EAP materials (textbooks and instructors' materials). For example, future research could focus on classroom observations and learners' interviews on their experience with exploring lexical bundles lists in their academic writing programme to provide a different angle on the treatment of lexical bundles.

The above proposed implication on merging the two approaches towards the teaching of lexical bundles is another interesting issue that emerged during my discussion on the pedagogical outcomes of lexical bundle treatment in EAP materials. The concern regarding a limited range of teaching activities is consistently emphasized in many studies, including the present study; hence, there has not been much evidence reported on the teaching of corpus-derived lists such as the written AFL sub-list and Hyland's (2008b) list. Therefore, I recommend that the above suggested plan for the teaching of lexical bundles during writing classes should be executed and tested, allowing EAP instructors to reflect on its feasibility and usefulness.

11.4 Closing remarks

The original aim of the present study was to conduct a comparative examination of the types, structures and functions of lexical bundles and the written AFL sub-list, from which four shared bundles were identified for analysis. In addition, other findings in this study has clarified the understanding of lexical bundles in the genre of EAP textbooks and instructors' materials. The study has recommended using different lists such as the AFL as a reference tool in the development of textbooks, instructors' materials and in the EAP syllabus. Despite the remarks made in the present study and in many other previous similar studies on the effectiveness of corpus-informed materials, it seems that the textbooks investigated in this study ignored the inclusion of recommended ready-made lists on lexical bundles within the textbooks. Therefore, it is perhaps the responsibility of the EAP course designers and writing instructors to be informed and to be more aware of the usefulness of lexical bundles produced from corpus-driven data. Even more importantly, EAP course designers and writing instructors could strongly urge textbook publishers to include such ready-made lists and to provide plenty of exercises and tasks on lexical bundles. Furthermore, instructors could be directly involved and use these reference tools as mandatory supplementary materials in their writing classes, through providing a sufficient number of tasks. This will lead to an overall improvement in the pedagogical treatment of lexical bundles in EAP materials.

REFERENCES

Ädel, A. (2006) *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamins.

Ädel, A. (2010a) Just to give you kind of a map of where we are going: a taxonomy of metadiscourse in spoken and written academic English. *NJES: Nordic Journal of English Studies*, 9(2), 69–97.

Ädel, A. (2010b) Using corpora to teach academic writing: challenges for the direct approach. *In*: M. Campoy-Cubillo *et al*. (eds) *Corpus-based approaches to English language teaching*. 1st edn. London: Continuum. 39–55.

Ädel, A. and 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.

Ädel, A. and Mauranen, A (2010) Metadiscourse: diverse and divided perspectives. *Nordic Journal of English Studies*, 9(2), 1–11.

Altenberg, B. (1998) On the phraseology of spoken English: the evidence of recurrent word combinations. *In*: Cowie, A. P. (ed) *Phraseology: theory, analysis, and applications*. Oxford: Oxford University Press. 101–122.

Alyousef, H. S. (2015) An investigation of meta discourse features in international postgraduate business students' texts: the use of interactive and interactional markers in tertiary multimodal finance texts. *SAGE Open*, 5(4), 1–10.

Angouri, J. (2010) Quantitative, qualitative or both? Combining methods in linguistic research. *In*: Litosseliti, L. (ed) *Research methods in linguistics.* London: Bloomsbury. 29–45.

Anonymous. (2014) *Academic English Preparatory Course syllabus terms 1, 2 & 3.*

Anthony, L (April 23, 2018a) AntConc (Version 3.5.7) *(Windows, Macintosh OS X, and Linux)*. Available at: http://www.laurenceanthony.net/software/antconc/releases/AntConc357/help.pdf (Accessed: 2 May 2018).

Anthony, L. (2018b) AntConc (Version 3.5.6) [Computer program]. Tokyo, Japan: Waseda University. Available at: http://www.laurenceanthony.net/ (Accessed: 2 February 2015).

Aull, L. L. and Lancaster, Z. (2014) Linguistic markers of stance in early and advanced academic writing: a corpus-based comparison. *Written Communication*, 31(2), 151–183.

Baker, P. (2006). *Using corpora in discourse analysis.* London: Continuum.

Baker, P. (2010) Corpus methods in linguistics. *In*: Litosseliti, L. (ed) *Research methods in linguistics*. London: Bloomsbury. 93–113.

Barton, E. L. (1995) Contrastive and non-contrastive connectives: metadiscourse functions in argumentation. *Written Communication*, 12(2), 219–239.

Becker, J. (1975) The phrasal lexicon. *In*: Nash-Webber, B. and Schank, R. (eds.) *Theoretical issues in natural language processing 1.* Cambridge, MA: Bolt Beranek & Newman report no.3081, AI report no. 28.

Bermel, N. and Knittl, L. (2012) Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech, *Corpus Linguistics and Linguistic Theory*, 8(2), 241–275.

Biber, D. (1993) Representativeness in corpus design. *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing*, 8(4), 243–257.

Biber, D. (2006) *University language: a corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D. (2009) A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3), 275–311.

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

Biber, D. and Conrad, S. (1999) Lexical bundles in conversation and academic prose. *In*: Hasselglad, H. and Oksefjell, S. (ed) *Out of corpora: studies in honor of Stig Johansson*. Amsterdam: Rodopi. 181–189.

Biber, D. and Conrad, S. (2001) Quantitative corpus-based research: much more than bean counting. *TESOL Quarterly*, 35(2), 331–336.

Biber, D. and Reppen, R. (2002) What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition,* 24(2),199–208.

Biber, D. *et al.* (1998) *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

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

Biber, D. *et al.* (2004) If you look at…: lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371–405.

Bolinger, D. (1976) Meaning and memory. *Forum Linguisticum*, 1, 1–14.

Burton, G. (2012). Corpora and course books: destined to be strangers forever? *Corpora*, 7(1), 91–108.

Bybee, P. J. and Hopper, J. (2001) Introduction to frequency and the emergence of linguistic structure. *In*: Bybee, P. J. and Hopper, J. (ed) *Frequency and the emergence of linguistic structure.* Amsterdam: John Benjamins, 1–24.

Bybee, P. J. and Eddington, D. (2006) A usage-based approach to Spanish verbs of 'becoming'. *Language*, 82(2), 323–355.

Bychkovska, T. and Lee, J. (2017). At the same time: lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes*, 30, 38–52.

Byrd, P. and Coxhead, A. (2010) On the other hand: lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL*, 5, 31–64.

Cattell, J. M. (1886) The time it takes to see and name objects, *Mind*, 11(41), 63–65.

Chen, L. (2010) An investigation of lexical bundles in ESP textbooks and electrical engineering introductory textbooks. *In*: Wood, D. (ed) *Perspectives on formulaic language: acquisition and communication*. 1st edn. New York/London: Continuum. 107–125.

Chen, Y. and Baker, P. (2010) Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, 14(2), 30–49.

Chen, Y. and Baker, P. (2016). Investing critical discourse features across second language development lexical bundles in rated learner essays, CEFER B1, B2 and C1. *Applied Linguistics*, 37(6), 849–880.

Chomsky, N. (1964). *Current issues in linguistic theory.* London: Mouton & Co.

Conklin, K. and Schmitt, N. (2008) Formulaic sequences: are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, 29(1), 72–89.

Conrad, S. (1996). Investing academic texts with corpus-based techniques: an example from biology. *Linguistics and Education*, 8, 299–326.

Cortes, V. (2004) Lexical bundles in published and student disciplinary writing: examples from history and biology. *English for Specific Purposes*, 23(4), 397–423.

Cortes, V. (2006) Teaching lexical bundles in the disciplines: an example from a writing intensive history class. *Linguistics and Education*, 17(4), 391–406.

Cowie, A. P. (1998) *Phraseology: theory, analysis, and applications*, Oxford: Clarendon: Oxford University Press.

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

Coxhead, A. (2008) Phraseology and English for Academic Purposes: challenges and opportunities. *In*: Meunier, F. and Granger, S. (eds) *Phraseology in foreign language learning and teaching.* Amsterdam/Philadelphia: John Benjamins, 149–159.

Coxhead, A. and Byrd, P. (2007) Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing*, 16(3), 129–147.

Coxhead, A. and Nation, P. (2001). The specialised vocabulary of English for Academic Purposes. *In*: Flowerdew, J. (ed) *Research perspectives on English for Academic Purposes*. Cambridge: Cambridge University Press.

252–267.

Crismore, A. *et al.* (1993) Metadiscourse in persuasive writing: a study of texts written by American and Finnish university students. *Written Communication*, 10(1), 39–71.

Csomay, E. (2012) Lexical bundles in discourse structure: a corpus-based study of classroom discourse. *Applied Linguistics*, 34(3), 369–388.

De Cock, S. (1998) A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics*, 3(1), 59–80.

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

Dörnyei, Z. (2007) *Research methods in applied linguistics*, 1st edn. Oxford: Oxford University Press.

Ellis, N. C. (1996) Sequencing in SLA: phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, 18(1), 91–126.

Ellis, N. C. (2008) The periphery and the heart of language. *In*: Meunier, F. and Granger, S. (eds.) *Phraseology in foreign language learning and teaching.* Amsterdam: John Benjamins. 1–13.

Ellis, N. C. *et al*. (2008) Formulaic language in native and second language speaker psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, 42(3), 375-96.

Erman, B. (2007) Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics*, 12(1), 25–53.

Erman, B. and Warren, B. (2000) The idiom principle and the open choice principle. *Text*, 20(1), 29–62.

Evison, J. M. (2010) What are the basics of analyzing a corpus? *In*: McCarthy, M. J and O'Keeffe, A. (eds) *The Routledge handbook of corpus linguistics*. Abingdon: Routledge. 122–135.

Firth, J. R. (1951) Modes of meaning. *Essays and studies (The English*

Association), *1934–1951*. London: Oxford University Press. 118–149.

Flowerdew, L. (2004) The argument for using English specialized corpora to understand academic and professional settings. *In*: Connor, U. and Upton, T. (eds) *Discourse in the professions: perspectives from corpus linguistics*. Amsterdam: John Benjamins. 11–33.

Flowerdew, L. (2012) *Corpora and language education.* Basingstoke: Palgrave Macmillan.

Flowerdew, L. (2015) Corpus-based research and pedagogy in EAP: from lexis to genre. *Language Teaching*, 48(1), 99–116.

Foster, P. (2001) Rules and routines: a consideration of their role in the task-based language production of native and non-native speakers. *In*: Bygate, M. *et al*. (eds) *Researching pedagogic tasks: second language learning, teaching, and testing*. San Francisco: Pearson Education. 75–93.

Gardner, D. and Davies, M. (2014) A new academic vocabulary list. *Applied Linguistics*, 35(3), 305–327.

Gouverneur, C. (2008) The phraseological patterns of high-frequency verbs in advanced English for general purposes: a corpus-driven approach to EFL textbook analysis. *In*: Meunier, F. and Granger, S. (eds) *Phraseology in foreign language learning and teaching.* Amsterdam: John Benjamins. 223–243.

Grabe, W. and Zhang, C. (2016) Reading-writing relationships in first and second language academic literacy development. *Language Teaching: Surveys and Studies*, 49(3), 339–355.

Granger, S. (2002) A bird's eye view of learner corpus research. *In*: S. Granger *et al*. (eds) *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins. 3–33.

Granger, S. and Meunier, F. (2008b) Phraseology in language learning and teaching: where to from here? In Meunier, F. and Granger, S. (eds) *Phraseology in foreign language learning and teaching.* Amsterdam: John Benjamins. 247–252.

Halliday, M. A. K. (1978) *Language as social semiotic.* London: Edward Arnold.

Harwood, N. (2002). Taking a lexical approach to teaching principles and problems. *International Journal of Applied Linguistics*, 12(2), 139–155.

Harwood, N. (2005) What do we want EAP teaching materials for? *Journal of English for Academic Purposes*, 4(2), 149–161.

Harwood, N. (2010). *English language teaching materials: theory and practice.* Cambridge: Cambridge University Press.

Harwood, N. (2014). *English language teaching textbooks: content, consumption and Production.* Basingstoke: Palgrave Macmillan.

Hernandez, S. (2013) Lexical bundles in three oral corpora of university students. *Nordic Journal of English Studies*, 13(1), 187–209.

Hewings, M and Thaine, C. (2012) *Cambridge Academic English: An integrated skills course for EAP*. Cambridge: Cambridge University Press.

Hewings, M. (2012). *Cambridge Academic English: An integrated skills course for EAP*. Cambridge: Cambridge University Press.

Hilpert, M. and Diessel, H. (2017) Entrenchment in construction grammar. *In*: Schmid, H. J. (ed) *Entrenchment and the psychology of language learning.* Washington: American Psychological Association, 57–74.

Hoey, M. (2005) *Lexical priming: a new theory of words and language*. Abingdon: Routledge.

HP Customer Support - Software and Driver Downloads (n.d.) *Readiris Pro 15*. Available at: https://support.hp.com/emea_middle_east-en/drivers/selfservice/hp-deskjet-ink-advantage-4620-e-all-in-one-printer-series/5157451/model/5157452 (Accessed: 1 December 2016).

Hyland, K and Tse, P. (2004) Metadiscourse in academic writing: a reappraisal. *Applied Linguistics*, 25(5), 156–177.

Hyland, K. (1994) Hedging in academic writing and EAP textbooks. *English for Specific Purposes*, 13, 239–256.

Hyland, K. (1999) Academic attribution: citation and the construction of disciplinary knowledge. *Applied Linguistics*, 20(3), 341–367.

Hyland, K. (2002) Directives: argument and engagement in academic writing. *Applied Linguistics*, 23, 215–239.

Hyland, K. (2004) *Disciplinary discourse: social interactions in academic writing*. Ann Arbor, MI: University of Michigan Press.

Hyland, K. (2005) *Metadiscourse: exploring interaction in writing*. London/New York: Continuum.

Hyland, K. (2008a) Academic clusters: text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, 18(1), 41–62.

Hyland, K. (2008b) As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4–21.

Hyland, K. (2009) *Academic discourse*. London: Continuum.

Hyland, K. (2010) Metadiscourse: mapping interactions in academic writing. *Nordic Journal of English Studies*, 9(2), 125–143.

Hyland, K. (2012) Bundles in academic discourse. *Annual Review of Applied Linguistics*, 32, 150–169.

Hyland, K. (2017) 'Metadiscourse: What is it and where is it going?', *Journal of Pragmatics,* 113, 16–29.

Hyland, K. and Hamp-Lyons, L. (2002) EAP: issues and directions. *Journal of English for Academic Purposes*, 1, 1–12.

Hyland, K. and Tse, P. (2007) Is there an "academic vocabulary"? *TESOL Quarterly*, 41(2), 235–253.

Jablonkai, R. (2010) English in the context of European integration: a corpus-driven analysis of lexical bundles in English EU documents. *English for Specific Purposes*, 29, 253–267.

Jespersen, O. (1924) *The philosophy of grammar*. London: George Allen & Unwin.

Jiang, N. and Nekrasova, T. M. (2007) The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, 91(3), 433–445.

Jones, M. and Haywood, S. (2004) Facilitating the acquisition of formulaic sequences: an exploratory study in an EAP context. *In*: N. Schmitt (ed) *Formulaic sequences.* Amsterdam: John Benjamins. 269–292.

Jordan, R. R. (1990) *Academic writing course*, 2nd edn. London: Collins ELT.

Karabacak, E. and Qin, J. (2013) Comparison of lexical bundles used by Turkish, Chinese, and American university students. *Procedia – Social and Behavioral Sciences*, 70, 622–628.

Kazemi, M. *et al*. (2014) The impact of teaching lexical bundles on improving Iranian EFL students' writing skills. *Procedia – Social and Behavioral Sciences*, 98, 864–869

Khedri, M. *et al.* (2013) An exploration of interactive metadiscourse makers in academic research article abstracts in two disciplines. *Discourse Studies*, 15(3), 319–331.

Koester, A. (2010). Building small specialised corpora. *In*: O'Keeffe, A and McCarthy, M. (eds) *The Routledge handbook of corpus linguistics.* Abingdon: Routledge.

Koprowski, M. (2005) Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal*, 59(4), 322–332.

Kuhi, D. and Behnam, B. (2011) Generic variations and metadiscourse use in the writing of applied linguists: a comparative study and preliminary framework. *Written Communication*, 28(1), 97–141.

Laviosa, S. *et al.* (2017) *Textual and contextual analysis in empirical translation studies.* WorldCat [Online]. Available at: http://public.eblib.com/choice/publicfullrecord.aspx?p=4615420 (Accessed: 25 June 2018).

Lee, D. (2001) Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, 5(3), 37–72.

Levy, S. (2008) *Lexical bundles in professional and student writing*. Berlin: VDM Verlag.

Lewis, M. (1993) *The lexical approach: the state of ELT and the way forward*. Hove: Language Teaching Publications.

Lewis, M. (1997) *Implementing the lexical approach*. Hove: Language Teaching

Publications.

Lewis, M. (ed) (2000) *Teaching collocation: further developments in the lexical approach*. Hove: Language Teaching Publications.

Li, J. and Schmitt, N. (2009) The acquisition of lexical phrases in academic writing: a longitudinal case study. *Journal of Second Language Writing*, 18(2), 85–102.

Liu, D. (2012) The most frequently-used multi-word constructions in academic written English: a multi-corpus study. *English for Specific Purposes*, 31(1), 25–35.

Mair, C. (2006) *Twentieth century English: history, variation and standardization*. Cambridge: Cambridge University Press.

Martinez, R. (2013) A framework for the inclusion of multi-word expressions in ELT. *ELT Journal*, 67(2), 184–198.

Martinez, R. and Schmitt, N. (2012) A phrasal expressions list*.* *Applied Linguistics*, 33(3), 299–320.

Mauranen, A. (1993) Contrastive ESP rhetoric: metatext in Finnish-English Economics texts. *English for Specific Purposes,* 12(1), 3–22.

McCulley, G. A. (1985) Writing quality, coherence, and cohesion. *Research in the Teaching of English*, 19(3), 269–282.

McEnery, T. and Hardie, A. (2012). *Corpus linguistics.* Cambridge: Cambridge University Press.

McEnery, T. and Wilson, A. (2001) *Corpus linguistics: an introduction*, 2nd edn. Edinburgh: Edinburgh University Press.

McIntosh, C. *et al.* (eds) (2009) *Oxford collocations dictionary for students of English*. Oxford: Oxford University Press.

McKee, S. (2012) Managing transitions into UK higher education: experiences of international students from a pre-sessional English course. *Investigations in University Teaching and Learning*, 8, 30–38.

Miao, H. (2014) An investigation of formulaic sequences in multi-modal Chinese college English textbooks. *Journal of Language Teaching and Research*, 5(6), 1308–1314.

Millar, N. (2011) The processing of malformed formulaic language. *Applied Linguistics*, 32(2), 12948.

Miller, G. (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97.

Nattinger, J. and DeCarrico, J. (1992) *Lexical phrases and language teaching*. Oxford: Oxford University Press.

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

Nelson, G. *et al.* (2002) *Exploring natural language: working with the British component of the international corpus of English*. Amsterdam: John Benjamins.

Nelson, M. (2010). Building a written corpus: what are the basics? *In*: O'Keeffe, A. and McCarthy, M. (eds) *The Routledge handbook of corpus linguistics.* Abingdon: Routledge.

O'Keeffe, A. *et al.* (2007) *From corpus to classroom: language use and language teaching*. Cambridge: Cambridge University Press.

Oakes, M. P. (1998) *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

O'Donnell, M. B. *et al.* (2013) The development of formulaic sequences in first and second language writing. *International Journal of Corpus Linguistics*, 18(1), 83–108.

O'Keeffe, A. and Farr, F. (2003) Using language corpora in language teacher education: pedagogic, linguistic and cultural insights. *TESOL Quarterly*, 37(3), 389–418.

Oshima, A. and Hogue, A. (1999) *Writing academic English*, 3rd edn. London: Longman.

Paltridge, B. (2004) Academic writing. *Language Teaching*, 37(2), 87–105.

Pawley, A. and Syder, F. H. (1983) Two puzzles for linguistic theory: native-like selection and nativelike fluency. *In*: Richards, J. C. and Schmidt, R. W. (eds) *Language and communication*. London: Longman. 191–230.

Quirk, R. *et al*. (1985) *A comprehensive grammar of the English language*. New York: Longman Group Limited.

Reppen, R. (2010). Building a corpus: what are the key considerations? *In*: O'Keeffe, A. and McCarthy, M. (eds) *The Routledge handbook of corpus linguistics.* Abingdon: Routledge.

Robinson, P. *et al*. (2014) Attention and awareness in second language acquisition. *In*: Grass, S. M. and Mackey, A. (eds) *The Routledge handbook of second language acquisition.* London: Routledge. 247–267.

Römer, U. (2010) Using general and specialized corpora in English language teaching: past, present and future. *In*: Campoy-Cubillo, M. C. *et al.* (eds) *Corpus-based approaches to English language teaching.* London: Continuum. 18–38.

Römer, U. and Wulff, S. (2010) Applying corpus methods to written academic texts: exploration of MICUSP. *Journal of Writing Research*, 2(2), 99–127.

Salazar, D. (2011) Lexical bundles in scientific English: a corpus-based study of native and non-native writing**.** Unpublished doctoral dissertation, Universitat de Barcelona.

Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing*. Amsterdam: John Benjamins.

Sánchez, P. (2014) Lexical bundles in biology: differences between textbooks and research articles. *Revista Española de Lingüística Aplicada*, 27(2), 492–513.

Saussure, F. de [1916] (1966) *Course in general linguistics*. New York: McGraw-Hill.

Schmid, H. J. (2017) A framework for understanding linguistic entrenchment and its psychological foundations. *In*: Schmid, H. J. (ed) *Entrenchment and the psychology of language learning*. Washington: American Psychological Association, 9–35.

Schmidt, R (1993) Awareness and second language acquisition. *Annual Review of Applied Linguistics*, 13, 206–226.

Schmidt, R. (1995) *Attention and awareness in foreign language learning*. Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.

Schmitt, N. and Carter, N. (2004) Formulaic sequences in action: an introduction. *In*: Schmitt, N. (ed) *Formulaic sequences.* Amsterdam: John Benjamins. 1–22.

Schmitt, N. *et al.* (2004) Knowledge and acquisition of formulaic sequences: a longitudinal study. *In*: Schmitt, N. (ed) *Formulaic sequences.* Amsterdam: John Benjamins. 55–86.

Sellen, D. (1982) *Skills in action*. Cheltenham: Hulton Educational.

Shahrokhi, M. and Moradmand, S. (2014) A comparative study of the use of collocation in Iranian high school textbooks and American English file books. *International Journal of Applied Linguistics & English Literature*, 3(3), 58–64.

Shaw, P. and Liu, E. (1998) What develops in the development of second-language writing? *Applied Linguistics*, 19(2), 225–254.

Simpson-Vlach, R. and Ellis, N. C (nd.) *Supplementary material is available at Applied Linguistics online.* Available at: https://academic.oup.com/applij/search-results?page=1&q=Simpson-Vlach%2C%20R.%20and%20Ellis&fl_SiteID=5135&SearchSourceType=1&allJournals=1 (Accessed: 2 October 2013).

Simpson-Vlach, R. and Ellis, N. C. (2010) An Academic Formulas List: new methods in phraseology research. *Applied Linguistics*, 31(4), 487–512.

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

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

Sinclair, J. (2005) Corpus and text: basic principles. *In*: Wynne, M. (ed) *Developing linguistic corpora: a guide to good practice*. Oxford: Oxbow Books. 1–16.

Stefanowitsch, A. and Gries, S. T. (2003) Collostructions: investigating the interaction of words and constructions, *International Journal of Corpus Linguistics*, 8(2), 209–243.

Storch, N. and Tapper, N (2009) The impact of an EAP course on postgraduate writing. *Journal of English for Academic Purposes*, 8, 207–223.

Stubbs, M. (2007) Quantitative data on multi-word sequences in English: the case of the word "world". *In*: Hoey, M. *et al*. (eds) *Text, discourse and corpora: theory and analysis.* London: Continuum. 163–189.

Sunderland, J. (2010) Research questions in linguistics. *In*: Litosseliti, L. (ed) *Research methods in linguistics.* London: Bloomsbury. 10–28.

Swales, J. M. and Feak, C. B (2000) *English in today's research world: a writing guide*. Ann Arbor, MI: University of Michegan Press.

Swales, J. M. (2002) Integrated and fragmented worlds: EAP materials and corpus linguistics. *In*: Flowerdew (ed) *Academic Discourse*. London: Longman. 150–164.

Tang, J. (2012) An empirical study on the effectiveness of the lexical approach to improving writing in SLA. *Journal of Language Teaching and Research*, 3(3), 578–583.

Terraschke, A. and Wahid, R. (2011) The impact of EAP study on the academic experiences of international postgraduate students in Australia. *Journal of English for Academic Purposes*, 10, 173–182.

Thaine, C. (2012). *Cambridge Academic English: an integrated skills course for EAP*. Cambridge: Cambridge University Press.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.

Tomlinson, B. (2011). *Materials development in language* teaching, 2nd edn. Cambridge: Cambridge University Press.

Tribble, C. (2015) Writing academic English further along the road. What is happening now in EAP writing instruction? *ELT Journal*, 69(4), 442–462.

Tsai, K. (2014) Profiling the collocation use in ELT textbooks and learner writing. *Language Teaching Research. Special Issue: Vocabulary Research and Pedagogy*, 1–18.

Underwood, G. *et al*. (2004) The eyes have it: an eye-movement study into the processing of formulaic sequences. *In*: Schmitt, N. (ed) *Formulaic sequences*. Amsterdam: John Benjamins. 153–168.

Vande Kopple, W. J (1985) Some exploratory discourse on metadiscourse. *College Composition and Communication*, 36(1), 82–93.

Wei, Y. and Lei, L. (2011) Lexical bundles in the academic writing of advanced Chinese EFL learners. *RELC Journal*, 42(2), 155–166.

Walter, E. *et al.* (3rd eds) (2008) *Cambridge Advanced Learner's Dictionary*. Cambridge: Cambridge University Press.

White, R. and McGovern, D. (1994) *Writing*. Mahwah, NJ: Prentice Hall.

Willis, D. (1990). *The lexical syllabus: a new approach to language teaching.* London: Collins Cobuild.

Willis, D. (2003) *Rules, patterns and words: grammar and lexis in English language teaching*. Cambridge: Cambridge University Press.

Wood, D. (2010) Lexical clusters in an EAP textbooks corpus. *In*: Wood, D. (ed) *Perspectives on formulaic language: acquisition and communication*. London/New York: Continuum. 88–106.

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

Wray, A. (1999) Formulaic language in learners and native speakers. *Language Teaching*, 32(4), 213–231.

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

Wray, A. and Perkins, R. P. (2000) The functions of formulaic language: an integrated model. *Language & Communication*, 20(1), 1–28.

Zarifi, A. and Mukundan, J. (2013) Selection and presentation of phrasal verbs in ESL textbooks. *Theory and Practice in Language Studies*, 3(10), 1821–1829.

Zimmerman, C. B. and Schmitt, N. (2005) Lexical questions to guide the teaching and learning of words. *The CATESOL Journal*, 17(1), 1–7.

Appendix A
Research questions

1. What are the most frequent four-word bundles profiled in textbooks and instructors' materials, in an EAP pre-sessional programme in the UK particularly aimed at teaching academic writing?

2. How are the most frequent four-word bundles classified functionally and structurally?

3. To what extent are the identified bundles in the EAP textbooks and instructors' materials based on data retrieved from corpus-driven lists such as the Academic Formulas List (AFL) (Simpson-Vlach and Ellis, 2010), in terms of the frequencies, structures and functions of the most frequent four-word lexical bundles?

4. Where in the EAP materials texts do the identified lexical bundles appear (for example, in readings and/or in instructions accompanying these readings)?

5. Do the EAP textbooks and instructors' materials present the identified four-word lexical bundles in tasks and exercises, for EAP learners?

Appendix B

Written AFL top 200 (sorted by two-factor FTW scores; adapted from Simpson-Vlach and Ellis, 2010)

Supplementary material is available at Applied Linguistics online.

		Speech		Writing		
		Raw freq.	Freq. per million words	Raw freq.	Freq. per million words	FTW
1.	on the other hand	86	40	251	119	2.84
2.	due to the fact that	5	2	27	13	2.64
3.	on the other hand the	6	3	50	24	2.55
4.	it should be noted	0	0	36	17	2.51
5.	it is not possible to	1	0	31	15	2.44
6.	a wide range of	9	4	66	31	2.42
7.	there are a number of	11	5	30	14	2.41
8.	in such a way that	20	9	23	11	2.32
9.	take into account the	5	2	24	11	2.27
10.	as can be seen	0	0	32	15	1.79
11.	it is clear that	6	3	69	33	1.72
12.	take into account	17	8	41	19	1.70
13.	can be used to	11	5	95	45	1.64
14.	in this paper we	0	0	29	14	1.64
15.	are likely to	16	7	129	61	1.61
16.	in the next section	0	0	32	15	1.60
17.	a large number of	16	7	47	22	1.59
18.	the united kingdom	2	1	54	25	1.57
19.	on the basis of the	8	4	48	23	1.57
20.	that there is no	10	5	67	32	1.56
21.	over a period of	10	5	27	13	1.55
22.	as a result of the	11	5	35	17	1.55
23.	can be seen in	1	0	36	17	1.52
24.	a wide range	13	6	69	33	1.51
25.	there are a number	13	6	30	14	1.47
26.	it is interesting to	0	0	32	15	1.47
27.	it is impossible to	1	0	25	12	1.47
28.	it is obvious that	0	0	23	11	1.46
29.	it is possible to	5	2	101	48	1.46
30.	it is not possible	2	1	38	18	1.45
31.	been carried out	1	0	37	17	1.45
32.	can be found in	0	0	39	18	1.45
33.	it is important to	3	1	92	43	1.40
34.	was carried out	1	0	56	26	1.39
35.	is likely to be	7	3	81	38	1.38
36.	wide range of	10	5	77	36	1.37
37.	the same way as	10	5	32	15	1.37

38.	due to the fact	5	2	27	13	1.36
39.	in accordance with the	4	2	26	12	1.36
40.	it is necessary to	2	1	56	26	1.35
41.	the other hand	88	41	254	120	1.35
42.	can be seen	12	6	185	87	1.35
43.	it is likely that	0	0	39	18	1.31
44.	such a way that	20	9	23	11	1.22
45.	to carry out	16	7	62	29	1.22
46.	it is possible that	1	0	40	19	1.22
47.	with respect to the	13	6	78	37	1.20
48.	give rise to	7	3	41	19	1.18
49.	carried out by	4	2	43	20	1.17
50.	whether or not the	6	3	38	18	1.13
51.	in the present study	0	0	23	11	1.11
52.	should be noted	0	0	38	18	1.07
53.	be carried out	3	1	38	18	1.06
54.	the other hand the	6	3	51	24	1.06
55.	does not appear	3	1	27	13	1.04
56.	his or her	6	3	71	34	1.01
57.	is not possible to	1	0	32	15	0.99
58.	shown in figure	0	0	84	40	0.96
59.	be used as a	1	0	36	17	0.95
60.	for the purposes of	3	1	50	24	0.95
61.	be regarded as	2	1	85	40	0.94
62.	to ensure that the	0	0	37	17	0.93
63.	allows us to	16	7	32	15	0.93
64.	it has been	26	12	168	79	0.92
65.	little or no	6	3	33	16	0.90
66.	carried out in	1	0	53	25	0.90
67.	to distinguish between	2	1	45	21	0.88
68.	in accordance with	12	6	55	26	0.88
69.	they do not	13	6	118	56	0.88
70.	at this stage	14	7	70	33	0.88
71.	is based on the	7	3	47	22	0.88
72.	shown in table	0	0	63	30	0.87
73.	in the absence of	10	5	86	41	0.86
74.	we have seen	11	5	56	26	0.83
75.	to determine whether	4	2	33	16	0.82
76.	in the context of	16	7	121	57	0.79
77.	a high degree	3	1	28	13	0.78
78.	the difference between the	18	8	30	14	0.78
79.	an increase in the	12	6	28	13	0.78
80.	it is possible	12	6	175	83	0.77
81.	can be achieved	0	0	36	17	0.77
82.	insight into the	0	0	34	16	0.77
83.	can be expressed	3	1	49	23	0.75
84.	we assume that	10	5	43	20	0.75
85.	they did not	12	6	56	26	0.73
86.	there has been	18	8	70	33	0.72
87.	on the part of	17	8	66	31	0.70

88.	in this paper	9	4	132	62	0.70
89.	the purpose of this	4	2	28	13	0.70
90.	less likely to	11	5	48	23	0.68
91.	a large number	19	9	49	23	0.67
92.	can easily be	0	0	32	15	0.67
93.	with regard to	9	4	85	40	0.66
94.	there are several	12	6	38	18	0.66
95.	over a period	10	5	30	14	0.66
96.	in this case the	17	8	57	27	0.66
97.	in conjunction with	12	6	48	23	0.65
98.	at the time of	14	7	68	32	0.65
99.	we do not	8	4	81	38	0.64
100.	has been used	8	4	43	20	0.63
101.	appears to be	19	9	113	53	0.63
102.	to do so	49	23	116	55	0.63
103.	there are no	46	21	82	39	0.62
104.	on the other	166	77	311	147	0.62
105.	has also been	3	1	53	25	0.61
106.	it is worth	0	0	42	20	0.61
107.	can be found	2	1	69	33	0.61
108.	the next section	2	1	41	19	0.60
109.	are a number of	12	6	30	14	0.60
110.	this paper we	0	0	34	16	0.60
111.	be seen as	18	8	94	44	0.60
112.	be related to the	3	1	26	12	0.59
113.	to ensure that	11	5	94	44	0.59
114.	it is important	6	3	139	66	0.59
115.	be explained by	0	0	32	15	0.58
116.	same way as	11	5	32	15	0.58
117.	see for example	0	0	42	20	0.58
118.	the presence of a	3	1	50	24	0.58
119.	that it is not	7	3	37	17	0.58
120.	in some cases	40	19	68	32	0.58
121.	to the fact that	21	10	49	23	0.57
122.	high levels of	12	6	35	17	0.56
123.	most likely to	6	3	55	26	0.56
124.	it appears that	13	6	61	29	0.56
125.	it follows that	2	1	65	31	0.55
126.	can also be	13	6	111	52	0.55
127.	it is clear	6	3	83	39	0.54
128.	by virtue of	13	6	54	25	0.54
129.	the most important	46	21	112	53	0.53
130.	an attempt to	25	12	62	29	0.53
131.	it is impossible	2	1	36	17	0.53
132.	factors such as	0	0	29	14	0.53
133.	is consistent with	1	0	61	29	0.53
134.	total number of	5	2	42	20	0.53
135.	similar to those	0	0	47	22	0.52
136.	as part of the	17	8	55	26	0.52
137.	can be considered	0	0	38	18	0.52
138.	at the outset	6	3	24	11	0.51
139.	in more detail	7	3	27	13	0.51
140.	should not be	13	6	108	51	0.51
141.	could be used	9	4	41	19	0.51

142.	appear to be	15	7	99	47	0.50
143.	as a consequence	6	3	50	24	0.50
144.	in this article	6	3	59	28	0.50
145.	assumed to be	3	1	82	39	0.49
146.	in the form of	19	9	98	46	0.48
147.	as a whole	57	26	92	43	0.48
148.	important role in	5	2	28	13	0.47
149.	it is interesting	2	1	38	18	0.46
150.	does not have	20	9	52	25	0.46
151.	none of these	12	6	32	15	0.46
152.	as shown in	1	0	139	66	0.45
153.	is likely to	19	9	169	80	0.45
154.	this means that	13	6	77	36	0.45
155.	be noted that	0	0	45	21	0.45
156.	be achieved by	0	0	28	13	0.45
157.	depends on the	39	18	93	44	0.44
158.	at least in	40	19	75	35	0.44
159.	a small number	9	4	25	12	0.43
160.	in table 1	0	0	62	29	0.43
161.	in most cases	7	3	37	17	0.43
162.	depending on the	30	14	62	29	0.41
163.	in both cases	11	5	36	17	0.41
164.	the validity of the	2	1	39	18	0.41
165.	small number of	10	5	38	18	0.40
166.	their ability to	16	7	40	19	0.40
167.	need not be	1	0	54	25	0.40
168.	needs to be	64	30	96	45	0.40
169.	have shown that	4	2	63	30	0.39
170.	it is necessary	5	2	71	34	0.39
171.	been shown to	5	2	66	31	0.39
172.	such as those	1	0	44	21	0.39
173.	are as follows	1	0	34	16	0.38
174.	for this purpose	3	1	31	15	0.38
175.	is determined by	7	3	48	23	0.38
176.	it is difficult	0	0	57	27	0.37
177.	even though the	18	8	44	21	0.37
178.	this does not	9	4	59	28	0.37
179.	was based on	16	7	40	19	0.37
180.	the nature of the	18	8	91	43	0.37
181.	in the course of	28	13	58	27	0.37
182.	degree to which	3	1	56	26	0.37
183.	be argued that	1	0	36	17	0.36
184.	in terms of a	18	8	32	15	0.36
185.	for this reason	6	3	44	21	0.36
186.	are based on	19	9	50	24	0.36
187.	in a number of	15	7	40	19	0.36
188.	two types of	14	7	45	21	0.34
189.	the total number	8	4	39	18	0.34
190.	is more likely	11	5	41	19	0.34
191.	which can be	14	7	120	57	0.34
192.	are able to	14	7	79	37	0.32
193.	be considered as	0	0	46	22	0.32
194.	be used to	18	8	163	77	0.31
195.	b and c	11	5	37	17	0.31

196.	depend on the	16	7	63	30	0.30
197.	is that it is	7	3	41	19	0.30
198.	is affected by	1	0	24	11	0.30
199.	should also be	4	2	38	18	0.30
200.	if they are	22	10	70	33	0.30

Appendix C

Modified written AFL sub-list

Total No. of N-Gram Types: 57

Total No. of N-Gram Tokens: 1400 per million words
Total No. of N-Gram Tokens: 2970 raw frequency

Rank	Four-word lexical bundles	Raw frequency	Freq. per million words
1.	*on the other hand*	251	119
2.	*it should be noted*	36	17
3.	*a wide range of*	66	31
4.	*take into account the*	24	11
5.	*as can be seen*	32	15
6.	*it is clear that*	69	33
7.	*can be used to*	95	45
8.	*in the next section*	32	15
9.	*in this paper we*	29	14
10.	*a large number of*	47	22
11.	*that there is no*	67	32
12.	*over a period of*	27	13
13.	*can be seen in*	36	17
14.	*there are a number*	30	14
15.	*it is interesting to*	32	15
16.	*it is impossible to*	25	12
17.	*it is obvious that*	23	11
18.	*it is possible to*	101	48
19.	*it is not possible*	38	18
20.	*can be found in*	39	18
21.	*it is important to*	92	43
22.	*is likely to be*	81	38
23.	*the same way as*	32	15
24.	*due to the fact*	27	13
25.	*in accordance with the*	26	12
26.	*it is necessary to*	56	26
27.	*it is likely that*	39	18
28.	*such a way that*	23	11
29.	*it is possible that*	40	19
30.	*with respect to the*	78	37
31.	*whether or not the*	38	18
32.	*in the present study*	23	11
33.	*the other hand the*	51	24
34.	*is not possible to*	32	15
35.	*be used as a*	36	17
36.	*for the purposes of*	50	24
37.	*to ensure that the*	37	17
38.	*is based on the*	47	22
39.	*in the absence of*	86	41
40.	*in the context of*	121	57
41.	*the difference between the*	30	14
42.	*an increase in the*	28	13
43.	*on the part of*	66	31

44.	*the purpose of this*	28	13
45.	*in this case the*	57	27
46.	*at the time of*	68	32
47.	*be related to the*	26	12
48.	*the presence of a*	50	24
49.	*to the fact that*	49	23
50.	*as part of the*	55	26
51.	*in the form of*	98	46
52.	*the validity of the*	39	18
53.	*the nature of the*	91	43
54.	*in the course of*	58	27
55.	*in terms of a*	32	15
56.	*in a number of*	40	19
57.	*is that it is*	41	19

Appendix D
Functional categorization of the four-word bundles in the written AFL sub-list in terms of frequency per million words

Categories	No.	Frequency per million words	Lexical bundles
1 Referential expressions			
	(1) Specification of attributes **(a) Intangible framing attributes**		
	1.	57	*in the context of (core)*
	2.	46	*in the form of*
	3.	43	*the nature of the (core)*
	4.	41	*in the absence of*
	5.	37	*with respect to the (core)*
	6.	31	*on the part of*
	7.	27	*in the course of*
	8.	27	*in this case the*
	9.	24	*the presence of a (core)*
	10.	22	*is based on the*
	11.	18	*the validity of the*
	12.	15	*in terms of a (core)*
	13.	12	*in accordance with the*
	14.	11	*such a way that*
	(b) Tangible framing attributes		
	15.	26	*as part of the (core)*
	16.	13	*an increase in the*
	17.	13	*over a period of*
	(c) Quantity specification		
	18.	31	*a wide range of*
	19.	22	*a large number of*
	20.	19	*in a number of*
	21.	14	*there are a number*
	(2) Identification and focus		
	22.	32	*that there is no*
	23.	19	*is that it is*
	24.	17	*can be seen in*
	25.	15	*as can be seen*
	(3) Contrast and comparison		
	26.	119	*on the other hand*

	27.	24	*the other hand the*
	28.	15	*the same way as*
	29.	14	*the difference between the*
	30.	12	*be related to the*
	(4) Deictics and locatives		
	31.	32	*at the time of*
Total	31	848	
2 Stance expressions			
	(1) Hedges		
	1.	38	*is likely to be*
	2.	18	*it is likely that*
	(2) Obligation and directive		
	3.	17	*it should be noted*
	4.	17	*to ensure that the*
	5.	11	*take into account the*
	(3) Expressions of ability and possibility		
	6.	48	*it is possible to*
	7.	45	*can be used to (core)*
	8.	19	*it is possible that*
	9.	18	*can be found in*
	10.	18	*it is not possible*
	11.	17	*be used as a*
	12.	15	*is not possible to*
	(4) Evaluation		
	13.	43	*it is important to*
	14.	33	*it is clear that*
	15.	26	*it is necessary to*
	16.	15	*it is interesting to*
	17.	12	*it is impossible to*
	18.	11	*it is obvious that*
Total	18	421	
3 Discourse organizing functions			
	(1) Metadiscourse/ Textual reference		
	1.	15	*in the next section*
	2.	14	*in this paper we*
	3.	11	*in the present study*
	(2) Topic elaboration: cause and effect		
	4.	24	*for the purposes of*
	5.	23	*to the fact that*
	6.	18	*whether or not the (core)*
	7.	13	*due to the fact*
	8.	13	*the purpose of this*
Total	8	131	
Total	57	1400	

Appendix E

Functional distribution of the four-word bundles in the written AFL sub-list

Functions	Types	%
REFERENTIAL EXPRESSIONS		
Intangible framing attributes (Specification of attributes)	14	24%
Tangible framing attributes	3	5%
Quantity specification	4	7%
Identification and focus	4	7%
Contrast and comparison	5	9%
Deictics and locatives	1	2%
STANCE EXPRESSIONS		
Hedges	2	4%
Obligation and directive	3	5%
Expressions of ability and possibility	7	12%
Evaluation	6	11%
DISCOURSE ORGANIZING FUNCTIONS		
Metadiscourse and textual reference	3	5%
Topic elaboration: cause and effect	5	9%
Total	57	100%

Appendix F
Structural categorization of the four-word bundles in the written AFL sub-list in terms of frequency per million words

Categories	No.	Frequency	Lexical bundles
(1) NP-based			
1	Noun phrase with *of*-phrase fragment		
	1.	43	*the nature of the*
	2.	31	*a wide range of*
	3.	24	*the presence of a*
	4.	22	*a large number of*
	5.	18	*the validity of the*
	6.	13	*the purpose of this*
2	Noun phrase with other post-modifier fragments		
	7.	24	*the other hand the*
	8.	15	*the same way as*
	9.	14	*the difference between the*
	10.	13	*an increase in the*
	11.	11	*such a way that*
Total	11	228	
(2) PP-based			
1	Prepositional phrase with embedded *of*-phrase fragment		
	1.	57	*in the context of*
	2.	46	*in the form of*
	3.	41	*in the absence of*
	4.	32	*at the time of*
	5.	31	*on the part of*
	6.	27	*in the course of*
	7.	26	*as part of the*
	8.	24	*for the purposes of*
	9.	19	*in a number of*
	10.	15	*in terms of a*
	11.	13	*over a period of*
2	Other prepositional phrase (fragment)		
	12.	119	*on the other hand*
	13.	37	*with respect to the*
	14.	27	*in this case the*
	15.	23	*to the fact that*
	16.	15	*in the next section*
	17.	14	*in this paper we*
	18.	12	*in accordance with the*
	19.	11	*in the present study*
Total	30	589	
(3) VP-based			

1	Anticipatory *it* + verb phrase/adjective phrase		
	1.	48	*it is possible to*
	2.	43	*it is important to*
	3.	33	*it is clear that*
	4.	26	*it is necessary to*
	5.	19	*it is possible that*
	6.	18	*it is likely that*
	7.	18	*it is not possible*
	8.	17	*it should be noted*
	9.	15	*it is interesting to*
	10.	12	*it is impossible to*
	11.	11	*it is obvious that*
2	Passive verb + prepositional phrase fragment		
	12.	22	*is based on the*
	13.	18	*can be found in*
	14.	17	*be used as a*
	15.	17	*can be seen in*
	16.	12	*be related to the*
3	(Verb phrase +) *that*-clause fragment		
	17.	19	*is that it is*
4	*That*-clause fragment		
	18.	32	*that there is no*
5	(Verb/adjective +) *to*-clause fragment Predicative adjective + *to*-clause		
	19.	45	*can be used to*
	20.	38	*is likely to be*
	21.	17	*to ensure that the*
	22.	15	*is not possible to*
6	Pronoun/noun phrase + *be* (+…)		
	23.	14	*there are a number*
7	Adverbial clause fragment		
	24.	15	*as can be seen*
Total	24	541	
(4)	Other expressions		
	1.	18	*whether or not the*
	2.	13	*due to the fact*
	3.	11	*take into account the*
Total	3	42	
Total	57	1400	

Appendix G
Structural distribution of the four-word bundles in the written AFL sub-list

Structure	Types	%
NP-BASED		
1-Noun phrase with *of*-phrase fragment	6	10%
2-Noun phrase with other post-modifier fragments	5	9%
PP-BASED		
1-Prepositional phrase with embedded *of*-phrase fragment	11	19%
2-Other prepositional phrase (fragment)	8	14%
VP-BASED		
1-Anticipatory *it* + verb phrase/ adjective phrase	11	19%
2-Passive verb + prepositional phrase fragment	5	9%
3-(Verb phrase +) *that*-clause fragment	1	2%
4-*That*-clause fragment	1	2%
5-(Verb/adjective +) *to*-clause fragment		
Predicative adjective + *to*-clause	4	7%
6-Pronoun/noun phrase + *be* (+…)	1	2%
7-Adverbial clause fragment by the subordinator	1	2%
OTHER EXPRESSIONS	3	5%
Total	57	100%

Appendix H
Information sheet for pre-sessional course leader/director

Research project title: Profiling lexical bundles in EAP textbooks and instructors' materials on an EAP pre-sessional course

1. Self-introduction

My name is Reem Fattani and I am an international postgraduate student from the School of English at the University of Sheffield.

2. The topic of the research

The purpose of the research is to investigate *lexical bundles* (Biber *et al.*, 1999), a type of formulaic sequence, in EAP materials aimed at teaching academic writing. Lexical bundles are sequenced multi-word combinations that are formed from three, four, five, and six words, which are used as fixed phrases. For example, the lexical bundle *as well as* is made up of a three-word expression, while *on the other hand* is a four-word bundle.

The aim of this research is to investigate lexical bundles found in EAP materials such as textbooks, instructors' handouts, and required reading texts that are intended for teaching academic writing in the course. The objective of the study is to profile the most frequent four-word bundles that are found in the EAP materials, and to compare the bundles found to those presented in a ready-made list called the Academic Formulas List (AFL) (Simpson-Vlach and Ellis, 2010). The AFL list is a strong and useful work on lexical bundles that provides the most commonly used bundles in academic discourse.

3. Methodology

This study will use the following methods to achieve the aims and objectives of this research:

1- Corpora compiling: I will create three corpora compiled from EAP materials. The corpora are as follows: a textbooks corpus and an instructors' handouts and required readings corpus. The aim is to profile the most frequent four-word bundles found in these different genres by using a specialized computer software program called AntConc.

2- Generating the lexical bundles: from each corpus, I will generate a list of the most frequent four-word bundles. This means I will have a *textbook(s) bundles list* of the most frequent four-word bundles and *instructors' materials bundles list* with *required reading* of the most frequent four-word bundles. These lists will be generated to help to conduct the comparison with the AFL.

3- Data analysis: I will perform a functional and structural analysis on the most frequent four-word bundles identified in the three lists and compare these bundles to bundles from the AFL list. The framework used by Biber *et al*. (2004) to classify bundles according to their meaning and function in texts is referred to as a functional classification. For example, writers use the bundle *on the other hand* when they want to organize their text by indicating contrast and comparison. In addition, according to Biber *et al*. (1999), structural classification presents the structure of bundles, which are not complete structural units. Instead, they are seen as fragmented phrases. For example, the bundle *on the other hand* is structurally classified as a prepositional phrase. This type of analysis will help the researcher see if there are significant similarities or differences in terms of frequency, function and structure among the three lists and in comparison to the AFL list.

References

Anthony, L. (2018b) AntConc (Version 3.5.6) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

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

Biber, D. *et al*. (2004) If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371–405.

Simpson-Vlach, R. and Ellis, N. C. (2010) An Academic Formulas List: New methods in phraseology research. *Applied Linguistics*, 31(4), 487–512.

4. The participants

The researcher needs to collect some information and EAP materials from the director and instructors of the course. First, I will ask them the following questions: To compile the three corpora for the present study, I need to know:

1- How many classes are offered in the course?

2- How many instructors teach these classes?

3- What are the textbook(s), handout(s), and required readings that are presented to learners during the teaching of academic writing?

4- Do the textbooks in the course have an integrated, multi-skills syllabus or not?

5- Do all instructor(s) use the same textbook(s), handouts, and required readings or not?

6- What other materials do instructors use (e.g. online texts, recommended book(s), etc.) in the teaching of academic writing?

Second, based on the answers, the material will be gathered from my participants. The material will include: textbook(s), instructors' handouts, and required reading texts that are used in the course and which are presented to students during the classes focusing on teaching academic writing.

The decision they make with regard to their participation in this research is entirely up to them. If instructors decide to take part in this research, they will sign a consent form and will be provided with an information form to keep. Even after giving consent to participate in the research, they can still withdraw from the project at any time without giving any reasons or further explanations.

5. Duration of the project

The duration of the PhD project is not related to the duration of the project within the centre. This means that once the leader and instructor(s) have provided all copies of the materials requested for the present study, they will not be subject to further participation in the study.

6. The manner in which the data will be used

The project takes a corpus-based approach, where the researcher will compile corpora to analyze lexical bundles derived from the EAP materials received from the course. The data received will be stored in electronic files and processed using special computer software for lexical handling, related to corpora compiling and data analysis.

7. Data after the study

After this project has been completed, all papers and data will be destroyed or returned to the course if applicable. Instructors' handouts will also be destroyed.

8. The benefits to arise from this research

It is hoped that by the end of this project, we will have a clear understanding of the most frequent four-word lexical bundles found in EAP materials and across different genres, in terms of teaching academic writing. This research will also shed some light on whether EAP materials are employing similar or different lexical bundles to those bundles provided by empirically derived lists such as the AFL list, which is targeted specifically at English for Academic Purposes.

9. Confidentiality

All the information, materials, handouts, and data that the researcher collects will be kept strictly confidential. Instructors' and directors' names will not be identified in the research or in any publications arising from the research.

10. The project has been ethically reviewed

This project has been ethically reviewed and approved by the School of English at the Faculty of Arts and Humanities, University of Sheffield.

11. My contact details for further information:

Name:
Email:
Phone

Instructors and the course leader will be given a copy of an information sheet and they will sign a consent form to keep.

Appendix I

Information sheet for pre-sessional course instructors

Research project title: Profiling lexical bundles in EAP textbooks and instructors' materials on an EAP pre-sessional course

Dear course instructor,
You are invited to take part in this research. Before you make your final decision about whether to participate, it is important for you to clearly understand the reasons behind this research project. Please take your time to read the following information carefully. You may ask any question if anything is not clear or if you would like more information. Take time to think and choose whether or not you wish to be a part of this research. Thank you for your time and for reading this information.

1. The purpose of the project

The purpose of the research is to investigate *lexical bundles* (Biber *et al*., 1999), a type of formulaic sequence, in EAP materials aimed at teaching academic writing. Lexical bundles are sequenced multi-word combinations that are formed from three, four, five, and six words, which are used as fixed phrases. For example, the lexical bundle *as well as* is made up of a three-word expression, while *on the other hand* is a four-word bundle.

The aim of this research is to investigate the lexical bundles found in EAP materials such as textbooks, instructors' handouts and required reading texts that are intended to teach academic writing in the pre-sessional course. The objective of the study is to profile the most frequent four-word bundles that are found in EAP materials, and to compare the bundles found to those presented in a ready-made list called the Academic Formulas List (AFL) (Simpson-Vlach and Ellis, 2010). The AFL list is a strong and useful work on lexical bundles that provides the most commonly used bundles found in academic discourse.

Methodology: this study will use the following methods to achieve the aims and objectives of this research:

1- Corpora compiling: I will create three corpora compiled from EAP materials. The corpora are as follows: a textbooks corpus and an instructors' handouts corpus, including required readings. The aim is to profile the most frequent four-word bundles found in these different genres by using a specialized computer software program called AntConc.
2- Generating the lexical bundles: from each corpus, I will generate a list of the most frequent four-word bundles. This means I will have a *textbooks bundles list* of the most frequent four-word bundles and an *instructors' materials bundles list* of the most frequent four-word bundles, including *required reading* of the most frequent four-word bundles. These lists are generated to help to conduct the comparison.

3- Data analysis: I will perform a functional and structural analysis on the most frequent four-word bundles identified in the three lists and will compare these bundles to bundles from the AFL list. The framework used by Biber *et al.* (2004) to classify bundles according to their meaning and function in texts is referred to as a functional classification. For example, writers use the bundle *on the other hand* when they want to organize their text by indicating contrast and comparison. In addition, according to Biber *et al.* (1999), structural classification presents the structure of the bundles, which are not complete structural units. Instead, they are seen as fragmented phrases. For example, the bundle *on the other hand* is structurally classified as a prepositional phrase. This type of analysis will help the researcher to see if there are significant similarities or differences in terms of frequency, function and structure among the three lists and in comparison to the AFL list.

2. The participants

The researcher needs to collect some information (questionnaires) and EAP materials from the instructors of the course. First, I will ask you the following questions:
To compile the data for the present study, I need to know:

1- How many classes are offered in the course?
2- How many instructors teach these classes?
3- What are the textbook(s), handout(s), and required readings that are presented to learners during the teaching of academic writing?
4- Do the textbooks in the course have an integrated, multi-skills syllabus or not?
5- Do all instructor(s) use the same textbook(s), handouts, and required readings or not?
6- What other materials do instructors use (e.g. online texts, recommended book(s), etc.) in the teaching of academic writing?

Second, based on your answers, the material will be gathered from you. The material will include: textbook(s), instructors' handouts, and required reading texts that are used in the course and which are presented to students during the classes focusing on teaching academic writing. The researcher also needs to administer a small questionnaire to collect some information from the instructors of the EAP course.

3. Your choice is respected

The decision you make with regard to your participation in this research is entirely up to you. If you decide to take part in this research, you will sign a consent form and you will be provided with a copy of this information form to keep. Even after giving consent to participate in the research, you can still withdraw from the project at any time without giving any reasons or further explanations.

4. Your acceptance to take part means:

You should know that this research is going to be conducted during the course. Your responsibilities will be to provide a copy/copies of textbook(s), copies of handouts, and copies of required readings in any form, either via email or personally. Once all data requested for the present study are submitted, your involvement is completed.

5. The benefits to arise from this research

It is hoped that by the end of this project we will have a clear understanding of the most frequent four-word lexical bundles found in EAP materials and across different genres. In addition, the comparison with a useful AFL list may shed some light on the similarities or differences between this list and the lexical bundles that are provided to students attending EAP courses. The AFL list will be particularly relevant for courses teaching academic writing.

6. Confidentiality

All the information, material, handouts and data that the researcher collects will be kept strictly confidential. Instructors' or directors' names will not be identified in the research or in any publications arising from the research.

7. The project has been ethically reviewed

This project has been ethically reviewed and approved by the School of English in the Faculty of Arts and Humanities at the University of Sheffield.

My contact details for further information:

Name:
Email:
Phone

You will be given a copy of this information sheet and you will sign a consent form to keep.

University of Sheffield

Appendix J

Participant consent form

Title of Research Project: Profiling lexical bundles in EAP textbooks and instructors' materials on an EAP pre-sessional course

Name of Researcher:

Participant Identification Number for this project: Please initial box

1. I confirm that I have read and understand the information sheet/letter
   explaining the above research project and I have had the opportunity to ask questions about the project. □

2. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason and without there being any negative consequences. In addition, should I not wish to answer any particular question or questions, I am free to decline. □

3. I understand that my responses will be kept strictly confidential. I give permission for members of the research team to have access to my anonymized responses. I understand that my name will not be linked with the research materials, and I will not be identified or identifiable in the report or reports that result from the research. □

4. I agree to take part in the above research project. □

________________________ ________________
____

Name of Participant Date Signature
(*or legal representative*)

_________________________ ________________
____

Name of person taking consent Date Signature
(*if different from lead researcher*)
To be signed and dated in the presence of the participant

_________________________ ________________
____

Lead Researcher Date Signature
To be signed and dated in the presence of the participant

Copies:
Once this has been signed by all parties the participant should receive a copy of the signed and dated participant consent form, the letter/pre-written script/information sheet and any other written information provided to the participants. A copy of the

signed and dated consent form should be placed in the project's main record (e.g. a site file), which must be kept in a secure location.

Adapted from the consent form provided by the University of Sheffield.

Appendix K

Instructors' materials (Samples)

No.	Instructors	Word count	Name of text	Type of text
	Instructor (1)			
1.		818 w	'A brief history of Sheffield'	Reading passages
2.		536 w	'Navigating our way through computer files uses the same brain structures as a dog finding its bone' http://www.sheffield.ac.uk/news/nr/navigating-way-through-computer-like-dog-with-bone-1.515267	Reading passages
3.		270 w	Academic Reading and Writing • Extending writing project essay –draft 1 writing • Getting started with writing • Generating and listing ideas – brainstorming • Selecting and grouping ideas into a logical plan • Writing an outline	Instructor's handout
4.		939 w	Academic Reading and Writing • Cultural considerations about writing • Paragraph coherence and cohesion • Using linking words in academic writing • Reading for extended writing	Instructor's handout
5.		1527 w	Academic Extended Writing • Grammar for paraphrase in extended writing • Purposes and results in order **to**, so as **to**, in order	Instructor's handout, including grammar parts

			that, so that, in such a way that • Passive modal verbs, infinitives and gerunds • Reduced relative clauses • Some conventions of written academic English • Extended writing criteria	
6.		444 w	Academic Extended Writing • 'Thermal sea water desalination based on self-heat recuperation' • Reading and note taking • 'Ecotourism: A panacea or a predicament' • Reading and exercises • Features of academic style • Some conventions of written academic English • Homework: start writing the first draft of your extended writing assignment in the coming week.	Instructor's handout, including writing conventions parts
7.		397 w	Academic Reading and Writing • Tutorials + writing sample • Tutorials + vocabulary self-study • Study skills for extended writing • Welfare lecture review • *Cambridge Academic English*	Instructor's handout
8.		356	Academic Reading and Writing • Course book introduction:	Instructor's handout

			Cambridge Academic English Intermediate • Academic orientation: reading and writing in academic English	
9.		457	Academic Reading and Writing • Presentations: Week 12 info • Lecture quiz review: ‘Anti-social children: Causes and consequences’ • EAP skills writing referencing task • Reading: *Cambridge Academic English: Indications and trends*	Instructor’s handout, including quiz, and writing conventions parts
10.		1056 w	Academic Reading and Writing • Reading comprehension on Google Classroom • *Cambridge Academic English Intermediate* • Review: Academic orientation pair work • Vocabulary: How do you learn? • Reading for key terms and guessing meaning in context • Grammar –ing forms present simple • Reading for your course, basic learning style	Instructor’s handout, including grammar parts
11.		634 w	Academic Reading and Writing • Vocabulary tech lesson with Nick • Academic reading circles • Article topic: ‘Using social media in order to	Instructor’s handout

			learn English' - 'Learning the Queen's English… on your mobile phone?'	
12.		415	Academic Reading and Writing - *Cambridge Academic English Intermediate*: Grammar review - Academic reading circle article on plagiarism - 'Carbon copies' by Peter Wilby in *The Guardian* Higher Education supplement - Summariser - Academic Word Person - Connector - Discussion	Instructor's handout
13.		564 w	Academic Reading and Writing - *Cambridge Academic English Intermediate*: Problems in the natural world - Reading: 'Understanding essay questions' - Extended writing project question analysis	Instructor's handout
14.		738 w	Academic Reading and Writing - Class on text analysis for vocabulary - *Cambridge Academic English Intermediate*: Problems in the natural world - Writing: Paragraph building - Homework: Upload texts from the extended writing book on eco-tourism or	Instructor's handout, including writing conventions parts

			desalinisation into Oxford Textchecker and select some words to put on the webcorp site	
15.		216 w	Academic Reading and Writing • Course information student handbook quiz • 'Getting to know you' activities EAP questionnaire • Google Classroom introduction • Reading 'A brief history of Sheffield' • Homework • Needs analysis questionnaire	Instructor's handout
16.		1484 w	• Dr Jeannette Littlemore – Department of English – University of Birmingham • Publications • Metaphor and the non-native speaker (Littlemore *et al.*, 2011)	Instructor's handout, including writing conventions parts
17.		489 w	Academic Reading and Writing • *Cambridge Academic English Intermediate*: Grammar and vocabulary review • Reading circle: 'Navigating our way through computer file uses the same brain structures as a dog finding its bone?' • Lecture quiz Antisocial children: Causes and consequences	Instructor's handout, including grammar parts, and writing conventions parts
18.		475 w	From: Paterson, K. (2013) *Oxford grammar for EAP*. Oxford: Oxford University	Grammar handouts

			Press (with Roberta Wedge) • Passive modal verbs, infinitives and gerunds	
19.		870 w	‘What is materials science?’ http://www.strangematterexhibit.com/whatis.html	Reading passage
20.		756 w	‘Cementation furnace steel making in Sheffield’ • Grammar and writing • Past simple passive in processes • ‘Steel manufacture using a cementation furnace’ • Plaque text (adapted by the instructor) • Reduced relative clauses	Reading passage + grammar handout
21.		1124 w	*Hewings Advanced Grammar*, Cambridge University Press, pp. 162–163 • Purposes and results: **in order to, so as to**, etc.	Grammar handouts
22.		1370 w	‘Carbon copies’	Reading passage
23.		819 w	‘Learning the Queen’s English … on your mobile phone?’	Reading passage
24.		366 w	Writing Modal Answer	Instructor’s handout
25.		293 w	Academic Reading and Writing • Punctuation in academic writing (IELTS writing task)	Writing conventions handout
26.		198 w	Book Quiz • *Cambridge Academic English Intermediate*	Instructor’s handout
27.		108 w	Book Quiz • *Cambridge Academic English Intermediate*	Instructor’s handout
28.		204 w	Academic Reading and	Instructor’s

			Writing • Tech training presentation techniques • Grammar and writing How steel was made in a cementation furnace Past simple passive voice/Reduced relative clauses/Infinitive of purposes • Vocabulary practice	handout, including grammar parts
29.		356 w	Academic Reading and Writing • *Cambridge Academic English*: Academic English • Reading and vocabulary • Book quiz • Vocabulary York/National Railway Museum/crossword	Instructor’s handout, including book quiz
Instructor (2)				
1.		957 w	Paraphrasing Review	Writing conventions handout
2.		111 w	Academic Reading: Discuss the following in groups: • What type of texts (academic and general) do you read in English? • How much time do you spend each week reading English? • What is a paragraph?	Instructor’s handout
3.		104 w	Reading Skills for Academic Study • What sort of things do you read in your own language? • What do you read in English?	Instructor’s handout

			http://www.uefap.com/reading/exercise/types/into.htm	
4.		528 w	Academic Reading • Choosing what to read • Reading with purpose • How to approach an academic text	Instructor's handout
5.		3568 w	'Adapting to climate change in small island developing states'	Article
6.		709 w	Referencing: an introduction • Including other people's writing in your work • Direct quotation	Writing conventions handout
7.		2577 w	Academic Reading Circles • 'Plagiarism in Japanese universities: Truly a cultural matter?'	Reading article
8.		671 w	Academic Writing • Quotations and referencing	Writing conventions handout
Instructor (3)				
1.		629 w	Academic Reading Circles: Text summariser • Read the article and highlight/underline the main ideas Connector • Read the article and make connections between the claims/arguments/findings in the article and your own experience/knowledge of the subject or real-life events connected to the subject Academic Word Person • Read the article and look for words, phrases, and/or collocations that are new or difficult to understand, or that are important to the	Instructor's handout

			understanding of the text Text Analyst • Read the article and find two or three important, interesting or difficult passages	
2.		11833 w	‘An analysis of friendship networks, social connectedness, homesickness, and satisfaction levels of international students’	Article
3.		5986 w	‘Does higher education promote independent learning?’	Article
4.		8019 w	‘Plagiarism in Japanese universities: Truly a cultural matter?’	Article
5.		11387 w	‘*“Nowhere has anyone attempted… In this article I am to do just that”* A corpus-based study of self-promotional *I* and we in academic writing across four disciplines’	Article
6.		8073 w	‘Why EAP is necessary: A survey of Hong Kong tertiary students’	Article
Instructor (4)				
1.		12654 w	• Normalization from verbs • Talking about cause and effect • Short paragraphs: Effects of colour, Gold, English, Dolphins, A Mardi Gras Custom, Olympic Athletes, Genetic Engineering, More Over, DVD. Here Comes BD!, Drugs and Olympic Games 1 • Connectors • Transition signals, transition words and phrases and conjunctive adverbs • Using outside	Instructor’s handouts, including grammar parts, and writing conventions parts (comprised in one document)

			sources: Plagiarism Citing sources Quotations Reporting verbs and phrases Sequence of tenses rules • Writing practice Language study (using keywords correctly) Reading quickly to find information Noticing collocations	
Instructor (5)				
1.		611 w	Presentation Questions Topic: Risk and hazards • 'Consumers' and physicians' perceptions about high tech wearable health products'	Instructor's handouts

References

Anonymous (n.d.) *Navigating our way through computer files uses the same brain structures as a dog finding its bone*. Available at: https://www.sheffield.ac.uk/news/nr/navigating-way-through-computer-like-dog-with (Accessed: 2 October 2018).

Bailey, S. (2006) *Academic writing : A handbook for international students*, 2nd edn. Abingdon: Routledge.

Barraclough, K. (1976) Post-medieval archaeology. *The Development of the Cementation Process for the Manufacture of Steel*, 10(1), 65–88 .

Betzold, C. (2015) Adapting to climate change in small island developing states. *Climatic Change*, 133(3), 481–489.

Evans, S. and Green, C. (2007) Why EAP is necessary: A survey of Hong Kong tertiary students. *Journal of English for Academic Purposes*, 6(1), 3–17.

Gillett, A. *et al.* (2009) *Inside track to successful academic writing*. Harlow: Pearson.

Glendinning, E. and Holmström, B. (2004) *Study reading: A course in reading skills for academic purposes*, 2nd edn., Cambridge : Cambridge University Press.

Gow, L. and Kember, D. (1990) Does higher education promote independent learning? *Higher Education*, 19(3), 307–322.

Harwood, N. (2005) '*Nowhere has anyone attempted ... In this article I aim to do just that'*: A corpus-based study of self-promotional *I* and *we* in academic writing across four disciplines. *Journal of Pragmatics*, 37(8), 1207–1231 .

Hendrickson, B. *et al.* (2011) An analysis of friendship networks, social connectedness, homesickness, and satisfaction levels of international students. *International Journal of Intercultural Relations*, 35(3), 281–295.

Hewings, M. (2005) *Advanced grammar in use*, 2nd edn. Cambridge : Cambridge University Press.

Jorden, R. R. (1999) *Academic writing course*, 2nd edn. Harlow: Longman.

Lambert, T. (n.d.) *A brief history of Sheffield, England.* Available at: http://www.localhistories.org/sheffield.html (Accessed: 13 October 2018).

Littlemore, J. *et al.* (2011) Difficulties in metaphor comprehension faced by international students whose first language is not English. *Applied Linguistics*, 32(4), 408–429.

Mizuno, H. *et al.* (2013) Thermal sea water desalination based on self-heat recuperation. *Clean Technologies and Environmental Policy*, 15(5), 765–769.

Oxford University Press (n.d.) *English File.* Available at: https://elt.oup.com/student/englishfile/?cc=gb&selLanguage=en (Accessed: 2 October 2018).

Paterson, K. (2013) *Oxford grammar for EAP: English grammar and practice for academic purposes with answers*. Oxford: Oxford University Press.

Paterson, K. and Wedge, R. (2013) *Oxford grammar for EAP: English grammar and practice for academic purposes with answers*. Oxford: Oxford University Press.

Suphan, N. and Yigit, Y (2015) Consumers' and physicians' perceptions about high tech wearable health products' *Procedia - Social and Behavioral Sciences*, 195(3), 1261–1267.

Trucano, M. (2010) *Learning the Queen's English ... on your mobile phone?* Available at: https://blogs.worldbank.org/edutech/learning-the-queens-english-on-your-mobile-phone (Accessed: 2 October 2018).

Wheeler, G. (2009) Plagiarism in the Japanese universities: Truly a cultural matter? *Journal of Second Language Writing*, 18(1), 17–29.

Wilby, P. (1998) Carbon copies: Sunderland academic alleges plagiarism; review body puts it down to regrettable oversight. Whatever happened? *The Guardian*, 5 May, p.iv.

Appendix L

A semi-structured questionnaire

I. Background information:

1. Name:
2. Gender:
3. Nationality:
4. Native language:
5. Qualifications:

II. Work experience:

6. For how many months/years have you been an English language teacher?

III. Experience in teaching writing:

7. For how long have you been teaching writing?

Institution:	Name of writing module/level	Number of years teaching this type of module

8. What workshops, conferences, or seminars have you attended on teaching second language writing, if any?

IV. Lexical bundle questions:

What are lexical bundles?

- They are considered one type of formulaic sequence.
- They are sequenced multi-word combinations that are formed from three, four, five, and six words, and are used as fixed phrases.
- For example, the lexical bundle *as well as* is made up of a three-word expression, while *on the other hand* is a four-word bundle.

Relative to the explanation provided above, the following set of questions is intended to elicit information on your views and usage of lexical bundles, connected to the kind of materials you use during writing classes. Please read and answer.

9. How often do you pay attention to lexical bundles when deciding what materials, handouts, and reading passages to use in writing classes?
    - A. I always pay attention to lexical bundles.
    - B. I sometimes pay attention to lexical bundles.
    - C. I irregularly pay attention to lexical bundles.
    - D. I rarely pay attention to lexical bundles.
    - E. I never pay attention to lexical bundles.

If your answer to question **9** is **A** or **B**, go to and answer questions **10** and **12**.
If your answer to question **9** is **C**, **D** or **E**, go to and answer questions **11** and **12**.

10. In your opinion, which of the following reason(s) made you pay attention to the presence of lexical bundles in the materials, handouts, and reading passages that you use during writing classes? (You can choose MORE THAN ONE answer.)
    - A. Lexical bundles are a focus of the course syllabus.
    - B. I was aware of lexical bundles from previous syllabuses.
    - C. I was aware of lexical bundles from my own reading.
    - D. Lexical bundles are a useful feature in academic writing.
    - E. Other (Please specify).

11. In your opinion, what are the reason(s) behind not focusing on the presence of lexical bundles in the materials, handouts, and reading passages that you use during writing classes? (You can choose MORE THAN ONE answer.)
    - A. I do not know what lexical bundles are.
    - B. Lexical bundles are not a focus of the course syllabus.
    - C. Lack of time.
    - D. I do not think lexical bundles are a relevant feature in the materials that I use.
    - E. I only focus on lexical bundles during teaching writing but not in my handouts.
    - F. Other (please specify)

12. Regarding the handouts or materials used in writing classes, I have …
    (You can choose MORE THAN ONE answer if applicable to you.)
    1. used a ready-made vocabulary list such as the Academic Word List (AWL).
    2. used a formulas list such as the Academic Formulas List (AFL).
    3. used other lists such as those found at the end of a chapter or

textbook.

4. used other lists that may be provided by English centres.
5. never used any vocabulary lists.
6. never used any formulas.

Appendix M

Instructors' materials weeded-out bundles

No.	Frequency	Range	Bundles
1.	6	2	*grammar for eap oxford*
2.	6	2	*k oxford grammar for*
3.	6	3	*native speakers of English*
4.	6	2	*oxford grammar for eap*
5.	6	2	*paterson k oxford grammar*
6.	6	2	*to in order to*
7.	5	2	*cambridge cambridge university*
8.	5	2	*hong kong journal of*
9.	4	2	*a m roy eds*
10.	4	2	*additionally of the American*
11.	4	2	*albany state university of*
12.	4	2	*buranen a m roy*
13.	4	2	*cambridge university press cambridge*
14.	4	2	*eds perspectives on plagiarism*
15.	4	2	*in l buranen a*
16.	4	2	*l buranen a m*
17.	4	2	*m roy eds perspectives*
18.	4	3	*non native speakers of*
19.	4	2	*only of the Japanese*
20.	4	2	*opposed to only of*
21.	4	2	*roy eds perspectives on*
22.	4	2	*to only of the*
23.	4	3	*www sciencedirect com science*
24.	3	2	*and non native speakers*
25.	3	2	*com science article pii*
26.	3	2	*first year students at*
27.	3	2	*http www sciencedirect com*
28.	3	2	*if you don t*
29.	3	2	*in academic writing in*
30.	3	2	*j the effect of*
31.	3	2	*kong journal of second*
32.	3	2	*language writing journal of*
33.	3	2	*of academic writing and*
34.	3	2	*of first year students*
35.	3	2	*of the fact that*
36.	3	2	*of the world the*
37.	3	2	*science article pii s*
38.	3	2	*sciencedirect com science article*
39.	3	2	*second language writing journal*
40.	3	2	*since the mid s*
41.	3	3	*someone else s words*
42.	3	2	*the stock market crash*
43.	3	2	*writing journal of second*
Total			43 bundles

Appendix N

Instructors' materials excluded and overlapping bundles

Instructors' excluded bundles: context-dependent/based			
Context-dependent/based bundles: a sequence containing context-related word(s), usually incorporating proper nouns or they are sequences that appeared in the same texts provided by two or three instructors on a particular topic.			
No.	Frequency	Range	Context-based bundles
1.	25	2	*english for academic purposes*
2.	21	2	*journal of second language*
3.	20	2	*of second language writing*
4.	9	2	*english for specific purposes*
5.	7	2	*students whose first language*
6.	6	2	*and intellectual property in*
7.	6	2	*of the american students*
8.	6	2	*of the japanese students*
9.	6	2	*plagiarism and intellectual property*
10.	6	2	*students in hong kong*
11.	6	2	*the united states and*
12.	6	2	*whose first language is*
13.	5	2	*japan and the united*
14.	5	2	*students were asked to*
15.	4	2	*a student can commit*
16.	4	2	*considered a moral transgression*
17.	4	2	*first language is not*
18.	4	2	*in a japanese university*
19.	4	2	*in a postmodern world*
20.	4	2	*in japanese higher education*
21.	4	2	*in the developing world*
22.	4	2	*in the middle east*
23.	4	2	*in the two languages*
24.	4	3	*in their native language*
25.	4	2	*intellectual property in a*
26.	4	2	*language is not English*
27.	4	2	*of new york press*
28.	4	2	*on plagiarism and intellectual*
29.	4	2	*perspectives on plagiarism and*
30.	4	2	*property in a postmodern*
31.	4	2	*state university of new*
32.	4	2	*students at hokkaido university*
33.	4	2	*suggest that japanese students*
34.	4	2	*the anatomy of dependence*
35.	4	2	*the japanese students who*
36.	4	2	*united states and japan*
37.	4	2	*university of new York*
38.	3	2	*by the time they*
39.	3	2	*his or her academic*
40.	3	2	*humanities and social sciences*
41.	3	2	*or her academic career*
42.	3	2	*plagiarism is not considered*

43.	3	2	*research in higher education*
44.	3	2	*students are more likely*
45.	3	2	*students in the united*
46.	3	2	*students to express their*
47.	3	2	*studies in higher education*
48.	3	2	*university of michigan press*
Total			48 bundles

Instructors’ overlapping bundles

Overlapping bundles: are part of longer lexical bundles and because of automatic retrieval, longer lexical bundles are broken down into two or three shorter ones.

No.	Frequency	Range	Overlapping bundles
1.	5	2	*in japan and the*
2.	5	2	*is quite possible that*
3.	4	2	*are a number of*
4.	4	2	*not considered a moral*
5.	3	2	*about a third of*
6.	3	2	*be used in the*
7.	3	2	*due to the fact*
8.	3	2	*is the best way*
9.	3	2	*not considered a major*
10.	3	2	*passive form of the*
11.	3	2	*the development of the*
12.	3	3	*the time of his*
13.	3	2	*the time they enter*
14.	3	2	*time they enter university*
Total			14 bundles

Appendix O

Textbooks weeded-out bundles

No.	Frequency	Range	Bundles
1.	22	3	*a look at the*
2.	19	3	*and vocabulary grammar and*
3.	19	3	*grammar and vocabulary grammar*
4.	19	3	*vocabulary grammar and vocabulary*
5.	13	3	*in pairs compare your*
6.	13	4	*of the world s*
7.	13	3	*the text in a*
8.	11	3	*cambridge cambridge university press*
9.	11	3	*n c u a*ഉൗ
10.	11	3	*n c u the*
11.	11	3	*the writer s position*
12.	10	3	*for example in the*
13.	10	3	*study tip when you*
14.	10	4	*the writer s opinion*
15.	9	2	*a read the following*
16.	9	2	*a student s essay*
17.	9	2	*a work in pairs*
18.	9	3	*n c something that*
19.	9	4	*that there is a*
20.	9	2	*with the title discuss*
Total			20 bundles

Appendix P

Textbooks excluded bundles

Overlapping Bundles			
Overlapping bundles: they are part of longer lexical bundles and because of automatic retrieval, longer lexical bundle is fragmented into two or three shorter ones.			
No.	Frequency	Range	*Overlapping Bundles*
1.	28	2	*words in the box*
2.	21	2	*using the words in*
3.	14	2	*you have been asked*
4.	13	3	*a dictionary to help*
5.	13	3	*dictionary to help you*
6.	13	3	*in the written academic*
7.	13	3	*paragraph of the text*
8.	13	3	*shows that in the*
9.	13	2	*the following sentences using*
10.	13	3	*the written academic corpus*
11.	12	2	*are going to read*
12.	12	3	*given an essay with*
13.	11	3	*been given an essay*
14.	11	3	*have been given an*
15.	11	2	*the meaning of words*
16.	10	2	*the questions about the*
17.	10	2	*words in bold in*
18.	9	2	*a dictionary to check*
19.	9	3	*advantages and disadvantages of*
20.	9	3	*are common in academic*
21.	9	4	*is the difference between*
22.	9	2	*may be more than*
23.	9	2	*read the following extract*
Total			23 bundles

Appendix Q

Instructors' materials bundles list (in-context information)

Total No. of N-Gram Types: 79				In-context information		
Rank	Frequency	Range	Bundles	Instructional parts	Reading parts	Both parts
1.	18	3	*in the United States*		*	
2.	15	3	*at the end of*			*
3.	15	3	*on the other hand*		*	
4.	12	4	*the end of the*			*
5.	10	4	*it is important to*			*
6.	9	2	*the extent to which*		*	
7.	8	4	*as a result of*		*	
8.	8	2	*in the field of*			*
9.	7	2	*with the help of*		*	
10.	6	2	*and the United States*		*	
11.	6	2	*at the beginning of*			*
12.	6	2	*in the middle of*			*
13.	6	2	*title of the article*	*		
14.	5	3	*a great deal of*			*
15.	5	2	*are more likely to*		*	
16.	5	2	*form of the verb*	*		
17.	5	2	*is not considered a*		*	
18.	5	2	*it is possible that*		*	
19.	5	2	*it is quite possible*		*	
20.	5	2	*the beginning of the*			*
21.	5	2	*the best way to*			*
22.	5	2	*the results of the*		*	
23.	5	2	*to know each other*		*	
24.	4	3	*as one of the*		*	

25.	4	2	*as opposed to only*	*	
26.	4	3	*at the time of*	*	
27.	4	2	*at the university of*	*	
28.	4	3	*can be used to*		*
29.	4	4	*for a long time*	*	
30.	4	2	*from the department of*	*	
31.	4	2	*have shown that the*	*	
32.	4	2	*in the development of*	*	
33.	4	2	*is important to note*	*	
34.	4	2	*it is interesting that*	*	
35.	4	2	*the fact that the*	*	
36.	4	2	*the purpose of this*	*	
37.	4	3	*the relationship between the*		*
38.	4	2	*there are a number*		*
39.	4	2	*to be one of*	*	
40.	4	2	*to the fact that*	*	
41.	3	2	*a third of the*	*	
42.	3	3	*all over the world*	*	
43.	3	3	*as a consequence of*	*	
44.	3	2	*as part of the*	*	
45.	3	3	*be aware of the*	*	
46.	3	2	*be done by a*	*	
47.	3	2	*by no means the*	*	
48.	3	2	*can be used in*		*
49.	3	2	*considered a major issue*	*	
50.	3	3	*few and far between*	*	
51.	3	2	*he was unable to*	*	
52.	3	3	*in countries*	*	

			such as			
53.	3	2	*in order to avoid*		*	
54.	3	2	*in order to learn*	*		
55.	3	2	*in order to test*		*	
56.	3	2	*in relation to the*		*	
57.	3	2	*in the case of*		*	
58.	3	2	*in which it is*		*	
59.	3	2	*is more of a*		*	
60.	3	3	*it is difficult to*		*	
61.	3	3	*it is possible to*			*
62.	3	2	*lack of understanding of*		*	
63.	3	2	*may or may not*			*
64.	3	2	*of different types of*		*	
65.	3	3	*of his or her*			*
66.	3	2	*on a regular basis*		*	
67.	3	2	*on the one hand*		*	
68.	3	2	*part of a larger*		*	
69.	3	2	*purposes of this study*		*	
70.	3	3	*that he or she*		*	
71.	3	3	*that it is a*		*	
72.	3	2	*the form of the*			*
73.	3	2	*the name of the*			*
74.	3	2	*the passive form of*	*		
75.	3	2	*this has been the*		*	
76.	3	2	*this is not necessarily*		*	
77.	3	2	*through the use of*		*	
78.	3	2	*to be the most*		*	
79.	3	2	*you may need to*	*		

Appendix R

Textbooks bundles list (in-context information)

Total No. of N-Gram Types: 102				In-context information		
Rank	Frequency	Range	Bundles	Instructional parts	Reading parts	Both parts
1.	38	3	*focus on your subject*	*		
2.	34	4	*as a result of*		*	
3.	33	4	*it is important to*			*
4.	32	3	*look at the following*	*		
5.	31	4	*and answer the questions*	*		
6.	29	4	*the words in the*	*		
7.	29	3	*work in pairs and*	*		
8.	28	3	*essay with the title*	*		
9.	27	4	*at the end of*			*
10.	27	4	*what do you think*	*		
11.	25	3	*from the text in*	*		
12.	25	3	*use a dictionary to*	*		
13.	23	3	*in the text in*	*		
14.	23	4	*in the United States*			*
15.	23	3	*you are going to*	*		
16.	21	4	*the information in the*	*		
17.	20	3	*the words in bold*	*		
18.	19	4	*can you think of*	*		
19.	19	2	*the text on page*	*		
20.	18	4	*on the other hand*		*	
21.	17	4	*in a way that*			*
22.	17	3	*of the text in*	*		
23.	17	3	*research shows that in*	*		
24.	17	3	*the subject of the*	*		
25.	17	4	*the ways in which*			*
26.	17	2	*what you have read*	*		
27.	17	3	*you have been given*	*		
28.	16	3	*answer the following questions*	*		
29.	16	2	*answer the questions about*	*		
30.	16	4	*at the beginning of*			*
31.	16	3	*from the text on*	*		
32.	16	4	*the end of the*			*

33.	16	3	*the following extracts from*	*		
34.	15	3	*common in academic writing*	*		
35.	15	3	*have been asked to*	*		
36.	15	4	*in your own words*	*		
37.	15	4	*make a note of*	*		
38.	15	3	*the meaning of the*	*		
39.	14	4	*a wide range of*			*
40.	14	4	*do you think the*	*		
41.	14	3	*in bold in the*	*		
42.	14	2	*to check your answers*	*		
43.	14	3	*why do you think*	*		
44.	13	3	*at the same time*		*	
45.	13	2	*complete the following sentences*	*		
46.	13	3	*decide which of the*	*		
47.	13	4	*do you think are*	*		
48.	13	3	*look again at the*	*		
49.	13	4	*make notes on the*	*		
50.	13	4	*the way in which*			*
51.	13	2	*the way we do*		*	
52.	13	3	*the written academic corpus*	*		
53.	12	3	*are more likely to*			*
54.	12	3	*it is possible to*			*
55.	12	4	*one of the most*			*
56.	12	3	*read the text again*	*		
57.	12	4	*the beginning of the*			*
58.	12	4	*the rest of the*			*
59.	12	4	*what is the difference*	*		
60.	12	2	*with a similar meaning*			*
61.	11	3	*a similar meaning to*	*		
62.	11	2	*complete the sentences using*	*		
63.	11	2	*inferring the meaning of*	*		
64.	11	4	*look back at the*	*		
65.	11	3	*of the words in*	*		
66.	11	2	*the following extract from*	*		
67.	11	3	*the relationship between the*	*		
68.	11	3	*used in the text*	*		

69.	11	2	*which of the following*	*		
70.	10	3	*as part of a*			*
71.	10	3	*be more than one*	*		
72.	10	2	*check your answers in*	*		
73.	10	3	*discuss the following questions*	*		
74.	10	4	*do you agree with*	*		
75.	10	2	*each of the following*	*		
76.	10	2	*in the box to*	*		
77.	10	3	*in the form of*			*
78.	10	4	*is the most important*			*
79.	10	3	*research shows that the*			*
80.	10	2	*scientists and their work*			*
81.	10	2	*that something is true*		*	
82.	10	4	*the first paragraph of*	*		
83.	10	3	*the first part of*	*		
84.	10	3	*the phrases in bold*	*		
85.	10	4	*the title of the*	*		
86.	10	2	*used to refer to*			*
87.	10	3	*you think are the*	*		
88.	9	4	*a great deal of*		*	
89.	9	3	*a large number of*			*
90.	9	3	*a piece of writing*			*
91.	9	2	*article in a journal*	*		
92.	9	3	*be followed by a*	*		
93.	9	3	*can be used to*			*
94.	9	2	*from the same family*	*		
95.	9	2	*in the correct order*	*		
96.	9	2	*it is a good*	*		
97.	9	3	*the advantages and disadvantages*			*
98.	9	3	*the correct form of*	*		
99.	9	4	*the use of computers*		*	
100.	9	2	*the verbs in the*	*		
101.	9	4	*to write an essay*	*		
102.	9	3	*what you already know*	*		