A Corpus-Based Investigation of Lexical Bundles and Keyness in B1, B2 and C1 ESL Learners' Academic Writing

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by Hattan Hejazi

May 2022

Table of contents

List of figures

List of tables

Acknowledgements

I am very grateful to God for making it possible for me to conduct and complete this study.

I thank my supervisor, Dr Christian Jones, for his dedication and commitment to supervising me. I thank him for his patience and belief in my ability to conduct this study. To my second supervisor, Professor Paul Simpson thanks for your support and encouragement to complete my thesis. I wish to thank the English Department of the University of Liverpool. I am grateful to my friends and colleagues for always encouraging me.

To my family, including my Parents Helal and Fawzyah, my brothers Hani, Haitham, Ahmed and Husain, my sisters Najah and Nedaa, thank you for your love, support, and unwavering belief in me. Without you, I would not be the person I am today.

Above all, I would like to thank my wife Ghada for her love and constant support, for all the late nights and early mornings, and for keeping me sane over the past few months. Thank you for being my muse, editor, proofreader, and sounding board. But most of all, thank you for being my best friend. I owe you everything. I would like to express my deepest gratitude to my cute daughters Toleen, and Talah and to my son Ahmed for support, patience, kindness and smile through the journey.

Finally, despite my love for the English Language, the work reported in this thesis would not have been possible without the financial support from the Saudi Government, for which I am grateful.

To all of you I say, thank you very much!

Abstract

A CORPUS-BASED INVESTIGATION OF LBS AND KEYNESS IN B1, B2 AND C1 ESL LEARNERS' ACADEMIC WRITING

Hattan Hejazi

The current study aimed to investigate variations in, and the developmental use of lexical bundles (LBs) in argumentative essays by ESL learners at different proficiency levels. The study attempted to address two major research objectives. First, to investigate the use of LBs and keybundles in academic writing in ESL learners at the B1, B2 and C1 levels to produce empirical data concerning possible variations in frequency, structures and functions associated at the three levels. Second, track the developmental use of LBs in written essays by ESL learners at the B1, B2 and C1 levels over time through a longitudinal study, in order to provide empirical data to measure the relationship between LBs and language proficiency levels. A combined qualitative and quantitative methodological approach was employed to investigate the use of three- and four-word LBs and their grammatical distribution in the writing of ESL writers.

The findings of the present study were as follows. C1 writers tended to use a greater number of different LBs with greater frequency than those at B1 and B2 levels, and shared more features of written discourse than other levels. Therefore, the increased use of LBs can predict the learners' performance at least at high proficiency level C1. Furthermore, ESL learners tended to use more verb-based bundles and research-oriented bundles in their essays, similar to those found in the BAWE corpus. However, closer inspection of the concordance lines of the keybundles showed some informality in their writing, especially at levels below B2.

When time interacted with proficiency (CEFR level) to affect the use of LBs, the results showed that it is at the CEFR-B2 level when learners show development in the use of LBs, as showing a transition from using more informal written expressions (i.e., verb-based bundles) to more academic style (i.e., noun-based and preposition-based bundles). Overall, the findings suggest that LBs are considered a pivotal means of distinguishing academic writing by learners at different proficiency levels. The results have significant implications for the design of teaching material when teaching LBs in academic writing.

List of abbreviations

- **BNC**: British National Corpus
- **BAWE**: British Academic Written English
- **CEFR**: Common European Framework of Reference for Languages
- **CL**: Corpus Linguistics
- **EAP**: English for Academic Purposes
- **EFL**: English as a foreign language
- **ESL**: English as a second language
- **KWs**: Keyword
- **L1**: first language
- **L2**: second language
- **LBs**: Lexical bundles
- **LL**: Log-likelihood
- **NNS**: Non-native speaker
- **NS**: Native speaker
- **RC**: Reference corpus
- **WST:** WordSmith Tool

1 Introduction

This chapter reviews the motivation behind the thesis. The first section starts by detailing the preliminary background information concerning the context, and the subsequent section moves on to the rationale for exploring lexical bundles (LBs). The following sections introduce the aim of this thesis and its value. Following this, the research questions are addressed in this thesis. The chapter concludes with an overview, outlining the structure of this thesis.

1.1 Background of the study

Writing is acknowledged to be one of the most challenging tasks for L2 learners and its mastery is essential to enable learners to use the English language to state their ideas, opinions, construct arguments and integrate a variety of viewpoints (Harmer, 2001; Torrance and Galbraith, 2006; Hyland, 2009a). Thus, high proficiency in writing performance is a vital for establishing successful communication, being considered an important component of language development (Geiser and Studley, 2002; Powell, 2009). Consequently, writing is important not only to ensure English language competence, but also to succeed in learning where English is the language of instruction (e.g., Leki and Carson, 1994; Chou, 2011). This is because of the prevalence of using written work as a primary form of assessment. In the words of Hyland (2009b, p.2), "only through language, whether in the form of a dissertation, viva, essay assignment or unseen exam, can students consolidate and display their learning to university gatekeepers and so progress to graduation and beyond". Therefore, being proficient in academic writing is essential to ensure academic success.

According to *student statistics report 2019 (Englishuk, 2019)*, there has been an increase in the number of non-native learners studying English in the UK, around half a million international students of all ages come to the UK every year, aiming towards a native or native-like proficiency. This growing interest confirms the importance of EAP courses within the field of language development. However, the specific language-related difficulties non-native English speakers face in the area of EAP are numerous. Previous researchers have noticed that second language (L2) learners encounter challenges in all four language skills (i.e., reading, writing, listening, and speaking) (Ferris and Tagg, 1996; Huang, 2005; Snow and Uccelli, 2009), with the most significant being in the area of academic writing (Evans and Green, 2007; Zhang and Mi, 2010).

One way to succeed in academic writing is using formulaic language, which enables students to create natural and fluent spoken and written texts. The importance of formulaic language has been long established over the past six decades, as single words have been determined to belong to larger lexical units whose meanings and uses differ from their component parts (Howarth, 1998; Shirazizadeh and Amirfazlian, 2021). It is increasingly acknowledged that certain sequences of words have functions that play an important role in the mastery of the language (Nattinger and Decarrico, 1992; Cowie, 1998; Schmitt, 2004; Qin, 2014), contributing to approximately 20–50% of written discourse (Biber et al., 1999; Erman and Warren, 2000). These combinations of words can be retrieved automatically from memory, making them an abundant source of lexical information in the learning process (Pawley and Syder, 1983). As claimed by Peters (1983), by incorporating formulaic language into discourse, learners can avoid grammatical mistakes and perform language production more quickly than is possible when composing language word-by-word. These expressions also enable learners to organise their ideas in context, and facilitate fluent linguistic production and communication (Hyland, 2008b; Li and Schmitt, 2009; Ohlrogge, 2009). Moreover, formulaic language helps learners achieve coherence and reach a higher level of language proficiency in terms of fluency and accuracy. Therefore, the appropriate and frequent use of formulaic language is a component for advanced and fluent writing; conversely, the absence or misuse of such formulaic language is one major indication of a novice writer or a lack of expertise in the academic context (Mccann, 1989; Wray, 2002; Kecskes, 2016).

Despite its apparent importance in language competence, there are several types of formulaic language (Howarth, 1998; Wray and Perkins, 2000; Wray, 2008). A number of terms have been adopted in the literature when referring to this phenomenon. Wray (2002) compiled the expressions commonly used in the literature to describe the formulaic aspects of language, and identified more than 40 terms, which were used to refer to types of multi-word units. For example, collocations (e.g., Firth, 1957), multi-word units (e.g., Moon, 1998), recurrent word combinations (e.g., Altenberg, 1998), phraseology (e.g., Granger and Meunier, 2008), formulaic sequences/language (e.g., Schmitt et al., 2004), repetitive phrasal chunkiness (e.g., Cock, 2000), and lexical bundles (e.g., Biber et al., 1999).

Under the umbrella of formulaic language, a large group of multi-word sequences are typically transparent in meaning, and often structurally incomplete units, which are referred to in this thesis as lexical bundles. As stated by Biber and Barbieri (2007), these expressions form the vast majority of formulaic language sequences. LBs are sequences of three or more words that co-occur more frequently than would be expected by chance based on frequency and dispersion thresholds (*e.g., in the case of the*, *on the other hand*) (Biber et al., 1999, p.183). In other words, LBs are identified empirically according to their frequency occurrences in a register, rather than their structures. They are also associated with larger phrases and clauses, serving as frames for expressing new information (Biber et al., 2004). These expressions repeatedly occur within the same register, demonstrating that language is "register specific and performs a variety of discourse functions" (Allen, 2009; Biber and Barbieri, 2007). For example, Biber et al. (1999) found that academic writing is characterised by more referential bundles, whereas spoken language is full of stance bundles. Therefore, they help shape the meaning of the text, and contribute to our sense of distinctiveness in a specific register (e.g., Biber et al., 2004; Hyland, 2008a; Chen and Baker, 2010; Wood and Appel, 2014). As stated by Biber and Barbieri (2007), LBs might serve as handy short-cuts or frames through which writers can scaffold propositional meanings and ideas with relative ease.

The study of LBs has been a topic of interest for researchers and instructors in the linguistic field since Biber et al. (1999) introduced the notion of LBs in The Longman Grammar of Spoken and Written English. Previous research into LBs confirmed that they are widespread in written registers, serving as the "building blocks of discourse" (Biber et al., 1999; Biber et al., 2004; Hyland, 2008b; Li and Schmitt, 2009; Chen and Baker, 2016; Liu and Chen, 2020). This line of research has demonstrated that the acquisition of LBs is therefore not only significant for the development of academic writing skills, but also a key component to signifies writing expertise in a manner that is essential to academic fluency (Salazar, 2014). Guided by the requirement to help such learners develop competence in academic writing, there are many studies of LBs on the various context, such as LBs on disciplinary variations (e.g., Hyland, 2008b; Pourmusa, 2014), native and non-native English writing (e.g., Uysal, 2012; Pan et al., 2016), written texts produced by language learners and expert writers (e.g., Qin, 2014; Pan and Liu, 2019), and non-natives of different levels of proficiency (e.g., Staples et al., 2013; Chen and Baker, 2016). It is generally agreed that native speakers, non-native experts, and novice non-native writers draw on LBs of different types.

In general, these studies have yielded valuable insights into how English users at different proficiency levels use language. However, the findings of these studies remain mixed, and the variations between L2 learners from different proficiency levels in terms of their use of different forms, structural and functional patterns into which LBs are categorised is not yet clear. This might be due to different corpus sizes, methodologies, and texts in different academic registers, or heterogeneity in corpus design, which might affect the use of LBs. More specifically, there are discrepancies in previous studies in terms of determining learner proficiency levels, which makes it difficult, if not impossible, to generalize across research results of this genre.

Recently, a number of studies examining LBs in written discourse have started using the Common European Framework of Reference (CEFR) (Council of Europe 2001) to discriminate between the learners' levels, which is arguably one of the most influential frameworks in language education, articulating the development of language proficiency through a number of levels. A study by Chen and Baker (2016) applied the CEFR scales to determine proficiency levels. They used re-sampled argumentative essays from the published learner corpora, restricted only to those by L1 Chinese learners retrieved from the Longman Learner Corpus (LLC) published between 1990 and 2002. However, their findings might not be replicable in other L2 contexts due to their restriction to only Chinese learners. In addition, examining data collected two decades ago may not generalise to the current use of LBs as language might change due to certain conditions and context. This viewpoint is supported by Hyland and Jiang (2018), who reported a considerable change in the functional distribution of LBs in response over time. Therefore, choosing a newer corpus could shed light on the accuracy of learners’ levels.

In light of the above reasons, the present study is motivated by the limited number of studies examining the use of LBs in writing of different proficiency levels, creating a need for more investigation into the variation and development of LBs across ESL learners’ levels. To the best of my knowledge, no previous studies have examined the variations in and the developmental use of LBs among English language learners enrolled on English language courses. These writing classes are required for all L2 learners wishing to attend university in most native English-speaking countries, and LBs have a critical function in writing proficiency. This thesis is distinguished from previous studies by focusing on LBs in argumentative essays by ESL learners enrolled in English courses across proficiency levels. The thesis attempts to provide a better understanding of the use of LBs in ESL learners' argumentative essays.

1.2 Rationale for exploring lexical bundles

The reasons for conducting this research came from my decade of experience working with English as a Foreign Language (EFL) students in Saudi Arabia: no matter what their proficiency level, students always encounter difficulties with writing in English. As a non-native English speaker, I was aware that one of the main reasons for the low standard of students' written work is their limited linguistic competence. Students may also produce incoherent and poorly written text due to their lack of experience of writing argumentative essays. Flipping this around, as a student for whom English is a second language, I often struggled with writing argumentative essays.

In order to improve my writing skills, I managed to spend extensive time writing more English essays, before reviewing them to highlight areas that needed improvement. When comparing my initial performance at the beginning of the course with more recent work, the results were satisfactory. However, on the whole, my writing still lacked coherence and cohesion. Trying to identify the reasons for these problems led me to explore an important aspect of producing coherence and cohesion in a text, and making sense of a particular context, namely the use of LBs.

Whilst many features are clearly important for developing academic writing, proficient use of LBs can help writers become more logical and coherent. Unfortunately, examining my essays and those of my Saudi friends with regard to the use of LBs showed that even the most proficient students are not aware of the use of these expressions in their writing. Having pinpointed this problem, I was motivated to explore how ESL learners, particularly those enrolled in EAP courses, use LBs. I was also interested in the prevalence of LBs in academic writing, and how the appropriate use of LBs can help language learners become more proficient in the written register.

The above acted as motivation to review some relevant literature, which showed that the traditional view is that developing fluency and coherence in academic discourse is largely affected by the use of LBs (Biber et al., 2004; Biber and Barbieri, 2007; Nesi and Basturkmen, 2006; Kashiha and Heng, 2013; Shin, 2018). Hyland (2012, p.153) claims that LBs are important for speakers and writers for three main reasons:

(1) “their repetition offers users (particularly students) ready-made sets of words to work with; (2) they help define fluent use and therefore expertise and legitimate disciplinary membership; (3) they reveal the lexico-grammatical community-authorized ways of making-meanings”.

These advantages of LBs have motivated me to analyse LBs in ESL learners’ argumentative essays. The proficient use of LBs can help writers become more logical and coherent (Hyland, 2008a). Having a description of the variation in bundles between ESL learners’ levels can then help teachers create targeted lessons that may help their students become more proficient in their academic writing.

As described in the next chapter, several studies have examined the use of LBs in various genres, registers, disciplines, and more specifically related to this thesis, across L2 learners’ proficiency levels (e.g., Staples et al., 2013; Cooper, 2016; Chen and Baker, 2016; Ruan, 2017). However, they have not generally considered the potential impact of frequency usage when predicting writing proficiency or the characteristics of LBs exhibited at different ESL levels. In addition, some of the research findings have resulted in lists of LBs that seem to be register/discipline-specific to some extent and cannot be generalizable to the entire range of language variety.

As discussed in the previous section, no previous studies have examined variations in and the developmental use of LBs among English language learners enrolled on English language courses. Therefore, the aim of this study was to examine the use of LBs in ESL learners’ argumentative essays. As explained above, as an EFL teacher and non-native English speaker, I am aware of the challenges that most L2 writers typically encounter when writing in English. It is hoped that the findings of this thesis will help ESL/EFL teachers and students be more conscious of the role of LBs in producing coherence and cohesion in a text.

1.3 Goal of the thesis

In order to address the existing gaps in the previous LBs research, this thesis aims to examine the variation and the development use of LBs in ESL learners B1, B2 and C1 academic writing. Thus, this study fulfils two major research objectives. The primary objective is to investigate the use of LBs and keybundles within academic writing at three CEFR levels: B1, B2 and C1, to produce empirical data concerning possible variations in frequency, structures and functions associated with the bundles identified at three levels. That will be useful to understand language variation specifically in English language learners (rather than university students or expert writers) academic writing.

The second objective is to track the developmental use of LBs in argumentative essays by ESL learners over time at three CEFR levels, so as to provide empirical data to measure the relationship between LBs and language proficiency levels. The three CEFR levels were selected to fulfil the purpose of exploring the use of LBs at a wide range of ESL levels, so as to track development across those levels: B1, B2 and C1. Identifying LBs in ESL learners' sub-corpora can be important indicators for determining the development of language users within these discourse communities. Hence, their grammatical and discoursal features may become practically perspicuous. That will be useful for language pedagogy, comparative analysis, cross-level analysis and language development for current and future research endeavours.

To achieve the aim of this study, I have taken step by step methods to address the research questions as described below:

1. Three sub-corpora for two rounds of comparison: The first part of this study is cross-sectional research using a collection of argumentative essays written by intermediate and advanced English language learners. These essays were rerated and built up, differentiating three learners' proficiency levels sub-corpora (B1, B2 and C1), yielding approximately 50,000 words at each level. The second part is longitudinal research tracking nine ESL learners over six months, producing three ESL learners' sub-corpora, each composed of approximately 20,000 words.
2. Identification of target bundles: The study used *WordSmith software* (*WST*) to identify LBs in the sub-corpora. This software identifies three- and four-word LBs, following a specified frequency cut-off point and dispersion threshold. The program provides lists of the most frequent LBs in each of the ESL learners' sub-corpora.

3. Frequency-based analysis: The normalized frequency of bundles from the three sub-corpora were compared to analyse the quantity while the type suggested the variety of bundle use.
4. The British Academic Written English (BAWE) corpus is used as a reference corpus to establish a basis for comparison in the use of LBs between B1, B2 and C1 levels.
5. Classification of bundles and analysis of their distributions: Biber et al. (1999) structural taxonomy and Hyland (2008b) functional classification of LBs were adopted to provide information regarding the role of these expressions in the learners' essays. This analysis helped to identify the characteristics of the LBs used at each level.
6. A keyness analysis was conducted to explore the variability and distinguished characteristics of the ESL learners' levels in the written discourse.

1.4 Value of the thesis

Corpus-based studies examining formulaic language and particularly the use of LBs and keybundles is a fruitful area of language research. By conducting comparative research, researchers can discover the characteristics of the language that different language users use, as LBs are important indicators when determining language competence. That would be useful for EAP/ESP teachers to provide L2 learners with more authentic language instruction and awareness-raising activities. Such a claim clearly suggests the high pedagogical value of corpus-based studies that examine the use of LBs in specialized academic writing. Such pedagogical value may be better summarized by Hyland (2008a), who stated that "gaining control of a new language or register requires a sensitivity to expert users' preferences for certain sequences of words over others that might seem equally possible. So, if learning to use the more frequent fixed phrases of a discipline can contribute to gaining communicative competence in a field of study, there are advantages to identifying these clusters to better help learners acquire the specific rhetorical practices of their communities".

A number of corpus-based studies have thus been conducted to better understand and explore how discourse is constructed in different genres. Many of them have focused on the use of LBs in a variety of academic writing (e.g., Chen and Baker, 2016;

Yang, 2017; Nekrasova-Beker and Becker, 2020; Shirazizadeh and Amirfazlian, 2021). The present study adds to such a tradition by investigating the use of LBs and keybundles in ESL learner argumentative essays. It is assumed that the results acquired in this thesis will shed light on several points:

1. It provides a clear picture of the written language used by ESL learners at three CEFR levels, namely B1, B2 and C1, which will be obtained by studying the variations and the development use of LBs in their academic writing. This study therefore could contribute to our understanding of ESL students' preferences towards language proficiency.
2. The study examines the similarities and differences of LBs choices found in ESL learners writing of different levels using wordlist and keyword functions in *WordSmith* software. With the wordlist list function, the findings reveal the most frequently found LBs and their structural and functional distributions. Meanwhile, with the keyword function, the findings might shed the light on lexical bundles' choices either exclusively found or unusually frequent in the ESL learners writing, reflecting the level of the group. The lexical choices appearing exclusively in the ESL learners' writings and those frequently found in the ESL learners' writings relative to the BAWE might somehow reflect the features that are commonly found and deemed 'acceptable' within a specific professional research community.
3. The findings contribute to an expanded framework for comparison of how ESL learners at different proficiency levels use LBs in argumentative essays. This contribution to descriptions of argumentative written essays can assist students to become more aware of the existence and frequency of LBs in academic prose and the functions inherent to' this register.
4. The results of this study could be beneficial for EAP teachers and language instructors who wish to enhance their teaching and overall materials development. By drawing attention to weaknesses and strengths of English L2 writers in the use of LBs in writing argumentative essays, researchers and educators could help them use technical terms and LBs preferred by community members in specific register. Rather than being only a mechanism, such measures could be part of a screening process designed to help ESL learners whose language skills are unlikely to ensure a smooth transition from one level to the next.

To conclude, the study aims to examine the use of LBs and keybundles used in argumentative essays by ESL learners at three levels, namely B1, B2 and C1. The study will identify language users' specific characteristics based on learners' essays. In addition, discovering how ESL learners at different levels construct LBs structurally and use them to convey specific functions related to topics is expected to be of great value for language research.

1.5 Research questions

To achieve the aims of the study, the following research questions are addressed:

1. What are the most frequent three- and four-word bundles found in ESL learners B1, B2 and C1 levels argumentative essays?
2. What differences exist in the structures and functions of LBs in ESL learners B1, B2 and C1 argumentative essays and proficient student writers?
3. What are the characteristics of keybundles deployed in ESL learners' essays in comparison with the BAWE writers?
4. To what extent does an increased use of LBs correlate with learners' level of proficiency?

By discussing these issues, the present study will improve our understanding of ESL learners with different proficiency levels who are trying to improve their English language to either join university or become members of a native-like writing community. It will also increase our understanding of the relationship between the use of LBs and language development. Moreover, it will clarify how the use of LBs in ESL writing has changed over time, since the commencement of learning English as a second language at ELCs in the UK.

1.6 Structure of the thesis

There are six chapters in this thesis.

- **Chapter 2** reviews the relevant research on LBs, serving as a foundation for the present study by placing previous studies into sub-sections according to their research aims. An extensive review was carried out to identify the research gaps left by previous studies and present the potential for further research. This chapter also details the importance of formulaic language and LBs, argumentative essays, CEFR levels and reviewing the relevant literature.

- **Chapter 3** describes a pilot study conducted using a corpus of B2 and C1 learners' academic writing. This analysis provides the researcher with ideas, approaches, and clues to enhance the capability to gather clearer findings in the main study. It also permitted a thorough check of the planned statistical and analytical procedures, giving an opportunity to evaluate the usefulness of the data. It served to guide decisions about alterations to the data collection methods, and analysis processes.
- **Chapter 4** describes the methodology applied. It offers detailed information on data collection, data processing, and data analysis. It first draws on the process carried out to compile the sub-corpora used in the study; these are comprised argumentative essays compiled from a range of UK language centres, as well as data extracted from the BAWE corpus, which was used as a reference sub-corpus. This chapter then describes the methods applied to identify three-and four-word LBs from the learners' sub-corpora and extract key-bundles (these units used more than expected in the target data compared to a reference corpus). This chapter concludes with the analytical framework used to answer the research questions.
- **Chapter 5** presents the main findings. It focuses on the analysis and discussion of cross-sectional and longitudinal studies. It includes a comparison of the most frequent LBs across the sub-corpora, followed by a structural and functional analysis of the bundles identified. These bundles are then compared with the reference corpus data to explore the characteristics of LBs used in ESL learners' sub-corpora.
- **Chapter 6** summarises the thesis, drawing together the results from chapter 5. The chapter also details the contribution of the study and highlights the research limitations. It also presents the implications for language learning and recommendations for future research.

1.7 Conclusion

The aim of this chapter was to introduce the proposed research and the motivation behind the decision to study ESL learners' use of LBs. The chapter then provided background to the study and depicted the gap in the local context, goal of this study, contribution this study will attempt to make and the proposed research questions. It is hoped that this research will contribute to a deeper understanding of the use of LBs in ESL research.

2 Literature Review

2.1 Introduction

This chapter seeks to accomplish four objectives. Section 2.2 provides theoretical background information on the nature and role of formulaic language in spoken and written discourse. It also highlights the particular characteristics of formulaic language, the importance of these expressions in language, along with the terminology used to describe this phenomenon. The section closes by highlighting the concept of LBs and the importance of this phenomenon in language users. Section 2.3 presents a theoretical and analytical framework for LBs. It begins by defining LBs and what distinguishes from other formulaic language, followed by reviewing the identifying characteristics of LBs (Section 2.4). Section 2.5 introduces the concept of keyword/keyness analysis and how those words, phrases, bundles play a role in second language writing. Section 2.6 looks at the Common European Framework of Reference CEFR Levels of English (CEFR) and its relevance to language competence. Section 2.7 defines and describes the importance of the argumentative essays in EAP classroom. Section 2.8 explores the research carried out on LBs and illustrates the different criteria used to identify the LBs in discourse. Finally, section 2.9 discusses the issue of teaching LBs in EAP classroom.

2.2 Formulaic language

In general terms, formulaic language can be defined as "a sequence, continuous or discontinuous, or words of 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" (Wray, 2000, p.465). Based on this definition, the primary characteristics of formulaic language are as follows: first, it is composed of multiple words; second, it carries a specific meaning and/or function; and finally, it is prefabricated and/or stored and retrieved mentally as a single unit (Wood, 2015). It has long been a topic explored in corpus-based research, and a large and growing body of literature is devoted to exploring diverse structural types of formulaic language in different speech and written corpora (e.g., Oakey, 2002; Conrad and Biber, 2005; Ellis et al., 2008; O'donnell et al., 2013; Omidian et al., 2018; Garner et al., 2019; Reppen and Olson, 2020).

Firth (1957) was one of the first researchers to investigate formulaic language and introduced the term 'collocation' to refer to these units. Following Firth's research, various pieces of collocation research in linguistic analysis have been published, and researchers have begun using different terms to refer to this phenomenon. Wray (2002) compiled the most common expressions used in the literature to describe formulaic aspects of language, and found that there were more than 40 terms used to refer to types of multi-word units, for example: collocations (e.g., Firth, 1957), multi-word units (e.g., Moon, 1998), recurrent word combinations (e.g., Altenberg, 1998), phraseology (e.g., Meunier and Granger, 2007; Granger and Meunier, 2008), formulaic sequences/language (e.g., Schmitt, 2004; Krummes and Ensslin, 2015), repetitive phrasal chunkiness (e.g., Cock, 2000), and lexical bundles (LBs) (e.g., Biber et al., 1999; Kim, 2020). In addition, researchers have employed a number of methods for identifying multi-word sequences depending on how they are operationalised, such as frequency, length, and idiomaticity, to name but a few. Therefore, although these terms can be embraced under the notion 'formulaic language', they have been developed to identify different types of sequences.

Given its ubiquity, formulaic language comprises a large body of language, and the ability to master these expressions is a component of language competence. This finding has encouraged researchers to investigate the use of formulaic language from different perspectives (e.g., in spoken and written language, across various levels and various genres), mostly in the English language. Studies indicate that more than one-third of language is comprised formulaic language (Nattinger and Decarrico, 1992; Biber and Conrad, 1999; Conklin and Schmitt, 2008; Conklin and Schmitt, 2012). For example, Biber et al. (1999) found that 21% of all words used in academic writing are part of such bundles. The authors argued that "(...) much of our everyday language use is composed of prefabricated expressions" (Biber et al., 2004, p.372). Foster (2001, p.85) supports this claim suggesting that around 25–32% of the spoken language of native English speakers is composed of multi-word sequences. With an estimate of 55%, Erman and Warren (2000) argue for an even greater percentage of spoken language being composed of formulaic expressions. Because of the prevalence of these sequences in language, mastering formulaic language is essential in achieving language proficiency (Wray and Perkins, 2000; Wray, 2002; Biber et al., 2004; Schmitt, 2010; Rafieyan, 2018). Such expressions occur frequently in an academic context, and are therefore important for language users; this is explained in more detail below.

Pawley and Syder (1983) argued that these combinations of words can be restored automatically from memory, which makes them an abundant source of lexical information at the early stage of the learning process. Peters (1983) also claimed that by incorporating formulaic language into discourse, learners can perform language production more quickly than is possible when composing language word-by-word. A number of previous studies have also claimed that acquiring these expressions not only helps learners to become fluent, but also to achieve a greater range and accuracy of these expressions (Nattinger and Decarrico, 1992; Howarth, 1998; Wood, 2006; Schmitt, 2010; Stengers et al., 2011; Henriksen, 2013; Saito and Liu, 2021). In one study, Saito and Liu (2021) investigated the production of a specific type of formulaic sequence, referred to as collocation, in L2 oral proficiency. Their findings showed a strong positive relationship between the length of speech and language proficiency, relying on the length of speech rather than the details of speech, may aid their production of language and sound more native-like when focusing. Schmitt (2010) also argued that the appropriate use of formulaic language increases learners' self-confidence as well as motivates learners by creating a sense of accomplishment, which might strengthen the learning process.

Coxhead and Byrd (2007) summarise the reasons behind the importance of formulaic language in academia as follows:

- The formulaic language 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;
- As a result of their frequent use, such sets become defining markers of fluent writing and are important for the development of writing that fits the expectations of readers in academia. (pp.134–135)

Despite their significance in language learning, mastering these expressions remains a challenge for L2 language learners, even at advanced levels. Previous research has found that L2 learners do not use formulaic language in the same fashion as native speakers (Laufer and Waldman, 2011; Foster, 2001). For example, Siyanova and Schmitt (2008) investigated the phraseological competence of adjective-noun collocations qualitatively and quantitatively. They examined the number of collocations used by both EFL learners and native speakers of English in their writing, as well as the accuracy and speed of their participants' judgements on whether the identified collocations may be considered native-like. The study found that, although both groups exhibited a similar number of collocations in their writing, their judgements regarding the given task demonstrated differences in terms of speed and accuracy; the quality of processing adjective-noun collocations also differed across the groups. This confirms that although L2 learners may make use of a large number of phraseological units in their language, this does not automatically equate to native-like language proficiency. One reason for this may be a lack of exposure, as the importance of formulaic language is often neglected in L2 language classrooms (Alali and Schmitt, 2012).

Despite the clear importance of formulaic language for language learners, it is necessary to note that not all formulaic language is the same, meaning the study of formulaic language can be hard to follow because of the various types of formulas developed, including idioms, proverbs, collocations, and LBs, to name but a few, each with their own features and behaviour (see section 2.3). Therefore, the present study will focus on a particular type of formulaic language, namely LBs, being recurrent contiguous lexical sequences (e.g., *on the other hand*) and first identified in the Longman Grammar of Spoken and Written English (Biber et al., 1999). By looking at one specific type of formulaic language, this study will use corpus data to examine the LBs and, consequently, understand phraseological patterning within the language. Given its pervasiveness in language than other formulaic language types, this motivates me to explore their usage in academic writing.

The study of LBs is important because they are widespread in different registers and they "are prominent due to their rigidity" that allow them to be a good standard for teaching and learning a foreign language, as they are easily identified (Biber et al., 1999, p.13). In other words, the importance of LBs has given support to the previous impressionist view that a large proportion of written or spoken language is made up of LBs, which have become "useful devices for the comprehension and construction of discourse" (Biber and Barbieri, 2007, p.284). Biber and Barbieri (2007, P.269) demonstrate that LBs form the vast majority of formulaic language sequences and function as building blocks in spoken and written discourse, an essential element of fluent linguistic production, and a key distinguishing feature of specific forms, registers, and genres. For example, Biber et al. (1999) found that 21% of the words in academic prose in LONGMAN learner corpus form part of such bundles. Thus, it is almost impossible to disregard LBs, due to their considerable occurrence in discourse across the conversation and academic prose registers. Hyland (2008a, p.41) also states that LBs have an important role in creating coherence, appearing more frequently than would be expected by chance, helping to shape meanings and "contributing to our sense of coherence in a text".

language learning and language fluency, researchers have examined the use of LBs in both oral and written language. It has been found that LBs are varied across registers, L1 writers, competence or expertise writers, disciplines, and levels (e.g., Biber et al., 1999; Hyland, 2008a; Öztürk, 2014; Chen and Baker, 2016; Reppen and Olson, 2020). Thus, the importance of studying LBs is evident in various broader areas, such as applied linguistics, second-language acquisition, language instruction, to name but a few.

All previous research on LBs has used corpora to identify the most common recurrent sequences of words and determine how those sequences can be interpreted as building blocks of discourse. A corpus provides a large amount of quantitative data and an opportunity to test ideas about specific language. It can also reveal instances of particularly rare or exceptional cases that could not be identified from looking at single texts.

This study seeks to build on previous research on LBs by examining those frequently occurring expressions in English essays written by ESL learners at B1, B2 and C1 language levels in the UK. Specifically, the researcher is interested in exploring how students from different levels use LBs in their writing and how their use is related to language levels. The term 'LBs' will be used as the main term to refer to this phenomenon throughout the thesis, as this is the term used by Biber et al. (1999) in a series of studies exploring LBs.

2.3 Defining lexical bundles

The term 'LBs' is narrowly defined as continuous three or more word sequences (e.g., *at the same time, one of the most*) which occur frequently in a corpus, to satisfy specified frequency and dispersion thresholds, for example, occurring at least 20–40 times per million words in three to six texts, or in at least 10% of the texts in a corpus (Biber and Barbieri, 2007; Chen and Baker, 2016). As stated by Hyland (2008a, p.44), LBs can be "identified empirically purely on the basis of their frequency rather than their structure". They are thus simply the highest frequency multi-word sequences in a register, identified automatically using corpus software (e.g., *WordSmith*), "regardless of their idiomaticity, and regardless of their structural status", thus, they become "prefabricated blocks" for discourse (Biber et al., 1999). The primary benefit of using this criterion of identification is that it eliminates the researcher bias in identifying a sequence as a bundle while also allowing them to create a replicable/verifiable, and possibly more viable, linguistic information source. This definition has been adopted in many studies of bundles, which has increased understanding of LBs' usage and provided standard identification criteria for such bundles (Biber et al., 2004; Cortes, 2004; Cortes, 2006; Hyland, 2008a; Hyland, 2008b; Grabowski, 2015; Liu and Chen, 2020).

The distinction between LBs and other formulaic language expressions, however, is very subtle due to small differences in the characteristics used to identify these expressions from corpora. Thus, if we compare, for example, LBs with two common formulaic language types in research (e.g., collocation, idiom), we find very few differences in the characteristics of these sequences.

For example, collocations which refer to the syntagmatic attraction between two (or more) lexical items (Lehecka, 2015), do not always appear as contiguous units, and their words do not need to be placed immediately next to each other. For example, the words in the phrase the *ingenious method* do not need to be connected to convey the meaning (*the **method** he used to end the fighting is **ingenious***. / *he used an **ingenious method** to end the fighting*). By contrast, LBs appear as contiguous units and usually part of a clause or phrase (Biber et al., 2004); this is only a possible feature in collocation but a core characteristic of LBs. In addition, collocations generally follow syntactic patterns (e.g., adjective + noun: *strong coffee*, *chilly night*) (Crossley and Salsbury, 2011), whereas LBs are identified disregarding any pre-defined linguistic categories (Ädel and Römer, 2012), and therefore often do not follow syntactic patterns (e.g., *one of the*). Rather, they are determined by their frequency of occurrence only, without any particular requirements for mutual expectancy of the words in the bundle.

LBs also differ from idioms in several ways. First, idioms have been defined as a group of words that occur in a more or less fixed phrase and whose overall meaning cannot be predicted by analysing the meanings of its constituent parts (e.g., *happy as a clam* and *set my teeth on edge*) (Simpson and Mendis, 2003). In other words, the meaning of an idiom cannot be derived from its individual component words (e.g., *piece of cake,* meaning 'easy' – none of the words in the idiom can be used to predict its meaning). By contrast, LBs are semantically transparent rather than idiomatic in meaning (e.g., *on the other hand*). In other words, most LBs are not idiomatic, and their meaning could be derived from the words that construct the bundle (Cortes, 2004). In addition, idioms are usually structurally complete, with fixed meanings. In contrast, LBs are statistical associations that can be extracted computationally. Therefore, they can be grammatically complete (e.g., *on the other hand*) or incomplete units (e.g., *one of the*) that occur across 'grammatical boundaries' (phrase or clause). Lastly, idioms are also less common in speech or academic writing than LBs (Biber et al., 2004; Biber and Barbieri, 2007). Therefore, LBs are more valuable to examine since users of language more frequently employ them.

There are several other characteristics that distinguish LBs from other forms of formulaic language. To better understand what LBs are, the following section provides a brief overview of their characteristics.

2.4 Characteristics of lexical bundles

It is important to identify the distinctive features of LBs in order that they can be analysed independently from other formulaic language types. There are several characteristics that indicate the nature of LBs and distinguish them from other types of formulaic language (Biber et al., 1999; Cortes, 2004; Hyland, 2008b; Chen and Baker, 2010; Ädel and Erman, 2012; Qin, 2014; Pan and Liu, 2019). This section will discuss these features in detail.

2.4.1 Frequency

The main characteristic that distinguishes LBs from other formulaic language expressions is their frequency of occurrences across texts in a corpus (Biber et al., 2004). LBs can thus be identified by applying a frequency-driven approach to a corpus made up of a large number of language productions of a specific register, with a specified frequency cut-off and dispersion criteria.

For word sequences to qualify as a LB, they must occur with high frequency in a corpus (Biber et al., 1999). As mentioned in the previous section, LBs commonly co-occur repeatedly in written or spoken language and are identified by their frequency rather than their structures; in this way, they are ‘prefabricated blocks’ for discourse. These expressions are stored in mind as a holistic chunk, then retrieved and used in language production (Wray, 2000; Wray and Perkins, 2000; Wray, 2002; Nekrasova, 2009). According to Biber et al. (2004), aside from their use for isolating and identifying LBs, frequency data also indicates the likelihood of any given multi-word sequence being stored and retrieved as an unanalysed sequence in the mental lexicon. LBs must be specified at a set frequency cut-off point (i.e., the minimum number of occurrences of a word cluster per million words). Previous research has suggested that the frequency cut-off point used to identify LBs ranges between 10 and 40 times or more in every million words (Biber et al., 2004; Biber and Barbieri, 2007; Hyland, 2008a; Chen and Baker, 2010; Jalali, 2015). However, the frequency cut-off point is somewhat arbitrary and “based on the aim and on the researchers’ evaluation of data manageability” (Chen, 2008, P.64), and there is no agreement in the literature on the correct cut-off point. For example, previous studies have used a minimum of 40 occurrences (e.g., Biber et al., 1999; Pan and Liu, 2019), 25 occurrences (e.g., Chen and Baker, 2010), 20 occurrences (e.g., Cortes, 2004; Hyland, 2008a; Hyland, 2008b), and ten occurrences (e.g., Shin, 2019) per million words. Many previous studies have also set different frequency cut-off points depending on the bundle size (i.e., the shorter the bundle, the higher the frequency threshold). For example, Cortes (2013) used a frequency threshold of at least 20 occurrences for four-word bundles, at least 10 occurrences for a five-word bundle, and 8 occurrences for six-, seven-, eight-, and nine-word LBs.

Gries (2008, p.423) points out that “it seems as if there is as yet no rigorous operationalisation of when something is frequent enough to be considered a unit in the above sense of the term.” Instead, researchers use their intuition to determine what is frequent enough or not. In other words, there is no specific frequency threshold, and the decision to set a frequency cut off-point is based on either previous research or researchers' intuition. Hunston (2002, p.147) also states:

> *How many examples of a three-, four- or five-word sequence are necessary for it to be considered a phrase [sic! I guess what is meant is "a phraseologism"; STG]? As this is not an answerable question [. . .]*

In small corpora, frequency cut-off points are always expressed in the form of a normalised frequency threshold per million words, converted into a raw frequency, which must be reached in a corpus of a given size. By multiplying this normalised threshold by the ratio between the corpus size and one million (e.g., cut-off point/corpus size*1million). In other words, a cut-off point of 40 occurrences in a million words is converted for a corpus of 50,000 words by multiplying 40 by 50,000, divided by one million). This normalisation provides a fair comparison between corpora of different sizes (Biber and Barbieri, 2007; Allan, 2016). However, several recent studies have argued that applying the same threshold in corpora of different sizes is problematic (Oakey, 2009; Hyland, 2012; Chen and Baker, 2016; Bestgen, 2018). For example, Hyland (2012, p.151) explains, "Such normalisation methods, which are widely used to compare individual words across different sized corpora, may, however, be unreliable when working with LBs, and more research is needed to establish their validity". Cortes (2015, p.205) also supported this view and said that "Comparison of bundles yielded by small corpora and large corpora has been shown to be problematic because applying the usual normalization formula results in unreliable figures."

Bestgen (2018) examined two, three- and four-word sequences in four corpora of different sizes to evaluate whether the frequency cut-off points that are commonly used for identifying LBs in corpora are high enough to ensure that the selected sequences are unlikely to result from chance. Bestgen noted that the corpus size is strongly associated with the efficiency of a normalised threshold. Therefore, he suggests that small corpora should use a higher cut off point to select high frequently LBs only, and avoid selecting bundles that could result from chance. Yet this approach does not avoid arbitrariness of the chosen thresholds, but it is important to take into account the size of corpora when applying the frequency cut-off points of different corpora sizes to achieve representativeness and comparability of the extracted bundles. Due to the smaller sub-corpora size in this study, I decided to follow Bestgen suggestion and adopted a high-frequency cut-off point for the small corpus, indicating that a narrower and more specific scope of target language tends to have only LBs that reflect formulaic language use in this context. On the other hand, a lower frequency cut-off point was adopted for the large corpus (i.e., reference corpus) to allow for a wider view of a language.

However, as Biber et al. (2004) stated, "frequency is only one measure of the extent to which a multiword sequence is prefabricated… We do not regard frequency data as explanatory. In fact, we would argue the opposite: frequency data identifies patterns that must be explained" (p. 376).

2.4.2 Dispersion criteria

Another criterion for identifying LBs is dispersion, which is important in order to "guard against idiosyncratic uses by individual speakers or authors" (Biber et al., 2004, p.75). To qualify as a bundle, the word sets must occur across multiple texts in a register. The dispersion threshold, therefore, provides a counterbalance of the identified bundles by requiring the identified bundles to have a wide distribution through the texts. According to Hyland (2012, p.152), the occurrence of LBs in multiple texts by various writers or speakers shows "some perceptual salience among users' conventionalisation within a particular discourse community". Furthermore, Wood (2015) suggests that the distribution/range criterion helps to reduce the possibility that certain bundle might be used more by a particular writer or a particular topic.

Similar to the frequency cut-off point, dispersion criteria also vary across the literature. For example, Biber (2006a) suggests that LBs should occur in three to five texts to account for author bias, while Hyland (2008b) set the threshold as 10% of the total corpus texts. In addition, depending on the corpus size, a different dispersion criterion was used in the previous studies. For example, a high dispersion cut-off point of five different texts in a corpus over 500,000 words (e.g., Esfandiari and Barbary, 2017; Pan and Liu, 2019); by contrast, a corpus of fewer than 200,000 might have a dispersion threshold of three different texts (e.g., Chen and Baker, 2010; Qin, 2014).

2.4.3 Length

The third parameter is the bundle size. In general, the length of LBs in previous research has varied due to the variety of structures and functions that are open to analysis (Biber et al., 1999; Cortes, 2002; Cortes, 2004; Biber and Barbieri, 2007; Hyland, 2008b; Chen and Baker, 2010; Durrant, 2015). Typically, though, four-word LBs (e.g., *one of the most*) were most commonly examined, followed by three-word bundles (e.g., *to sum up*). For example, Biber et al. (1999) found that 18 % of the words in academic prose were three-word LBs, which occurred 6,000 times per million words, with four-word bundles appearing 5,000 times per million words. Hyland (2008b) noted that four-word LBs are more common than other strings and carry more functional and structural value. By contrast, longer bundles such as five- and six-word LBs (e.g., *as a result of the, it can be concluded that the*) are less common, and their frequencies in a corpus drop off significantly (Biber et al., 2002; Allan, 2016). For these reasons, they have not been as commonly investigated. It is worth noting that three-word bundles are often too numerous and many of them are part of longer bundles, such as the bundle *on the other,* which usually forms part of the bundle *on the other hand*. However, some three-word bundles overlap with more than one four-word bundle, as in the case of *a lot of*, a common three-word bundle, which can be part of *a lot of people* or *a lot of information*; thus, interest in these is common. As Appel (2011b, p.69) observed, shorter bundles “seem to have become a standard unit of length in this type of research, but problems still persist” when dealing with the computer software. This is because the software divides the longer bundles into smaller overlapping bundles of three- or four-words. For example, *there are quite a* and *there are quite a lot of.*

The three discussed characteristics of LBs must be inputted into a corpus tool to provide the initial lists of LBs from a specific corpus (Biber et al., 2004; Hyland, 2008a; Chen and Baker, 2010; Jalali, 2015; Ruan, 2017).

2.4.4 Transparency

The fourth characteristic of LBs is idiomaticity. Most LBs are not idiomatic in their meaning; they are “semantically transparent and formally regular” (Hyland, 2008a, p.6). These bundles function as a unit, the meaning of which can be easily understood from the bundle component. For example, bundles, such as *it is important to* and *one* *of the most*, are entirely clear from the individual component words. On the other hand, idioms such *spill the beans* (reveal a secret) and *set my teeth on edge* (unpleasant) have idiomatic meanings and are rarely used in speech or academic writing.

2.4.5 Incompleteness

To date, there has been some agreement on the observation that bundles tend to be made up of syntactic fragments that extend across structural units, particularly in academic writing (Biber et al., 1999; Biber et al., 2004; Hyland, 2008b; Simpson-Vlach and Ellis, 2010). LBs are more likely to be structurally incomplete units, where they serve as frames for the expression of new information (e.g., *on the other hand* and *can be used*) that occur at the phrase and clause boundaries (Cortes, 2004). Most LBs bridge two structural units, starting with either a phrase or clause boundary, with the last word of the first bundle being the first word in the second lexical unit (e.g., *In this essay I* and *I will discuss*) (Biber, 2006a). Biber and Barbieri (2007) stated that LBs play an important role in discourse as they represent a kind of pragmatic head for larger phrases or clauses and function as a frame for new information.

Biber and Conrad (1999) found that just 15% of LBs in the spoken corpus represent complete units, as phrases or clauses, while less than 5% of bundles used in academic writing are regarded as complete units. Even though LBs are not usually structurally complete, they have been shown to serve an important discourse function (Biber et al., 1999; Cortes, 2004). They can also occupy different positions in a text. For example, verb-based bundles (e.g., *am going to*) and dependent clause fragments (e.g., *if they want to*) are often associated with spoken language. In contrast, noun-based bundles (e.g., *the importance of the*) and preposition-based bundles (e.g., *on the other hand*) are increasingly used in written language (Biber et al., 1999; Hyland, 2008b; Qin, 2014).

Biber et al. (1999) categorised LBs by their grammatical parts into 12 main structural categories in academic prose and 14 major categories in conversation, with some overlaps between the categories. This classification was modified by Biber et al. (2004), who proposed three main categories according to the main grammatical features (clausal or phrasal). The first category comprises LBs that incorporate verb phrase (VP) fragments (e.g., *it is important to, I do not know*); the second category incorporates dependent clause fragments in addition to simple verb phrase fragments (e.g., *to be able to, if you do that*), although these categories are described as clausal bundles; and the third category incorporates noun-phrase or prepositional-phrase fragments (e.g., *on the other hand, the number of the*), described as phrasal bundles. These classifications have been adopted widely in many studies in the same area (e.g., Biber et al., 1999; Biber et al., 2004; Hyland, 2008a; Vo, 2016). However, since Biber et al. (2004) taxonomy was based on spoken and written corpora, many researchers have modified it according to their study data or have classified LBs differently due to different registers of text or the subjective evaluation of concordance lines. For example, a new sub-category 'Verb phrase with active verb' was added to Biber et al.'s (2004) taxonomy in four studies (Chen and Baker, 2010; Qin, 2014; Lu and Deng, 2019; Pan and Liu, 2019). Chen (2008) suggests that it is pointless to distinguish between the first two categories of Biber et al. (2004) classification since they are both composed of verb components, and the use of dependent clauses in academic writing is uncommon.

Hyland (2008b) also adopted a taxonomy based on Biber et al. (1999) system, but employed fewer structural categories, replacing the 'Copula be + noun phrase/adjective phrase' subcategory with 'Copula be + noun phrase/adjective phrase' or 'Be + complement (noun phrase)', as shown in the table below.

Table 2.1. Structural classification of LBs.

Structural types	Sub-types
Verb-based	Anticipatory it + verb / adjective phrase
	Copula be + noun / adjective phrase
	Pronoun/NP + be
	First-person pronoun + dependent clause
	(verb/adjective +) to-clause
Noun-based	a. NP with of-phrase
	b. NP with other post-modifier
	Other noun phrases
Preposition-based	Prepositional phrase with embedded of-phrase
	Other prepositional phrase expressions

Hyland (2008b) simplified the taxonomy by grouping bundles into three major categories based on three grammatical structures (verb-based, noun-based and proposition-based), However, the taxonomy has been modified and developed as recommended by Biber et al. (2004) in order to place the identified bundles that could not be classified.

In addition to the structural correlations, LBs perform specific functions in discourse and have been classified according to their function correlations in discourse. One widely accepted taxonomy is that of Biber et al. (2004), which distinguishes three primary discourse functions: stance expressions, discourse organisers, and referential expressions; where each discourse function consists of several sub-categories. Stance bundles most often refer to the speaker's knowledge of or attitude toward the information in the following proposition. Referential bundles are defined as expressions that "make direct reference to physical or abstract entities, or to the textual context itself" (Biber et al., 2004, p.384). Finally, discourse organisers are defined as "relationships between prior and coming discourse" (Ibid). This study compared the LBs in classroom teaching and textbooks to those identified in previous research on conversations and academic prose.

On the grounds that the categories identified by Biber were intended to differentiate between spoken and written modes of discourse, and so were not all directly applicable to research focused on written registers. Hyland (2008b) developed the aforementioned functional taxonomy of LBs, with different designs applicable to the domain of academic writing (as an alternative to the range of registers considered by Biber et al. (2004), as illustrated below.

Table 2.2. Functional classification of LBs distribution. (Hyland, 2008a)

Functional types	**Sub-types**
Research-oriented	Help writers to structure their activities and experiences of the real world (corresponded to referential bundles).
• Location	Indicating time and place
• Procedure	Indicating methodology or purpose of research
• Quantification	Describing the amount or number
• Description	Detailing qualities or properties of the material
• Topic	Related to the field of research
Text-oriented	Concern with the organisation of the text and its meaning as a message or Argument (correspond to discourse orienting)
• Transition signals	Establishing additive or contrastive links between elements
• Resultative signals	Mark inferential or causative relations between elements
• Structuring signals	Text-reflexive markers which organize stretches of discourse or direct readers elsewhere in the text
• Framing signals	Situate arguments by specifying limiting conditions
Participant-oriented	Focused on the writer or reader of the text (correspond to stance bundles)
• Stance features	Convey the writers' attitudes and evaluations
• Engagement features	Address readers directly

To summarise the characteristics above and the observations made in this section, it can be concluded that LBs are a continuous set of three or more words, identified by their frequency and dispersion criteria. Furthermore, these expressions are usually incomplete structural units, mostly non-idiomatic in meaning, but carry out specific discourse functions. Therefore, they are much more than sequences of individual words put together by chance; these expressions have specific functions and meet specific communicative needs (Biber and Barbieri, 2007; Hyland, 2008a; Wood, 2015). They have essential features that contribute to a sense of distinctiveness in a specific register (Biber et al., 2004; Hyland, 2008a; Chen and Baker, 2010; Wood and

Appel, 2014). At the same time, the absence or inappropriate use of these bundles is a sign of a novice writer or a lack of language competence (Hyland, 2008b).

Although LBs' characteristics serve as a good starting point for any research examining this phenomenon, the thresholds used to identify LBs may need to be adjusted depending on the size of the corpus used and the level of specialisation being examined.

2.5 Keyword/Keyness analysis

Although there have been many studies of Lexical Bundles, there have been limited studies examining keywords in relation to LBs and how they characterise discourse patterns. Keyword analysis has been employed as a stepping-stone for several kinds of textual analyses, particularly corpus-based analysis (e.g., Freddi, 2005; Kwary, 2011; Seale, 2008). This method is widely used because it provides keyword lists, which can highlight important concepts and styles of the texts or corpora under investigation (Scott, 2009). In language studies, the term "keyword" has no clear definition. It has been defined by Scott and Tribble (2006) as indicators of a text's "aboutness" and style. Meanwhile, Scott (1997); Baker (2004); Scott (2008) have described Keywords as those words, phrases, bundles etc., whose frequency is unusually high or low in a given text or corpus when compared to a reference corpus, regardless of their importance within that corpus. Although they might not be the most significant terms in a text or corpus, they serve as a valuable means of describing a text or genre, and may be used to examine the lexico-grammatical characteristics of a corpus. Consequently, as argued by O'keeffe et al. (2007); Scott (2012), the study of keywords has significant value when used in linguistics and other relevant domains, such as language education or stylistics.

The analysis of keywords is one of the techniques facilitated by a computational tool that involves comparing two wordlists, one being the target corpus and the other a reference corpus, which is a somewhat larger collection of texts that is more general in nature than the target corpus. Setting an appropriate metric for keyness analysis would then provide a list of keywords, whose frequencies are statistically higher or lower in the study corpus than the reference corpus, and indicate that 'the larger the difference, the more "key" a word would be' (Gabrielatos, 2007, p.4). According to Scott (2015), keyness represents the quality of a word or phrase that has been identified as a key in a given corpus or text. Scott (2008, p.64) adds that a word is considered to be statistically key if ‘its frequency in the text, when compared with its frequency in a reference corpus, is such that the statistical probability as computed by an appropriate procedure is smaller than or equal to the *P-value* specified by the user’.

Therefore, because keywords are statistically significant in a text/ texts, they tend to indicate the text’s ‘aboutness’ – simply, what the text is about – as well as its style (Groom, 2010; Scott, 2010). Goźdź-Roszkowski (2011, p.35) also argues that keywords can ‘reveal not only a great deal about the subject matter, the “aboutness” of a particular genre, but they can also specify the salient features which are functionally related to the genre’. Thus, they are not only frequently regarded as ‘useful indicators of the characteristic style of a particular text or corpus’ (Groom, 2010, p.59), but also ‘often provide a way of identifying which words best distinguish the texts of a particular author or group of authors from another’(Hyland, 2012, p.68). Accordingly, the merit of using a keyword analysis is that it offers an empirical discovery method based on frequency and distribution, which will direct the researcher to important features of a text (in comparison with other texts), as keywords differ in accordance with different textual collections. In summary, the term keywords provides a frequency-based approach to determining the characteristics of a given text or genre by calculating the keyness.

In corpus linguistics research, there is growing interest in keyword analysis and in determining the lexico-grammatical features of the studied texts in various genre types (e.g., Scott, 1997; Cacchiani, 2011; Grabowski, 2013; Grabowski, 2015; Palmer, 2016; Pojanapunya and Todd, 2018; Mutiara, 2018). Keyword studies are not only conducted to identify literary style, but have also investigated language patterns in academic and scientific prose, providing valuable lexicon and grammar usage details for instruction in English for Specific Purposes (ESP). When investigating language patterns in academic writing, Goźdź-Roszkowski (2011) examined patterns of linguistic variation in American Legal English. The study conducted a keyword analysis to determine the common words in one legal genre compared to others. They found that some operative genres such as legislation and contracts contain more formulaic language, which is evidenced by the high occurrence of LBs. By contrast, the persuasive genres included a wider variety of lexical expression ‘as reflected in the use of a larger set of different low-frequency word types’ (ibid., p. 227).

Turning to the use of LBs, Grabowski (2015) presented a corpus-driven study of the use of LBs and the functions of keywords across four types of pharmaceutical texts (samples of patient information leaflets, summaries of product characteristics, clinical trial protocols, and chapters from academic textbooks on pharmacology) to determine the patterns of language use by the four pharmaceutical registers. That study revealed language patterns change considerably as a consequence of topic- and function-related differences between the text types. It also emerged that the use of keywords and LBs also varied across texts. For example, academic textbooks presented the largest number of keywords and drew heavily on specialist vocabulary from the field of pharmacology. By contrast, the other three sub-corpora were far more formulaic, as the writers 'follow specific guidelines for the macro-structure and the scope of information conveyed by these text types' (ibid., p. 27).

Another study by Palmer (2016) explored the lexical and phraseological resources associated with the communicative functions of grade comments of eleven sub-corpora. The study identified a possible connection in grade comments, between keyword-containing LBs and framing devices for generic features, specifically opening and closing moves. Moreover, the keywords showed the quality of meta-grammatical terminology utilized by learners to construct goals for each class.

Focusing specifically on academic writing, Pang (2009) used the keyword function of *WordSmith* tools to examine the use of LBs identified from LOCNESS compared to the LBs identified from the WECCL corpus. The results identified approximately seventy-two key bundles that were overused by native English speakers compared with non-native learners. In contrast, there were 245 key-bundles used by Chinese learners that were significantly underused by native English speakers. The study claimed that most of the overused LBs in the WECCL corpus were apparently content-based and had equivalents in the Chinese language.

A recent study by Mutiara (2018) explored LBs and keywords in psychology research articles. In terms of keyword analysis, the study found that some words have patterns in their use in the research articles and co-occur with LBs, which demonstrates how LBs and single words possess specific meanings in discourse. For example, when the word *gender* collocates with the word *differences*, it is sometimes followed by prepositional phrases that begin with *in*.

All the reviewed studies have demonstrated that the methods of keywords analysis have been conducted in various types of text, are useful to describe the contents of a document, and can be examined from various perspectives. Overall, keywords have contributed to our knowledge of the characteristics of language patterns. However, there has been little research on the use of LBs addressing the keyword variations across proficiency levels; in addition, keyness is still poorly understood, and much additional exploratory work needs to be done (Scott, 2010). Thus, and as a departure from most of the previous research identifying LBs in academic writing, this study attempts to examine this genre from a lesser known perspective: that is, to investigate the 'key lexical bundles' used in ESL learners' academic writing from different proficiency levels, and to explain any differences between the ESL learners' levels in terms of language use. The notion of keyness is one of the objectives that can provide useful insights into the characteristics of ESL learners' academic writing. Moreover, a comparison between the lists of keybundles helps identify differences between the ESL learners' levels written discourse.

In order to avoid confusion with commonly known terms, 'keyness analyses' is employed instead of 'keyword analyses', and the term 'key bundles' replaces 'keyword', since the study focusses on LBs rather than single word.

2.6 CEFR Levels of English and its relevance to language competence

The Common European Framework of Reference for Languages (CEFR) was developed by the Council of Europe and has been translated into thirty-nine languages (Coe, 2014). It is a transparent, coherent, and comprehensive reference instrument useful for directors, syllabus designers, teachers, teacher trainers, and proficient learners. It 'views users and learners of a language primarily as "social agents"; in other words, members of society who have tasks (not exclusively language-related) to accomplish in a given set of circumstances, in a specific environment and within a particular field of action' (Europe, 2001, p.9). The CEFR uses an 'action-oriented approach' for describing language proficiency, describing and evaluating what learners do, how they act, or compete for tasks (actions) with the language in a variety of contexts. These actions are codified according to the scales as CEFR descriptors. These are referred to as 'Can–Do' descriptors which define communicative competence first in terms of the learners' Can-Do in the L2 language (Byrnes, 2007).

According to the Council of Europe’s, the CEFR is much more than proficiency scales. The central function of the CEFR is as follows:

> “The Common European Framework provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe. It describes in a comprehensive way what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively.” (Coe, 2014).

This framework categorises learners into six proficiency levels (A1 and A2, B1 and B2, C1 and C2). These levels are roughly equivalent to Beginner, Intermediate, and Advanced, although the CEFR levels are more precise than these terms (and refer to them as Basic, Independent, and Proficient). This structure aims to ‘help language professionals further improve the quality and effectiveness of language learning and teaching (Europe, 2001) and has been widely used in a variety of language-related areas, such as language education and language testing. The framework allows teachers and language instructors to measure students’ progress at each level of language learning; these levels are combined to describe the four main language skills (Reading, Writing, Speaking, and Listening) of language capacity at each CEFR level (see Appendix A The six proficiency levels are important tools for defining a student’s competence in the target language; they are used to assess language in education and can be utilized to describe language competence for L2 learners – that is to say, students of a language other than their native tongue (Szudarski, 2017). The CEFR holistic scales aim to reflect the trajectory of L2 learners as they make improvements in a second or foreign language.

A number of studies have examined the lexical and grammatical features of a particular language at different CEFR levels and the rating procedure across different tasks and languages (e.g., Glaboniat et al., 2005; Huhta et al., 2014; Byrne, 2016; Chen and Baker, 2016). For example, Huhta et al. (2014) explored the quality rating procedures used in SAL research across multiple tasks and languages. The study used five different tasks in two languages (L2 Finnish and FL English) from informal to semi-informal to formal tasks. They asked nineteen moderators to assess the English and Finnish performances, all of whom were language education professionals, with some being experienced moderators. The study found that the moderators presented unique information, along with the fact that the procedures were reliable, and the scales applied were valid for the purpose of the study.

Chen and Baker (2016) also addressed the aspects of discourse through the use of LBs across the CEFR levels. Their study used the CEFR scale to define the learners' writing development by their proficiency level, using argumentative or expository essays written by L2 Chinese identified from the Longman Learner Corpus (LLC). The essays were rated according to a robust procedure from the manual for Relating Language Examinations by the Common European Framework of Reference for Languages (Europe, 2003). After the rating, the LLC was divided into three sub-corpora representing CEFR levels B1, B2 and C1. The studies revealed that the B2 level learners had begun to show signs of transition and were able to distinguish between formal and informal writing, whereas C1 writers were clearly showing a formal style typical of the written genre. Although the lack of idiomaticity was marked in B2 writing and still lingered in C1 writing, the evidence found in this study demonstrates that L2 proficiency growth can be assessed using formulaic sequences (e.g., lexical bundles, collocation). This perspective provides a further point of investigation for my thesis. Although research on second language acquisition in both written and spoken registers has focused on levelling proficiency following the CEFR, studies have also emphasised the need for further exploration into how the different levels are associated with particular characteristics of L2 proficiency. (Hulstijn, 2007). Certainly, it is more precise to evaluate learners' writing proficiency independently from their overall L2 English language proficiency level, since there is no evidence that the overall CEFR level is necessarily correlated to the different dimensions comprising L2 proficiency (i.e., writing proficiency). In this study, the CEFR written framework will be used to describe ESL learners' written language at various levels. The following sections review previous research on LBs in various aspects of written registers.

2.7 Argumentative essays

One type of academic writing has received relatively little research attention in the literature of LBs, although it is probably the most common genre of academic study (Mei, 2006; Wingate, 2012); the argumentative essay. This is the most frequent format that university students have to write in, particularly in the fields of arts, humanities, and social sciences (Hewings, 2010). Argumentative essays are defined as 'argumentative or expository in character, i.e. besides presenting facts, they have the aim to explain, analyse, and interpret these facts and, usually, to argue for a certain standpoint' (Altenberg and Tapper, 1998, p.83). Additionally, such an essay requires well-structured research, accuracy, details, and critical and logical thinking to support the ideas coherently (Parkinson and Musgrave, 2014). It uses logic and reason to prove that one idea is more legitimate than another; it attempts to persuade a reader to adopt a certain point of view or take a particular action. The argumentative essay can connect the linguistic and conceptual components of multiple genres in one essay. For instance, it may include spoken language characteristics such as first personal pronouns and short sentences and include some academic writing features such as long sentences, critical analysis, and noun phrases (Jaworska et al., 2015).

Argumentative writing skills have long been identified as important at various levels in academic studies (Applebee et al., 1994; Németh and Kormos, 2001), and 'general life' (Crowhurst, 1990, p.349). Although the nature of the essay varies considerably across and even within disciplines, developing an argument is regarded as a characteristic of successful writing (Lea and Street, 1998). In fact, argumentation plays a significant role because it is one of the indispensables 'soft skills' of life. Having a good argumentative ability is extremely advantageous. At the various language levels, for instance, great importance is attached to writing argumentative essays, where learners are required to show their writing competence by presenting their ideas and expressing their opinions and thoughts in a well-structured and appropriate form (Feak and Dobson, 1996; Varghese and Abraham, 1998). In addition, the ability to write argumentatively is also a required skill for higher education. For over a century, it has remained the 'gatekeeping mechanism within individual courses as well as at critical stages of passage through secondary schools and into college' (Heath, 1993, p.105). English medium universities evaluate this skill through various tests such as the International English Language Testing Systems (IELTS) and the Test of English as a Foreign Language (TOEFL). Learners need to 'provide general factual information, outline and/or present a solution, justify an opinion, and evaluate ideas and evidence' (Ucles, 2002). According to Nesi and Gardner (2006) survey of assessed writing in twenty disciplines, a commonly recognized value of the essay is its 'ability to display critical thinking and development of an argument within the context of the curriculum' (p. 108).

Argumentative essays are therefore by far the most common educational task students undertake in order to demonstrate secondary learning and writing proficiency (Moore and Morton, 2005; Wilcox and Jeffery, 2014), to determine placement for higher educational studies (Gere et al., 2013; Aull, 2015), and to develop writing within specific disciplines (Wingate, 2012; Dryer, 2013; Crossley and Mcnamara, 2014).

Students are expected to have the essential skills to write well-structured argumentative essays when entering university; they are therefore by far the most common writing tasks used to differentiate between novice and advanced writers. The increased demand for the number of L2 learners studying in English-medium universities has highlighted the importance of competent argumentation writing skills. Additionally, as stated by Aull and Lancaster (2014, p.153) ‘thousands of English first language (L1) and L2 students enter general writing courses and are assumed to transition into the university-level discourse in this way’. Students are expected to arrive at university already possessing certain writing skills regardless of their discipline, alongside having the ability to express their thoughts within a well-structured argument; however, previous research has shown that many L2 learners fall short of these expectations and face many challenges to their mastery of argumentative writing, with only small numbers of L2 students being rated as ‘competent or better’ writers (Mccann, 1989; Knudson, 1992).

The challenges faced by L2 learners when asked to write an argumentative essay are due to several recurrent problems. Hyland (1990) suggests that one such issue could be associated with an ‘inadequate understanding of how texts are organized’. This viewpoint has been confirmed by Butler and Britt (2011), who reveal that students struggle to write a complete and well-structured argumentative essay. In addition, Wolfe et al. (2009) examined a cognitive argumentation schema for written arguments and found that university students have difficulty in responding to the problem’s writing prompts, or do not know how to express their opinions verbally. Another suggested obstacle is a tendency for the writer to demonstrate bias in order to support his own opinion (Perkins, 1985; Knudson, 1992; Knudson, 1994; Wolfe and Britt, 2008). Finally, learners often have a vague idea of the skill of argumentation, which could be linked to a lack of understanding of the concept of the argumentative essay.

In order to tackle these challenges and demonstrate proficiency in writing, learners require evidence from multiple sources to support their ideas logically, to increase their knowledge of the topics in question, and to have the ability to verbalize their thoughts efficiently (Riley and Reedy, 2005; Nippold and Ward-Lonergan, 2010). Therefore, argumentative writing is a complex task that requires both cognitive and linguistic skills. In order to improve their language abilities learners must master all of these skills, alongside developing and logically defending a specific position (Campbell and Filimon, 2018).

Research on argumentative essay writing in most studies of university writing in general has primarily confirmed the context of genre. It has explored, for instance, the history of the assessment of writing, student writing performances vis-à-vis the topic of the task, and whether the argumentative essay is among the genres that students would be expected to encounter (Haefner, 1992; Heath, 1993; Destigter, 2015; Burstein et al., 2016). In addition, research on applied linguistics, particularly English for Academic Purposes (EAP), has shown the crucial use of argumentative essays in university writing tasks (Qin and Karabacak, 2010; Dastjerdi and Samian, 2011; Campbell and Filimon, 2018). Anada et al. (2018) investigate the rhetorical features of IELTS essays, and argue that certain features such as argumentation and the argument structure should be taken into account to determine the quality of students' essays. The study reveals that these features are important to determine the quality of the organization of the IELTS essays, and also highlighted the significance of this organization in delivering the intended message. In another study, Jo (2017) examined the quality of rated argumentative essays by adolescents from three different cultures: Russia, China, and the USA. Interestingly, the study found that the linguistic aspects were essential for increasing the quality of the essays, while other aspects such as rhetorical questions, examples used, and appeals made little or no difference to their quality.

Previous studies have also confirmed that each genre of text possesses its own distinctive linguistics features (Biber and Conrad, 2009; Biber and Egbert, 2018). For example, some studies have found that the characteristics of verb choice are specific to certain types of writing, such as news texts and editorial newspaper articles (Oktavianti and Ardianti, 2019; Oktavianti and Adnan, 2020). Academic essays also have their own linguistics features compared to other genres of texts. In formulaic language, LBs are ubiquitous in academic texts, showing that these word sequences are 'building blocks of discourse' (Biber et al., 1999; Biber et al., 2004; Hyland, 2008b; Li and Schmitt, 2009; Chen and Baker, 2016; Liu and Chen, 2020). Yang (2017) investigated LBs in narrative and argumentative writing by Chinese EFL Learners. The study used data from WECCL (Written English Corpus of Chinese

Learners), alongside a sub-corpus of SWECCL (Spoken and Written English Corpus of Chinese Learners) compiled by (Qiufang et al., 2003). The study claimed that learners used LBs more frequently in argumentative essays than in narrative writing. It also revealed that learners prefer to write argumentative texts as they are more highly structured than narration, which might assist them to use more LBs in order to express their opinions. It could thus be assumed that the argumentative essay is represented more than other genres in learners' writing development.

Previous LB research examined L2 argumentative essays by comparing them either with native speakers or professional writers, or with their counterparts at different levels (e.g., Chen, 2009; Juknevičienė, 2009; Dontcheva-Navratilova, 2012b; Ruan, 2017; Shin, 2019). These studies show how LBs are associated with genre, L1 impact, disciplines, and learners' levels. For instance, Shin (2018) investigated LBs in L1 Korean-speaking EFL learners and L1 native English speakers' argumentative essays produced in the first year of university study. The research compiled two non-native sub-corpora and a native sub-corpus for LB analysis. The authors observed that such LBs are also frequent in novice argumentative essays, regardless of the students' first language. It is therefore intriguing to examine the use of these expressions among L2 learners in argumentative essays.

2.8 Previous research on lexical bundles

The study of LBs in spoken and written discourse has been a topic of interest for researchers and instructors in the linguistics field since Biber et al. (1999) first coined the term 'LBs' in the Longman Grammar of Spoken and Written English. Many researchers have analysed LBs typical of academic writing in a variety of written texts, such as students' writing and published academic articles (e.g., Bal, 2010; Kwary et al., 2017; Liu and Chen, 2020). This body of research has confirmed that specific sets of bundles are widely used in written academic discourse (Neely and Cortes, 2009; Hyland, 2012) and that their accurate use has been a signal of mastery of language fluency (Cortes, 2004; Hyland, 2008a; Huang, 2015). Therefore, LBs have been shown to differ markedly in their use from one study to another.

Paquot and Granger (2012) stated that since LBs differ in size (e.g., three to six words), and the operational methods used to identify LBs may vary from one study to another, it is not easy to make comparisons between studies. However, some common features can be recognised in the literature. For instance, some studies found that less proficient writers use more LBs than their expert writers' counterparts (Hyland, 2008b), but other studies also found that lower proficiency levels' learners tended to use fewer LBs than advanced learners (Chen and Baker, 2016). Other studies also claimed that speakers use more clausal bundles in conversation, consisting of verb phrase fragments (e.g., *am going to the*) and dependent clause fragments (e.g., *if they want*). In contrast, written language writers use more phrasal bundles, such as noun-based (e.g., *the importance of the*) or prepositional-based (e.g., *on the other hand*) (Biber et al., 1999; Hyland, 2008b; Qin, 2014). Therefore, LBs have been shown to differ markedly in their use from one study to another.

Previous studies of LBs can be categorised into three groups based on their perspective: first, the descriptive approach, examining LBs in specific registers (Karabacaka and Qinb, 2013). Second, a comparative approach, analysing LBs in L1 and L2 (e.g., Ädel and Erman, 2012; Liu, 2012; Pan et al., 2016); third, the psycholinguistic approach (e.g., Wray, 2002; Nekrasova, 2009). There are also several different approaches that have used in LBs analysis; for example, comparison by genre, register, proficiency level, age, and gender; or comparing their structural and functional distributions, or discourse markers. The findings from the broader literature on LBs usage in written academic discourse are discussed in detail in the following sections.

2.8.1 Lexical bundle variation in academic writing

An increased number of studies have investigated the use of LBs in academic writing. These studies have had different purposes and looked at LBs in different discourses, such as in different disciplines (e.g. Hyland, 2008a; Durrant, 2015); different registers (e.g. Biber, 2006b; Biber and Conrad, 1999); different academic genres (e.g., research articles, theses and dissertations) (e.g. Jalali and Zarei, 2016; Hyland, 2008a); different degrees of writing expertise, especially in ESL/EFL settings (e.g. Römer, 2009); and different populations (i.e., native versus non-native) (e.g. Römer, 2009; Jones et al., 2013; Qin, 2014; Chen and Baker, 2016; Shin, 2018). The results of these studies indicate the importance of these linguistic features in acquiring language. According to Hyland (2008b), LBs shape the meaning of the text and contribute to a sense of distinctiveness in a specific register. Thus, LBs can lead to fluent writing and be a sign of proficient writers in a specific discourse community (Cortes, 2004; Hyland, 2008a;

Hyland, 2008b). This section provides an overview of the variation of LBs across different writing genres.

One of the earliest studies in this area focused on register variation of LBs usage. Biber et al. (1999) compiled a corpus of over 40 million words of American and British English texts from conversation, fiction, newspapers, academic prose, non-conversational speech, and general prose to provide a picture of the wide communication system of the English language. The study revealed that the frequency and form of LBs varied between the conversation and written registers, where the former contained more such prefabricated sequences than the latter. Further research by Biber et al. (2004) reported that twice as many LBs are used in university lectures as are used in conversation, and four times as many LBs as in textbooks. More recently, Gray and Biber (2013) used a corpus-driven approach to identify four-word LBs in corpora of conversation and academic writing; the sub-corpora used in the study were taken from the Longman Spoken and Written Corpora. The authors confirm previous results that there were more four-word bundles in conversation than in academic writing; in contrast, there were more different bundle types in academic writing than in conversations.

LBs can also vary across a discipline. For instance**,** Cortes (2004) compared the written productions of university students who were native English speakers with published journal articles. The corpus of over 2 million words was drawn from two main disciplines: history and biology. The study revealed that students rarely used the LBs identified in the corpus of published writing. She also found that research articles in biology, one of the 'hard' fields, employed bundles much more than articles in history, a 'soft' field. Hyland (2008b) examined the forms, structures, and functions of four-word LBs in a 3.5-million-word corpus consisting of three sub-corpora of doctoral dissertations, master's theses, and published research articles in four disciplines (biology, electrical engineering, applied linguistics, and business studies). In this study, the cut-off point used to identify LBs was a minimum frequency of 20 occurrences per million words and an occurrence in at least 10% of texts. Hyland (2008b) found that only 10% of the top 50 four-word bundles occurred across all four disciplines.; 10 of the top 50 LBs occurred across three disciplines.

In contrast, the electrical engineering texts contained the greatest range of bundles, accounting for the rest of the top 50 bundles and more than half of the list. Biology, on the other hand, had the lowest proportion of LBs. According to Hyland (2008b), the greater reliance on LBs by an engineer could be a result of the relatively abstract and technical nature of the communication. This study revealed that authors from different disciplines used LBs differently to improve their arguments, convince the readers, and establish their credibility, resulting in a higher proportion of LBs being used. Following this study, Alipour and Zarea (2013) identified three- and four-word LBs to determine the differences in research articles from three disciplines (physics, computer engineering, and applied linguistics) and non-native corpus of applied linguistics research articles written in English by Iranian authors. The result was similar to that of Hyland (2008b), indicating that LBs were used in engineering disciplines more frequently than in other disciplines. However, the three disciplines were different in the use of LBs. For example, the bundles *due to the* was used more frequently in the physics texts than in other disciplines, and the bundle *some of the* occurred more commonly in disciplines other than physics. Interestingly, while the three-word bundles were identified in the engineering disciplines more frequently than in the other two disciplines, the four-word LBs were more frequently used in physics than in the other disciplines. On the other hand, computer engineering texts had the lowest proportion of four-word LBs. This is confirmation that there is disciplinary variation, as the three disciplines used LBs in different ways, demonstrating that LBs are genre-specific rather than discipline-specific. A recent study by Reppen and Olson (2020) examined disciplinary variations in LBs across nine disciplines (architecture, business, culinary science, digital arts, fashion design, film, hospital industry, interior design, and studio arts). Although the study found that LBs were common within and across the academic disciplines, 84% of the identified bundles were found in only one or two of the nine disciplines; only nine of 776 bundles were shared across all nine disciplines.

The other source of variation in the use of LBs examined so far is L1 versus L2 academic writing. The distinction between L1 and L2 English writers has been reported to exist in a wide range of academic texts of university students' writing (e.g., Chen and Baker, 2010; Salazar and Joy, 2011; Dontcheva-Navratilova, 2012a; Salazar, 2014; Pan et al., 2016; Bychkovska and Lee, 2017; Shin, 2019). Several studies have found that non-native learners' use of LBs is often problematic (e.g., Li and Schmitt, 2009). Although some studies have shown that non-native learners can produce formulaic sequences as native speakers (e.g., Nesselhauf, 2005), there is evidence that the limited repertoire of LBs they know leads them to overuse some bundles (Granger, 1998a). For instance, Chen and Baker (2010) examined the LBs used by Chinese non-native speakers and native speakers using the frequency-driven approach. The study used L1 and L2 learners’ sub-corpora from the BAWE corpus, and one corpus of L1 published academic texts from Freiburg-LOB Corpus of British English ('FLOB')(Maire, 1999). The study found fundamental differences in LBs usage between the three groups. For example, L2 learners overused some bundles rarely used in academic writing; the most frequent LBs in both L1 sub-corpora were rarely or never used by L2 learners. For example, L1 learners in both published and student writing adopted a wide repertoire of hedges, whereas L2 learners used a narrow set of hedges. Interestingly, native speakers used more diverse LBs than L2 learners, and the range of LBs and the number of tokens used increased with advanced writing proficiency. However, the authors indicated that the results might be affected by the corpus size, as larger corpora tend to show fewer bundles. For this reason, their study findings contradict other studies (e.g., Cock, 2000; Hyland, 2008b). Regarding the question of to what extent the increased use of LBs correlates with learners’ proficiency level, Chen and Baker (2010) failed to determine whether the frequency of LBs was associated with proficiency level. This idea requires further research, providing further motivation for this present study.

In a comparable study focusing on students’ writing, Ädel and Erman (2012) investigated the use of LBs by advanced L1 speakers of Swedish and native English speakers, all undergraduate students of linguistics. The study found a significant difference in the type/token measures between L1 speakers of Swedish and native speakers. The result confirmed the general pattern found by Chen and Baker (2010) that non-native writers produce fewer and less varied LBs. For example, unintended ‘this’ (*this can be seen*), existential ‘there’ (*there is no evidence*), passive voice and hedges bundles (Ädel and Erman, 2012). Römer (2009) investigated LBs in native speakers, non-native speakers of English, and expert academic writing (act as a reference corpus) and reported results that conflict with Chen and Baker (2010) and Ädel and Erman (2012). Römer (2009) explored whether, with respect to the use of LBs in written academic English, nativeness is an issue. The study examined the LBs used by native and non-native speakers in Apprentice Academic Writing (AAW) in the disciplines of linguistics and English (language and literature) to answer the question of whether nativeness has an effect on language proficiency in a controlled environment. Apprentice academic writing samples (AAW) are defined as “unpublished pieces of writing that have been written in educational or training settings” (Scott and Tribble, 2006, p.166). The study compared a list of 280 frequent discontinuous lexical sequences, p-frames and n-grams (of different lengths) derived from three corpora (native AAW, non-native AAW, and expert academic writing) to determine whether nativeness and expertise had an impact on language patterning. The first corpus used in Römer’s study was the Cologne-Hanover Advanced Learner Corpus (CHALC), a subset of 45 essays and papers written by upper-level university students. The second corpus is the Michigan Corpus of Upper-level Student Papers (MICUSP_EL), comprised 191 English and linguistic papers. The third corpus was a collection of 30 published research articles from the field of linguistics taken from the Hyland Corpus (Hyland, 1998), comprising 210,000 words. The first and second corpora are similar in terms of discipline, students’ level, and text type, and they only differ in terms of student writers’ native-speaker status. On the other hand, the Hyland _ Ling corpus differs from the other two corpora in terms of the academic writing expertise level and the number of years the authors had spent in academia. All three corpora together totalled 610,000 words. When examining advanced students, the results provided insight beyond the difference between native and non-native, indicating that experience or expertise is more important than nativeness. The study found that both native and non-native speakers developed their academic writing competence in similar ways and very few differences were found between the groups in their use of LBs. However, the lack of expert academic English phraseological items in both groups’ texts indicates that they may need to acquire academic writing conventions. The L2 versus expert comparison is more important than the L1 versus L2 analysis for understanding language development in regard to the use of LBs.

In a similar study, Pan et al. (2016) compared English telecommunications research articles written by L1 English and L1 Chinese professionals. The study used a high-frequency threshold of 40 occurrences per million words and a dispersion threshold of five texts in the native corpus and ten texts in the non-native corpus (since the L2 corpus was twice as large as the L1 corpus). They found that non-native articles included different structures to L1 articles. In addition, non-native writers used a greater number of lexical bundle types and tokens than the native writers did. There were 55 types of four-word LBs in L1 texts, totalling 1,845 tokens; by contrast, 71 types of four-word bundles were identified in L2 texts, totalling 26,489 tokens. The results support Römer (2009) argument that nativeness affects the use of lexical bundle items in written academic English.

Prior LBs studies also confirmed the view that academic writing is complex. Indeed, it becomes increasingly complex and requires greater effort as learners advance in their studies, and even novice native English writers have some difficulties using LBs appropriately (Ortega, 2003; Pan et al., 2016; Gray and Biber, 2013). Various studies have compared university students' writing versus published academic writing to identify differences in the use of LBs between novice experts and writers (e.g., Cortes, 2002; Cortes, 2004; Hyland, 2008a; Tribble, 2011; Jalali and Zarei, 2016). For example, Cortes (2004) compared LBs in the writing of published authors and university students at three levels: lower undergraduate level, undergraduate upper level, and graduate level. The findings suggested that, in general, LBs are used much more frequently by published authors than by university students. It was surprising that undergraduate and graduate students rarely used the target bundles. The author assumed that university students might have used alternative expressions rather than using LBs to serve the same function, to avoid errors and because they felt more confident using those expressions (Cortes, 2004). For example, using *alternatively* rather than the *on the other hand*; or, *because* instead of *can be explained by*. The result might be connected to what Hasselgren (1994, p.237) described as 'lexical teddy bears', as learners tend to rely on high-frequency occurrence, and on words with which they are familiar to express ideas and to avoid grammatical mistakes.

Hyland (2008a) also explored the form, structures, and functions of four-word clusters in three electronic corpora – master's theses, PhD dissertations, and research articles – to show the importance of clusters in differentiating between genres. The study found that research articles contained 71 different clusters occurred more than 20 per million, while the master's theses corpus contained the highest number of clusters, with 149 different clusters, followed by 95 different clusters in the PhD dissertations. The study revealed that master's level writers used a wide range of clusters with much greater frequency than PhD writers, who in turn exceeded expert writers. Moreover, master's thesis writers and PhD writers commonly used bundles not found in professional academic papers or appeared far less frequently. Hyland, therefore, suggests that these apprentice writers rely more on clusters in developing their arguments compared with the advanced learners. This finding is similar to that of Cortes (2002), who found that most of the LBs identified in the published author texts were far less frequently or never used by students. She also found that student writers used different LBs than experts did, arguing that students seemed to be familiar with a small number of bundles and used them frequently.

In the same vein, Wei and Lei (2011) examined the LBs in a corpus of 20 doctoral dissertations in applied linguistics and a corpus of journal articles published by professional writers. A total of 154 bundles were identified in the Chinese learners' corpus, totalling 7,548 individual bundles, whereas only 87 bundles were identified in the professional writers' corpus, totalling 4,245 individual bundles. This result is similar to that of Hyland (2008a), where the advanced Chinese ESL learners used more and a wider range of LBs than professional writers did. Another example comes from a study conducted by Jalali and Zarei (2016), who examined the use of four-word LBs in three corpora of doctoral theses, master's dissertations, and research articles. Four-word bundles had to occur at least 25 times and in five different texts to be counted as a bundle. In contrast to Hyland (2008a);Cortes (2002), the study results did not show any differences between the three genres in LBs' usage. Jalali and Zarei (2016) argue that the variations in the use of LBs in one discipline within different genres are more than those found in one genre of different disciplines. However, since reviewed studies did not reach to an agreement on whether LBs are more frequently found in professional texts than in text written by university students, this issue requires further investigation.

One of the common claims of previous LBs research is that these expressions are not easy to master through simple exposure to spoken or written discourse (Cortes, 2004; Biber and Barbieri, 2007; Hernández, 2013; Karabacaka and Qinb, 2013). Biber and Barbieri (2007) propose that further research is needed to examine the extent to which LBs are naturally acquired without exposure to their usage in context. However, several LBs research argued that LBs need to be overtly taught, and learners should be exposed to the use of more LBs in all environments (e.g., Cortes, 2004; Biber and Barbieri, 2007; Hernández, 2013; Karabacaka and Qinb, 2013). Appel and Wood (2016) claimed that providing less proficient writers with materials related to the test topic helps L2 learners supplement their limited linguistic resources to cope with their difficulties using formulaic sequences. In this study, two reading articles related to the test topic were given to the L2 learners to help them become familiar with the subject matter and were used as reference materials during the test. The results revealed that low-L2 learners exhibited greater use of longer LBs (four and five words) than medium- and high-L2 learners. More of these identified bundles were also found in the reading articles included as part of the test materials. Karabacaka and Qinb (2013) compared the use of LBs in the argumentative writing of three groups of university writers: Turkish, Chinese, and American. The study claimed that even advanced non-native learners had difficulty acquiring some LBs through simple exposure, suggesting that explicit teaching may be needed to improve the learning process.

Cortes (2006) focused on teaching LBs to university students in a writing-intensive history class, conducting pre-and post-instruction analyses of teaching the target bundles to the students in the history class. The study claimed that the systematic exposure to these target bundles was not sufficient to make a significant difference in the use of LBs, and suggested the need to conduct "a longitudinal study of the same group of students at different stages in their careers in the discipline to investigate their formulaic language development." These reviewed studies confirmed that LBs are not acquired naturally, and even simple exposure to these expressions did not increase the production of LBs. Therefore, LBs are far from being simple expressions, as they have been shown to be rarely used by students.

As for the structural correlations of LBs, the findings reported in the literature are also varied. Biber et al. (2004) compared LBs across spoken and written registers: conversations, classroom teaching, textbooks, and academic prose. They found that spoken registers differ in bundle structures and functions from written texts. They found that conversations and classroom teaching are composed of mainly verb phrase bundles while noun-phrase and preposition-phrase based bundles are preferred in textbooks and academic prose.

Focusing specifically on academic writing, Ucar (2017) compared published articles of non-native and native English speakers. They reported that native writing included slightly more *NP with of-phrase*, while non-native academic writing used more *PP* than native English texts. Similarly, Bychkovska and Lee (2017) explored LBs in L1-English and L1-Chinese undergraduate students' argumentative essays. They found that native English writers relied more on NP-based bundles (e.g., *the* *nature of the*) while VP-based bundles (e.g*., it is possible*) were more common in non-native writing corpora, yet their writing was still more phrasal.

Turning to the use of LBs across proficiency levels, Juknevičienė (2009) compared the LBs produced by university learners of three different proficiency levels. Structurally, the analysis showed that lower proficiency writers tended to use more VP-based bundles while the higher-level learners used more NP-based bundles in their academic writing. The result was similar to Chen and Baker (2016), who compared LBs across three proficiency levels: B1, B2 and C1. They found that lower-level learners share more features with conversation while the more proficient writers share more characteristics with academic prose. It can be seen from the analysis of bundles' structures of the reviewed studies, professional and native writers' writing are mainly full of grammatical patterns associated with formal and academic language, while novice writers relied more on informal conversational language. These findings are useful as they may reflect the whole picture of L2 writers' discourse features.

In addition to their structural taxonomy, LBs have been examined according to their function correlations in discourse. Biber et al. (2004) comparison of LBs across spoken and written registers revealed that conversations mainly rely on stance bundles (e.g*., I want to, I think that*), but textbooks and academic text comprised a greater number of referential bundles (e.g., *the importance of the, in the case of*). As Conrad and Biber (2005) clarify, such differences occur because academic texts focus on presenting primarily factual information while spoken language highlights interpersonal interactions.

Focusing specifically on academic writing, Hyland (2008a) examined clusters in corpora of research articles, doctoral dissertations, and master's theses from four disciplines (electrical engineering, business studies, applied linguistics, and microbiology). The study found great variations in the use of LBs across the sub-corpora. While the hard science (electrical engineering and microbiology) writing mainly composed of research-oriented bundles (corresponding to 'referential' bundles in Biber' framework) (e.g., *There are a lot*), soft science corpora (applied linguistics and business) were dominated by text-oriented bundles (corresponding to 'stance expression' in Biber' framework (e.g., *I would like to*). This finding is in line with Durrant (2015) and Omidian et al. (2018), who reported that hard sciences comprise more research-oriented bundles, whereas soft sciences comprise more ‘text-oriented and ‘participant oriented bundles.

Among the studies focusing on different degrees of writing expertise, especially in EFL settings, Chen and Baker (2010) reported that expert writers use more ‘referential’ bundles than L1 and L2 students. In contrast, a wider variety of ‘discourse organisers’ bundles (e.g., *first of all*) are used by L1 and L2 students to elaborate or clarify a topic than by English L1 experts. Another study by Nkemleke (2012) explored the functions of LBs in 150 non-native postgraduate dissertations written over a period of 3 years (2007–2009). The results indicated that postgraduate students were aware of a broad range of LBs, and their functional distribution across texts appeared to be skewed in most cases. In addition, the study reported that postgraduate students’ writing was dominated by research-oriented LBs and fewer text-oriented bundles. This finding is in line with Hyland (2008a), who found that master’s students’ discourse was characterised by heavy use of research-oriented clusters, while clusters in doctoral dissertations and research articles were more likely to be participant-oriented or text-oriented. The results of the previous research, therefore, indicate that LBs in academic writing are full of research-oriented bundles, though there are some variations between language users.

A common issue that appeared when classifying LBs is the difficulties of determining their categories, due to the multifunctionality that some bundles have in a single context. Previous researchers have encountered this issue when assigning LBs to functional categories (Cooper, 2016; Leedham, 2011; Pecorari, 2009). In general, the process of classifying LBs according to their functional distribution is subjective, and it is important to look at the concordance lines to see the bundles in context and address the target bundles ‘multi-functionality’. For example, Pecorari (2009) addressed the multifunctionality issue and the importance of looking at the concordance lines to see the bundle in context if it has more than one function, and stated that “if 50% of given bundle tokens were related to a given function, the bundle was assigned to that category”. Furthermore, Byrd and Coxhead (2010) gave examples of multifunctional LBs, such as the bundle *at the end of,* which can be used as either a time/location or structuring signal. In order to classify bundles into the correct sub-categories, the present study followed Pecorari (2009) method of addressing the issue of multi-functionality in LBs identified in the ESL learners’ sub-corpora. The exact procedure for functional classification applied in this study will be described in section4.8.

Additionally, previous LBs research has found a strong relationship between structural and functional distributions. For instance, Biber et al. (2004) found that most stance bundles were linked to verb-based bundles, whereas referential expressions were dominated by noun-based or preposition-based bundles. These results were also supported by several previous research such as Chen (2008) and Hyland (2008b), who examined the use of LBs in academic writing.

Recently, Beng and Keong (2017) examined the structural and functional types of LBs using a specialised corpus from the Malaysian University English Test. The results also showed a strong connection between noun-based and research-oriented bundles, and dependent clauses were strongly linked to text-oriented function. In contrast, verb-based bundles usually performed a participant-oriented function. From these findings, it can be seen a strong relationship between only noun-based bundles and research-oriented bundles. At the same time, the significant association of other structures and functions categories remained uncertain, since they were associated with various categories by language users.

These considerable variations of LBs in frequency, structures, and functions across registers, disciplines, and genres, as well as the different use of LBs by L1, L2 and professional writers found in the reviewed studies, underscore the pedagogical value of studies that examine the use of LBs in specific genres of academic writing. Moreover, such studies can be used as practical models for novice writers because, as proposed by Hyland "Bundles occur and behave in dissimilar ways in different disciplinary environments and it is important that EAP course designers recognise this, with the most appropriate starting point for instruction being the student's specific target context" (Hyland, 2008a, p.20). The current study and the ones reviewed in the following sections may contribute to such work.

2.8.2 The use of lexical bundles across proficiency levels

A number of studies have examined the use of LBs across L2 learner proficiency levels to identify noteworthy differences in their use. However, as proficiency information is not always readily available, the available research has resorted to level of the study, (e.g., university levels), used learners' responses in high-stakes exams (e.g., IELTS), or using a common standard for determining learner proficiency (e.g., CEFR levels) on data retrieved from an available learner's corpora (e.g., Chen, 2009; Staples et al., 2013; Qin, 2014; Appel and Wood, 2016; Li and Volkov, 2017; Ruan, 2017).

Yet, the findings obtained from these studies have presented mixed and uncertain results. For example, Qin (2014) investigated five-word LBs in different university graduate student levels (first-year MA programme, second-year MA programme, first and second-year PhD programme, beyond second-year PhD programme, and expert level). One interesting finding was that the more advanced L2 learner writers were, the more their use of LBs increased. Although the frequency of LBs by second-year L2 PhD (level four) was higher than by expert writers, the study showed a steady increase in the use of lexical bundle types across the levels. As suggested by Qin (2014), the increased use of LBs in second year L2 PhD students compared to expert writers might be due to the overuse of certain bundles in their writing.

In another study, Ruan (2017) examined the use of four-word LBs in Chinese undergraduate written assignments, using a total of 777 texts collected from students across different years of study at the university. The study found a pattern in the use of LBs: as students' progress to higher levels, they can use a wider range of LBs in their academic writing. However, this finding needs further investigation as the study did not consider proficiency levels when building the corpora. This issue has been addressed in the present study, as I used the CEFR levels to differentiate between the sub-corpora in order to identify the similarities and differences between the levels.

Contrary, Staples et al. (2013) investigated idiomaticity using LBs in written responses across three proficiency levels in the Test of English as a Foreign Language (TOEFL iBT) in a controlled environment. The study was concerned with the use of LBs in the area of language testing research, and found similarity in LBs usage in terms of function across proficiency levels, though the less proficient learners used a greater amount of bundles overall. This result supports Römer (2009) argument that proficiency level is an important determinant of learners' use of LBs and that even experts and native speakers may need help to improve their academic writing competence. A recent study by (Appel and Wood, 2016) also explored the use of three-, four-, and five-word LBs in academic essays written by low-, medium-, and high-L2 writers collected from Canadian Academic English Language (CAEL) assessment. Two reading articles related to the test topic were given to the L2 learners to help them become familiar with the subject matter, and were used as reference materials during the test. The results revealed that low-L2 learners used more longer LBs (four and five words) than medium- and high-L2 learners. However, it appears that many of these bundles were related to the source materials in the test, with a narrower range than the more proficient writers. Appel and Wood (2016) concluded that providing less proficient writers with materials related to the test topic helps L2 learners supplement their limited linguistic resources to cope with their difficulties using formulaic sequences.

Similarly, Vo (2016) used an English Placement Test corpus to examine the use of vocabulary distribution and LBs in three non-native English students' writing, taken from three different levels. The study focused on the use of four-word LBs using a frequency threshold of 10 times in 50,000 words in the sub-corpus and used in at least five texts. The frequency analysis of LBs revealed that high-proficiency writers used fewer LBs than lower-level writers.

In contrast, Li and Volkov (2017) corpus-based study focused on LBs used by test-takers of different English proficiency levels, and found that higher proficiency level writers used more varied LBs than lower proficiency level writers.

Although the use of test-taker and university students level data have yielded valuable insights into how LBs are used by English users of different proficiency levels, the absence of using a common standard for differentiating between the learners' level still makes the results uncertain and cannot be generalized across L2 learners' research.

A recent study by Chen and Baker (2016) examined criterial discourse features through the use of LBs in a corpus of graded essays of different CEFR levels (B1, B2, and C1) written by Chinese learners of English retrieved from the Longman Learner Corpus. The study revealed that lower proficiency writers used more LBs, whereas the bundles employed by higher proficiency writers were more academic in style. In addition, the authors found that EFL learners exhibited different structures and functional patterns. For example, B1 level writing appeared to have the highest proportion of VP-based bundles, whereas C1 showed the highest combined proportion of NP- and PP-based bundles. However, when examining the functional distribution, the results showed a very similar distribution of bundle functions across CEFR levels, with an increase of discourse organiser and referential expression bundles across the level. The results were in line with Staples et al. (2013), where L2 learners exhibited more discourse organiser and referential expression bundles in their writing.

In another study, Vo (2016) examined the use of LBs English Placement Test corpus of three non-native English students' writing, taken from three different levels. The study found that the three groups produced more NP-based and PP-based bundles than VP-related bundles. This finding is contrary to Chen and Baker (2016), as the three levels were used more referential, followed by stance bundles. In a recent study, Ruan (2017) examined the use of four words LBs in Chinese university students' academic writing across different levels of studies. The study found that Year 1 learners relied heavily on NP-based and VP-based bundles in their writing, while Year 2 and Year 4 learners used more PP-based bundles. Functionally, the results were similar to Chen and Baker (2016) that L2 learners from different proficiency levels used more discourse organisers and referential expression bundles in their writing. However, looking closely at the differences between the levels revealed that Year 1 learners used more referential expressions and stance expressions than other students, while Year 2 and Year 4 learners used more discourse organisers than Year 1 students.

In summary, despite the small number of studies examining the use of LBs across learners' levels, the available studies revealed significant differences in the use of LBs in L2 writing across proficiency levels. Furthermore, while it seems that a strong set of LBs would likely improve academic scores and learner confidence, the challenge these items pose needs a learning strategy. Finally, these results suggest a complex picture that offers sometimes conflicting account of whether LBs differ between writing at different proficiency levels. In the light of these findings, further research is necessary to provide evidence of the relationship between the use of LBs L2 learners proficiency levels.

2.8.3 Review of longitudinal studies on lexical bundles in academic writing

Despite the abundance of studies on LBs in academic writing, a small number of studies have directed their attention to the developmental pattern of the usage of LBs usage over time (e.g., Cooper, 2016; Candarli, 2020).

In one of the latest studies, Cooper (2016) examined the interaction of LBs to academic performance over time. The study examined the use of four-word LBs in a corpus comprising assignments written by first-year psychology over a period of three years. The results detect an interaction between the density of LBs and the academic performance among undergraduate students, as high-level learners found to use more of the target bundles than low-level. The researcher suggested that the increased use of LBs across learners' levels might be due to a consequence of the difference in writing style across the academic year. The findings were more optimistic, attesting to a considerable change in the learners' production of LBs and to a high degree of variability among learners over time. However, the study did not detect any significant differences in the use of structural and functional types by various student groups according to the academic year. The findings thus remain somewhat unclear, due to the confounding effects of text differences, which may affect the usage of LBs.

Recently, Candarli (2020) examined the changes of LBs, in terms of structural categories and discourse functions in 98 advanced second language (L2) learners during their first-year at an English medium university in a non-English-speaking country. The findings revealed changes in the frequency of different functional and structural categories of LBs over time. However, the study uncovered only significant increased use in noun-based and referential expressions bundles in the L2 writers' essays at Month 9. Candarli (2020) suggests that the distribution of different categories of LBs in the L2 writers' essays aligned more closely with the distributional characteristics of LBs in English academic prose, since Biber et al. (2004) found that academic writing in English relies on noun based and preposition-based bundles, the majority of which serve as referential expressions.

While the above research so far has suggested development and changes for some aspects of LBs use by L2 writers, it is imperative to deepen our understanding of the patterns of development from novice to more mature writing through a longitudinal study in a common L2 academic writing context.

2.8.4 Discussion on reviewed studies

As reviewed in this chapter, a number of studies have investigated the use of LBs in academic writing. These studies have had different purposes and examined LBs in differing discourses of writing such as different disciplines (e.g., Reppen and Olson, 2020), registers (e.g., Biber et al., 1999), academic genres (e.g. research articles, theses and dissertations) (e.g., Hyland, 2008a), degrees of writing expertise, especially in

ESL/EFL settings (e.g., Römer, 2009), and populations (i.e. native versus non-native) (e.g., Wei and Lei, 2011). One of the key findings associated with LBs, is that developing language competence in academic writing is considerably affected by their utilisation (Biber et al., 1999; Biber and Barbieri, 2007; Hyland, 2008b; Ruan, 2017). General interpretations have indicated that LBs are of significance to native and non-native English writers due to the heavy reliance on these expressions in academic writing. However, becoming competent in the usage of LBs is complex, for both L1 and L2 writers, such that mastery of the use of LBs distinguishes expert writers from novice writers, regardless of their L1 (Cortes, 2002, 2004, Hyland, 2008a, Pang, 2009). This is demonstrated by novices' lack of knowledge of LBs compared to their expert counterparts. For this reason, one of the distinctive features of novice learners' writing is the lack of inclusion, or inappropriate use of LBs.

Furthermore, despite the general agreement that there are differences in the use of bundles by novice and advanced L2 learners, the precise nature of these differences remains unclear. While some studies have found that less proficient learners use more LBs in their writing than proficient learners (e.g., Staples et al., 2013), others show an increased use of LBs in highly proficient learners' academic writing (e.g., Ruan, 2017).

Although these studies have enriched our knowledge about the use of LBs, little research has focused on the use of LBs among L2 learners with different levels of proficiency. Thus, it must be acknowledged that the quantity of literature related directly to L2 learners of different proficiency levels is limited and the studies available that focus on L2 learners' levels have mostly used university students' data (e.g., Vo, 2016), learners' responses in high-stakes exams (e.g., Staples et al., 2013), or re-sampled data from the published learner corpora (e.g., Chen and Baker, 2016). In general, these studies have yielded valuable insights into how language is used by English users at different proficiency levels. However, these studies' findings remain mixed, and the gaps distinguishing L2 learners from different proficiency levels' use of the various forms, structural, and functional patterns into which LBs are categorised remains unclear. This may be due to a difference in corpus sizes and methodologies, and texts in different academic registers, or due to the heterogeneity in corpus design, which might affect the use of LBs. More specifically, previous studies have discrepancies in determining learner proficiency levels which makes it difficult, if not impossible, to generalise across research results of this genre.

Although Chen and Baker (2016) used the CEFR for determining proficiency levels, which is arguably one of the most influential frameworks in language education nowadays, they only address argumentative essays restricted to L1 Chinese learners retrieved from the Longman Learner Corpus (LLC) published between 1990 and 2002. Therefore, further research is needed to explore such argumentative writing, including newer data of non-native writers from a variety of L1 backgrounds. This viewpoint is supported by Hyland and Jiang (2018), who reported a considerable change in the functional distribution of LBs in response over time. Therefore, choosing a newer corpus could shed light on the accuracy of learners' levels.

Additionally, to the best of my knowledge, approaches that have analysed the interactions of LBs of L2 writing across the proficiency levels over time have been relatively rare, the majority focusing on university students at English-medium universities (e.g., Cooper, 2016; Candarli, 2020). Therefore, little is known about the impact of a longer duration of L2 learning on the use of LBs in L2 writing.

Moreover, it should be noted that these studies focused on disciplinary academic writing and also did not employ a common standard framework for determining the learners' proficiency levels, therefore making generalisation about L2 writing not possible from the results of these studies.

With the above concerns in mind, this thesis is motivated by a lack of empirical research to analyse and compare LBs in ESL learners' argumentative essays of different proficiency levels. In other words, a need is felt to explore the variations and the development use of LBs across ESL learners of different proficiency levels, specifically among those who are joining English for Academic Purposes (EAP) courses. Therefore, the present study is intended to fill the above gaps in the literature. It first aims to examine how ESL learners of different proficiency levels use LBs in their writing, tracing the developmental use of LBs in argumentative essays across differing ESL learners' levels over time. It has been recommended elsewhere that efforts to trace any development in the use of LBs should take the form of a longitudinal study on a single group of learners, to provides information about changes to a set of research units during a period of time (Taris, 2000; Wang et al., 2017). Therefore, the second aim of this study is to explore the question of lexical development over time by conducting a longitudinal quantitative analysis of the LBs used by ESL learners from three different levels, B1, B2, and C1, over a period of six months of language study.

Based on the review studies in this chapter, the current study advances the existing research into the use of LBs in L2 academic writing in several ways: Firstly, the quantity of literature related directly to ESL learners of different proficiency levels is limited with few studies focusing on students on EAP courses. The context of this study was English for Academic Purposes courses in the UK, which typically aim to develop the confidence and skills students need to succeed in their academic studies; therefore, they are more likely representative of language learners, and their writing should clearly contribute to the existing data concerning the different use of LBs in academic writing. It is hoped the present study will provide an exciting opportunity to advance our knowledge of the use of LBs in ESL learners' academic writing. Secondly, the study traced the argumentative essays that assess ESL learners' writing performance. This made it possible to investigate to what extent the frequency of LBs contributes to ESL learners' levels, further exploring the relationship between the frequency of LBs in argumentative essays of ESL learners and their different levels of proficiency. Hence, the present study responds to the call to explore the relationship between the frequency use of LBs and proficiency levels (Chen and Baker, 2010) in learner corpus studies.

In addition, to the best of my knowledge, the only list that can fully reflect individual language proficiency is the *English Vocabulary Profile* (EVP) [1], which contains a considerable number of multi-word units (phrases, phrasal verbs, collocations and idioms), categorised according to the six levels of the CEFR scale. However, as LBs can be mixed up with other multi-word units, such as idioms and colocations, they may need to be dispersed in a separate list than other multi-words to be used more effectively (i.e., by researchers, teachers, etc.), and avoid confusion with other multi-word units. Therefore, using the (EVP) lists or any general corpus as a representative of the data may neither be sufficient for analysing LBs and nor capture ESL learners' argumentative writing performance.

In light of that, the current study employs carefully matched corpora of ESL writers, built on essays written in a single genre (argumentative essays), and in

[1] https://www.englishprofile.org/wordlists

response to CEFR levels, in order to compare the use of LBs more effectively by the three language level groups. Fourthly, although a large body of research has been conducted on LBs in academic writing, there is still little research exploring ESL learner variation and development in the use of bundles in English argumentative essays across different levels of proficiency. This strand of research is of great significance to the teaching of English academic writing to L2 learners, as it can reveal how ESL learners at different proficiency levels utilise LBs as a resource when composing their essays, thereby demonstrating their understanding of academic conventions.

2.9 Teaching Lexical bundles

Formulaic language such as LBs and collocations have proven to be 'pervasive in academic language use, a main element of fluent linguistic production, distinguishing novice and skilled use in both spoken and written registers' (Hyland, 2012, p.153). Coxhead and Byrd (2007, p.134-135) have proposed three primary reasons as to why formulaic language, in this case LBs, are important for students such as second-language learners (L2).

1. 'LBs are often repeated and thus 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. LBs become defining markers of fluent writing and are important for the development of the style of writing that fits the expectations of readers in the world of academia;
3. These bundles often lie on 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.'

The above reasons help clarify why LBs should be an explicit part of English language teaching, particularly in EAP, in order to ensure that students are linguistically prepared for the language they will no doubt encounter in their future academic careers. Hyland (2008a) insists that 'academic writing draws on a much larger stock of prefabricated phrases than either news or fiction' (p. 44).

Since LBs occur very frequently in language usage, it might be assumed that they can be acquired naturally and easily. However, as Biber and Barbieri (2007); Cortes (2006) point out, the acquisition and proper use of these expressions does not appear to happen naturally. Although professional writers make use of a wide variety of LBs to develop their arguments and persuade their readers, many LBs used by experts have rarely or never been used by students in different disciplines or genres and at different levels of expertise (Cortes, 2004; Hyland, 2008a; Chen and Baker, 2010). Therefore, one of the objectives of EAP teachers in writing classes is to encourage learners to use LBs to generate a discourse that is adequate for academic purposes. More to the point, as Hyland (2008a) believes, 'gaining control of a new language or register requires a sensitivity to expert users' preferences for certain sequences of words over others that might seem equally possible' (p. 5).

Many studies have suggested different ways of enabling students to use formulaic language. For instance, Lewis (2000) presents an edited volume with many innovative ways of teaching collocations. Pang (2010) also suggests different strategies and techniques that will enable L2 learners to expand their repertoire of academic rhetorical features to include these expressions.

In the field of EAP, early initiatives on teaching formulaic language are reported in (e.g., Jones and Haywood, 2004; Cortes, 2006; Li and Schmitt, 2009). The effectiveness of teaching LBs to improve English learners' writing skills has been confirmed by a number of previous studies (e.g., Kazemi et al., 2014; Alhassan and Wood, 2015; Shin and Kim, 2017). Jones and Haywood (2004) tracked a group of non-native speaker university students for their use of certain word combinations during a ten-week period. Although the study noted some minor progress in the production of formulaic language after the instruction, they reported high motivation and tendency towards the use of these expressions by the participants in their study. They also argued that students neglecting to use formulaic expressions in their academic writing may lead to a weak performance. Similarly, Li and Schmitt (2009) analysed written assignments of a Chinese MA student for the use of formulaic language over a period of ten months. The study found that the participant learned new formulas and their proper usages, particularly from academic reading materials during the course of the study, and also became more confident in using the expressions in her writing. They emphasized that learners could acquire formulaic language from exposure to a rich variety of sources, along with explicit instruction.

In another study, Cortes (2006) examined the teaching of LBs to a group of university students in a history class for five twenty-minute sessions. She noted an increase in students' awareness of and interest in these expressions, although the time was insufficient to make significant differences between pre- and post-instruction production of LBs. However, a recent study by Shin and Kim (2017) examined the potential for teaching the use of LBs with adult English language learners of varying proficiency levels. The data was collected over three weeks through pre/post tests in which participants wrote sentences using the core expressions. The results showed both low-and high-proficiency focus groups demonstrated significant improvement on the post tests. It could thus be argued that appropriate use of LBs can not only signify EFL learners' expertise, but also give a near-native inflection to their writings.

2.10 Conclusion

The chapter has illuminated a number of prominent issues that form the background of this study. Firstly, it presented a theoretical background detailing formulaic language, its characteristics, significance, and the prevalence of these expressions in academia. Secondly, it investigated the theories and research that form the background to this study of LBs, as a way to provide an indication of academic performance. The aim of this investigation has been to explore the main criteria in terms of which bundles are identified, and to provide explanations of how LBs varies from the use other formulaic language, particularly idioms and collocations. The chapter also looked at the keyword analysis concept and its role in linguistics analysis. Then, the CEFR descriptors and its relevance to language competence were explored, leading to an explanation of why they were selected for the analysis of this thesis. This chapter also points out a number of important features of the argumentative essay genre in the EAP classroom. Then, it reviewed the variations in LBs in academic writing, and the empirical findings reported, particularly those that relate directly to the focus of this thesis. This is followed by a review of the relevant research on LBs, and narrows down the research topic to the use of LBs among students of differing levels of English language proficiency. Finally, the final part of this chapter has dealt with the impact of teaching formulaic language and in particular LBs on academic writing.

The remainder of this thesis is divided into four chapters. Chapter 3 tests the methodological decisions used for analysing LBs on pilot data, before conducting the main analysis. Chapter 4 presents the methodological considerations in the analysis of the ESL learners' sub-corpora. Next, chapter 5 presents the results and discussion of the study data. Finally, chapter 6 presents the conclusion of this study.

3 Pilot Study

3.1 Introduction

This chapter aims to ensure the reliability and validity of the research results through conducting a pilot study. A pilot study, also known as a ‘feasibility study’, presents an opportunity to test the methods and procedures on a smaller scale, to establish their efficacy for a larger-scale study; this includes the research methodology and data analysis, as well as the research instruments, such as an observation context or corpus size. A pilot study is an important component of good study design as it can inform the researcher of the strength or weakness of the proposed study (Barnbrook, 1996). It also allows the identification of problems and shortcomings identified in the research design – for example, determining whether the number of participants is sufficient for answering the research questions; whether the data is promising; and whether the use of certain equipment will be feasible – so that they can be addressed and avoided in the main study. A pilot study might also provide the researchers with ideas and approaches that will enable them to yield clearer findings. It starts as informal experiments, where a series of steps are tried out on a handful of participants and lead to the successful completion of the main study.

Therefore, conducting a pilot study was important to this research for several reasons. Primarily, it helped test the adequacy of the research instruments and the soundness of the methods used to answer the research questions. In addition, building a corpus requires careful decision-making in terms of corpus type, representativeness, sample type, and sample size. Thus, the pilot study reduced the number of challenges related to data collection strategies for the initial construction of the ESL learners’ corpora, as it made it possible to redesign parts of the main study to overcome the shortcomings revealed in the pilot study. According to Biber (1993), an important consideration in building a corpus is the overall design. For example, the essays included in the corpus, number of essays, and length of the writing samples, each of which contributes to achieving ‘representativeness’. Therefore, being familiar with this series of steps helped to plan the larger study and improved the researcher’s chances of success in the main study.

After highlighting the aims of the chapter, section 3.2 discusses the decisions that were taken to achieve ‘representativeness’ in terms of the research population, students' levels, number of participants, and essays' rating so that they could be assigned to an appropriate level of learner corpora. Section 3.3 explains the data collection process followed for the ESL learners' sub-corpora, including the participants and the corpora used in this study. Section 3.4 introduces the longitudinal study conducted for this research, followed by a brief description of the reference corpus (RC) used for the analysis (Section 3.5). Section 3.6 explains the research questions followed by the analytical methodology used to analyse the data (Section 3.7). Lastly, section 3.8 presents the result and discussion of the pilot study.

In order to reduce repetition, only a brief description of the methodology of the pilot study will be given in this chapter; an in-depth discussion will be presented in the main study.

3.2 Pilot study corpora and research population

Two sub-corpora of written essay data from two different levels of learner, B2 and C1, were compiled. The learner corpus is a form of specialised corpora and has been the subject of much research (Granger, 2002; Gilquin et al., 2007). By compiling written samples, such as ESL learners' essays, researchers can generate a fruitful dataset on which to conduct various analyses. The decision to use learners' sub-corpora in the present study was based on their usefulness in exploring and identifying the similarities and differences in the use of recurrent word combinations across L2 proficiencies of "actual language in use" (Adolphs, 2006, p.97). A learner corpus composed of data comprises written or spoken data, or both, that has been produced by learners in the process of acquiring a second or foreign language (Mcenery and Xiao, 2011). Using a learner corpus is also useful to better understand second or foreign language learning approaches in corpus linguistics to reveal how ESL learners acquire the language. In this pilot study, learner sub-corpora were used to test the analytical procedures to predict an appropriate sample size and improve upon the study design prior to undertaking a full-scale research project.

For the purposes of this thesis, the data sets were drawn from ESL learners who had studied at the English Language Centres (ELCs) in the UK, mainly incorporating essays and examination transcripts contributed by the ELCs. The intention was that if the data collected for the pilot study proved helpful, it would be incorporated into the resulting B2 and C1 sub-corpora.

For a sub-corpus to be representative, the data should be controlled and match the research purpose. As the data was collected from two ELCs, this helped control the participants' age and nationality in the study, as students must be over 18 to enter the ELCs and are accepted from different nationalities. The written essays of students at two different levels, intermediate and advanced, were argumentative essays equivalent to the IELTS task 2 in terms of essays' titles.

The following decision concerned the Common European Framework of Reference for Languages (CEFR) levels to be analysed in the pilot study. The decision was made to collect essays written by ESL learners at two different proficiency levels, B2 and C1, for the sub-corpora to test the methodology and research design and determine their appropriateness for use in the full study. With regard to the relationship between the use of LBs and academic performance, language learners at these two levels are able to use the target language in academic and professional situations; thus, it was assumed that they would create well-structured and detailed texts, more so than students at other levels. As this pilot study also considers the overuse and underuse of LBs in the comparison between groups of different proficiency levels, B2 and C1 levels were assumed to be able to provide more words and LBs than other levels. Therefore, they would also provide a practical comparison between the levels to answer the research questions. Additionally, as the sub-corpora used in the pilot study were drawn from written argumentative essays, the number of words in the collected essays were quite similar between the two levels.

The third step was to decide the number of participants and essays collected for the corpus. As a rule of thumb, the sample in a pilot study should be representative of the target study population, based on the same inclusion/exclusion criteria as the main study (Thabane et al., 2010). However, Biber (1993) states that a thorough definition of the target population and decisions concerning the method of sampling are more important considerations than the sample size, as the sample size is not as important as sample representativeness. Therefore, the essays collected for the pilot study were compiled from participants writing academic essays to test their progress and be placed in higher levels at the ELCs if they meet the requirements. In total, 130 essays were collected over four months. One reason for collecting data from the ELCs was to enable control of age, cultural background, proficiency level and ensure that all essays meet the study requirements (See section 3.3). Due to the time limit and difficulty of collecting students essays from more different language centres, a small number of essays in each corpus was expected, as Baker (2010) states that a specialised corpus is easier to collect but can be small in size and restricted by different factors, such as genre, time, and place. The last step to meet the purpose of the study determination of the proficiency level. As the pilot study focused on the B2 and C1 levels, some of the essays allocated to different levels, such as A1, A2, B1 and C2, were excluded.

In accordance with the aim of this thesis, the pilot study was divided into two stages. In the first stage, the B2 and C1 sub-corpora of ESL learners were compared (in terms of frequency, structures and functions of LBs and keyness) to provide an overview of some of the linguistic features used so as to differentiate between the levels. The expectation was that such a comparison would enable the researcher to identify any relationship between the use of LBs and academic performance.

In the second stage, the study comes under second language development research, which compares learner language across proficiency levels. The longitudinal study investigated the development over three months of two ESL learners in terms of their use of LBs in academic essays across the different levels, to provide a picture of the increases in proficiency level.

3.3 Data collection

The first step in compiling the B2 and C1 pilot sub-corpora followed University of Liverpool ethical procedures, as detailed in the University Policy on Research Ethics (see section 4.5.1). After ethical approval had been given, the agreed procedure with ELCs teachers was to collect the essays after the placement test. The data were collected from two different students' levels (intermediate and advanced) and the essays were equivalent to the IELTS task2 in terms of the essays' title which were used to allocate students to the appropriate level of study in the ELCs. These texts were chosen because they represented different research and methodological aspects, thus highlighting some of the diversity present among ESL learners in the UK.

These essays were compiled and then rated following the manual for *Relating Language Examinations to the Common European Framework of Reference for Languages* (Europe, 2003).

Three teachers who were teaching IELTS preparation were trained to rate the essays using a Writing Assessment Scale developed by CEFR, as shown in Table 3.1. Those teachers were selected to re-rate the essays because of their experience in teaching IELTS preparation or as IELTS examiners. To avoid duplication, procedure for rating the essays will be explained in detail in section 4.5.5.2.

Table 3.1. Raters profiles.

Model	**Rater 1**	**Rater 2**	**Rater 3**
First Language	English	English	English
Current work	Teaching English of academic purposes+ IELTS preparations	Teaching English of academic purposes+ IELTS preparations	Teaching English of academic purposes+ IELTS preparations
Familiarity with CEFR	Excellent	Excellent	Excellent
Qualification	MA TESOL	MA TESOL	MA TESOL
Teaching Experience	5	9	7
IELTS Teaching Experience	3	5	2
Experience using the CEFR scales in marking written English	Yes	Yes	Yes

After the determination criteria, the essays were incorporated into four learners' sub-corpora representing CEFR levels A2, B1, B2, and C1 according to their rating. However, A1and A2 rated essays were excluded from the analysis as they are insufficient samples in this pilot study.

In order to build the B2 and C1 sub-corpora, student essays were retyped, and irrelevant information was cleaned (e.g., titles). The essays were then converted to txt. format to be analysed using *WST* to determine the frequency of use of LBs and keyness in academic writing. Finally, all the text files were renamed and compared manually against the actual student samples to ensure accuracy.

3.3.1 Participants of the study (cross-sectional)

In total, 42 ESL learners, contributing 130 essays, participated in the research to examine the use of LBs in their academic writing. The authors of the essays were intermediate and advanced English learners; each had been studying English for at least two months in the UK. Their ages ranged from 18 to 40 years, and they came from different L1 language backgrounds, although the majority (69%) of the participants were Arab or Chinese. Moreover, most of the students were female (70%), with only 30% being male. Table 3.2 presents an overview of the candidates for the pilot study. The students' proficiency, however, was estimated to range from lower-intermediate to advanced level on the basis of the levels they were assigned to at the institution.

Table 3.2. Description of the cross-sectional data.

ESL learners' profile	
Number of participants	42
Collected essays	130
Levels	Intermediate-advanced
Gender	70% Female, 30% Male
Average age	28
Time in the UK	At least two months
Nationality	Saudi, Chinese, Qatari, Omani, Italian, Spanish, Thai, Iraqi, Peruvian

3.3.2 B2 and C1 sub-corpora

In total, only 109 essays were appropriate for the pilot study and incorporated into either B2 or C1 sub-corpora; other essays were incorporated into other CEFR levels. B2 sub-corpus was restricted to 59 essays totalling 15,488 words, C1 sub-corpus consisted of 50 essays totalling 12,752 words, as displayed in Table 3.3.

Table 3.3. Description of B2 and C2 sub-corpora.

Modules	**B2 Corpus**	**C1 corpus**
Type of essays	Argumentative and explanatory	
Number of texts	59	50
The average length of the essays	262	255
Total number of words	15488	12752
Total type of words	2723	2513

3.4 Longitudinal study

The relationship between language competence and the number of LBs identified in a text is still debated. As discussed in section 2.8.3, a number of previous studies found that advanced L2 students relied more on LBs than lower-level students (e.g., Chen and Baker, 2010; Ädel and Erman, 2012). On the other hand, some studies found that the percentage of LBs decreased as students moved to higher levels (Hyland, 2008a; Staples et al., 2013). It should be noted that most of the previous studies have based on cross-sectional corpus data; only a small number of studies have focused on the development corpus (e.g., Cooper, 2016; Candarli, 2020). Thus, it is necessary to adopt a longitudinal study design to measure the development of LBs in ESL learners' academic writing across proficiency levels. It is my intention that this case study will shed light on this issue.

The longitudinal study tracked two ESL students over a period of three months to observe the development of their usage of LBs in their academic essays. The longitudinal study participants were two Saudi students at the upper intermediate level who then allocated to the advanced level after a month. They were aged 24 and 29 years, had studied English for over eight years. Students were asked to write at least four argumentative essays between 200 to 400 words as a weekly homework assignment. In total, the two participants provided 65 essays which constituted the longitudinal corpus. Following the same rating procedure applied in the cross-sectional study, all essays were assigned to raters, who categorized them under the appropriate sub-corpora. In total, 20 essays were incorporated into the B2 sub-corpus, totalling 5,007 words and C1 sub-corpus consisted of 40 essays totalling 10,597 words.

3.5 The reference corpus

Most corpus linguistics studies dealing with learner corpora use a comparative or reference corpus which designed to present general information about a specific language (e.g., the 100-million-word British National Corpus), and 'often used as a baseline in comparison with more specialized corpora' (Hunston, 2002, p.15) and more importantly, to minimize subjectivity and to guarantee the reliability of the results. In other words, the reference corpus must cover a wide range of the language to be used as a benchmark that the researcher can regard as a standard of comparison. Before choosing the reference corpus, there are three factors of concern: the purpose of the study, type and mode of texts (written or spoken) (Sinclair, 1991). Once these concerns have been addressed, the only requirement for an appropriate reference corpus appears to be that it should be larger than the target corpus to represent the characteristics of the vocabulary so that it can be used as the basis of a study.

A question thus arises as to what is a sufficient size for a reference corpus. In this connection, Tribble (1999) used two reference corpora, the one million-word FLOB Corpus and the 100 million-word BNC, to examine the top ten positive and negative keywords from two different word lists. The study found that the size of the reference corpus was unimportant and did not affect the result. In the same vein, Goh (2011) examined four factors related to the reference corpus that might affect the result of a keyword calculation (corpus size, genre, varietal difference, and diachrony), and found that genre and diachrony were the only two factors that influenced the results. This confirms the previous study's finding that the reference corpus size is unimportant. On the other hand, Berber-Sardinha (2000) compared five English corpora with reference corpora of different sizes to answer the question: How large must a reference corpus be? The study revealed that for a reference corpus to be sufficient, it should be five times larger than the target corpus, explaining that the larger the reference corpus, the more keywords would be found in the target corpus. It can be concluded that, what counts as large enough examined and would an acceptable size for a reference corpus is debatable.

In my point of view, the reason for using a reference corpus is to have a benchmark against which you discover the main features in the study corpus. This is so much related to the concept of salience. We need to know the degree of saliency of each feature in the study corpus in relation to the reference one. Therefore, it will not be a sufficient degree unless your reference is bigger in size.

By considering the aim, genre and size of the target ESL learners' sub-corpora, the BAWE corpus has chosen as a reference corpus. The BAWE was compiled from academic works written at universities in the UK as a part of a project entitled "An Investigation of Genres of Assessed Writing", and was a cooperation between the Universities of Warwick, Reading and Oxford Brookes. It represents British university students' (native and non-native) academic writing of four disciplinary areas: Arts and Humanities, Social Sciences, Life Sciences and Physical Sciences, across four years of study (first-year undergraduate and taught masters level), and representing 35 main disciplines. Various methods were employed to gather assignments for the BAWE project. It aims to fill a gap in corpus provision and facilitate for determining some of the linguistic features of university students writing. The corpus contains 2,761 texts of proficient assessed academic works written at universities in the UK (totalling 6,506,995 words), ranging in length from approximately 500 words to around 5,000 words. The corpus was collected from both L1 and L2 university students for identifying "the characteristics of proficient student writing produced for degree programmes in British universities" (Nesi et al., 2008, p.2). Although the BAWE corpus bringing together L1 and L2 writing and described as a learner corpus, the corpus considered as proficient university writing, regardless of their learners' first language or cultural background. The language drawn from the corpus is clearly shown the academic tone which enhanced by subject-specific focus. Taken together, the BAWE corpus cover a wide range of the language, and it is suitable to be used as a benchmark that the researcher can regard as a standard of comparison.

Therefore, the rationale for using the BAWE corpus for comparison with the ESL learners' sub-corpora emerge from the focus on the attributes of the students' writing within the corpus. In addition, it contains different genre, mainly students' assignments (e.g., essays, critiques, case study) so represent academic writing in general. Similar to the BAWE, the ESL learners' sub-corpora consist of essays submitted by learners in partial to fulfilment test requirements. Therefore, comparing the language of ESL learners with that of BAWE allows for the identification of lexical features that are specific to ESL as opposed to proficient writers.

However, huge size difference between the target corpus and the reference corpus makes the later unsuitable to serve as a reference corpus as it makes huge differences between the observed value and the expected value between corpora. As Hoey (2009, p.5) notes that the more observed and expected values differ, the more chi-square will 'tend to compute erroneous answers'. Therefore, to ensure comparability with a small dataset used in this study, only the first 65 texts of the BAWE corpus were selected for the investigation. This was a sufficient amount for a reference corpus and was used in the pilot study, comprising 163,091 words – this is more than five times greater than the target sub-corpora (B2 and C1), having 15,488 and 12,752 words, respectively. The BAWE corpus was downloaded from the online free version and analysed using *WordSmith* tool (*WST*) to calculate the frequency of occurrences of three- and four-word LBs and compare this with the bundles identified in the learners' sub-corpora. Table 3.4 provides an overview of the size of the reference sub-corpus used in this study.

Table 3.4. An overview of BAWE corpus.

Concepts	**Wide range**
Number of papers	65
Years	2004-2007
Type of text	student assignments
The average length of texts	2509
Word Types	13111
Total number of words	163091

3.6 Research questions

1. What are the most frequently used three- and four-word LBs in B2 and C1 sub-corpora?
2. What distinguishing features a keyness analysis tell about LBs identified in B2 and C1 sub-corpora in comparison with the BAWE?
3. How do LBs in B2 sub-corpus differ from C1 in terms of structures and function?
4. To what extent does the use of LBs correlate with the learners' level of proficiency?

The four research questions above were chosen based on a corpus-based methodology, which has been used in similar previous studies to compare structural and functional aspects of LBs across two corpora (e.g., Biber et al., 2004; Hyland, 2008b; Jablonkai, 2009; Ädel and Römer, 2012). The corpus-based approach allows for the analysis of more linguistic features than other approaches (Biber et al., 1998b). A corpus-based approach to research, according to Conrad (2004, p.69), "has allowed us to understand patterns of variation more comprehensively" and can also "describe variation in the use of a specific feature of the language, rather than to characterise a variety."

3.7 Analytical framework

The analysis used to answer RQ1 and RQ2 was provided by *WST* (Scott, 2012). To retrieve three- and four-word bundles from the given sub-corpora, the first step is choosing the frequency threshold. As discussed in section 2.4.1, previous research has suggested that the frequency cut-off point used to identify LBs ranges between 10 and 40 times or more in every million words (Biber et al., 2004; Biber and Barbieri, 2007; Hyland, 2008a; Chen and Baker, 2010; Jalali, 2015). However, the frequency cut-off point is somewhat arbitrary and "based on the aim and on the researchers' evaluation of data manageability" (Chen, 2008, P.64), and there is no agreement in the literature on the correct cut-off point. In this study, to ensure a stronger relationship between LBs and academic performance, a high-frequency cut-off point of four times per 100,000 words (40 times per million words) was selected to include highly used LBs in the analysis and eliminate low-frequency parameters.

In addition to frequency cut-off, dispersion criteria were also applying. It is recommended that a bundle must be found in at least three to five texts (Cortes, 2004; Chen and Baker, 2010; Biber and Barbieri, 2007), or in at least 10% of the texts (Hyland, 2008a) to avoid focusing on idiosyncratic uses by individual speakers of the texts (see section 2.4.2). In the present study, the converted frequency for the dispersion criteria of the B2 corpus is 0.6 times, which was rounded up to two occurrences in 15,488 words (two being the minimum frequency found in the literature) (Table 3.5). The converted frequency for the C1 corpus was 0.5 times, rounded up to two occurrences in 12,752 words, while the converted frequency for the reference corpus was 6.5, which was rounded up to seven occurrences in 163,091 words. The dispersion criteria for both sup-corpora and the reference corpus was set at three different texts, as shown in Table 3.5.

Table 3.5. Cut-off points in the cross-sectional study.

Corpus	Cut-off points in absolute frequency	Range of texts
B2	0.6 rounded up to 2 in 15488 words	3
C1	0.5 rounded up to 2 in 12752 words	3
BAWE	6.5 rounded up to 7 in163091 words	3
Normalized frequency of the cut-off point	Four times per 100,000 words	

After retrieving the corpus and setting the frequency and distribution criteria, *WST* provided lists of three- and four-word LBs for both sub-corpora. In order to narrow down the included LBs, some exclusion criteria were applied. As recommended by Chen and Baker (2010), all content-based bundles (e.g., proper nouns), such *Nowadays computer education,* were discarded as they do not reflect on the use of the general academic language. These bundles appeared in a limited number of texts and were therefore considered to be content-specific or represent an idiosyncratic use of language (Biber, 2006a, p.134). In other words, LBs of terminology and context-based phrases do not reflect students' use of the more generally applicable bundles that occur across a wide range of subject areas, and so impede the likelihood of reaching conclusions regarding the use of bundles outside one subject-specific area. This view is supported by Ädel and Erman (2012), who excluded "topic-related bundles" as they do not reflect the use of LBs by students of different levels. In order to exclude the content-based bundles, LBs that contained words directly connected to the topic of the essay were manually discarded from the lists.

Following Chen and Baker (2010), the overlapping LBs were also combined as one bundle to avoid duplication in the counting of high-frequency bundles. There are two types of overlapping. The first occurs when the three- or four-word bundles originate from longer bundles, such as five- or six-word bundles. For example, the three-word bundle *on the other,* constructed from the four-word bundle *on the other hand,* occurs 19 times in the B2 corpus. The second type of overlapping is where two bundles overlap and one of the phrases subsumes the other bundle, via complete submission; for example, *first of all* occurs 14 times, while *of all the* occurs only three times. These two bundles occur as subsets of the four-word bundle *first of all the.* To avoid inflation of the analytical results, overlapping bundles were combined to make a single longer unit, adding the fifth word in brackets. For example, the bundles *on the other hand* and *the other hand the* were combined to make one single bundle: *on the other hand + (the).* After refinement, the lists of high-frequency occurring bundles were compared in order to answer RQ1.

The second research question concerns about the keyness analysis. 'Keybundles' are those that occur unusually frequently in the target sub-corpora compared to the reference sub-corpus. The procedure to identify the keybundles began by using *WST* to compare the word lists of the most frequently used LBs in the B2 and C1 sub-corpora with the word list in the reference sub-corpus. Then, the keyness of the high-frequency LBs was examined to identify the bundles that occurred differently in B2 and C1 sub-corpora as in the reference corpus. The actual calculation included testing the keyness using the log-likelihood (LL) test and the effect size, which provides a more accurate result than other tests. The (LL) test aims to establish the degree to which the differences between high and low levels are significant (Scott, 2012) and the effect size reveal the size of a frequency difference. The main procedures implemented to test the keyness have been used in a range of previous studies (e.g., Scott and Tribble, 2006; Culpeper, 2009).

The *P-value* threshold for the log-likelihood test is arguably arbitrary, so a small threshold was set to obtain fewer keybundles; the threshold was 0.000001 (default threshold in *WST*), meaning there is a one in a million likelihood that it occurred by chance. In other words, "The smaller the P-value, the more likely that the word's strong presence in one of the sub-corpora is not due to chance but a result of the author's (conscious or subconscious) choice to use that word repeatedly" (Baker, 2006, p.125). Finally, a list of 'Keybundles' was provided by the *WordSmith* tool for unusually occurring bundles. A positive keyness of a bundle indicates the significant use of a bundle in a corpus, whereas a negative keyness means that a bundle is underused in the corpus. This quantitative analysis was followed by examining the structures and functions of LBs to answer RQ3.

From the structural and functional point of view, the LBs were classified structurally using Biber et al. (1999) taxonomy to answer RQ3. As discussed in section 2.4.5, the taxonomy has been modified and developed as recommended by Biber et al. (2004) in order to place the identified bundles that could not be classified under Biber et al. (1999) original structural taxonomy. Therefore, the framework was developed for the purpose of the present study to incorporate all the target bundles, including those that were not present in Biber' taxonomy. The complete adopted taxonomy used in this cross-sectional study will be presented in section 3.8.3 and for the longitudinal study in section 3.8.6.

In regard to the function of the bundles identified in the sub-corpora, Hyland (2008b) functional taxonomy, developed based on Biber et al. (2004) typology, was applied to classify the identified LBs in B2 and C1 sub-corpora and compare them with the BAWE sub-corpus to identify differences between the CEFR levels in terms of the variety in and accuracy of use of LBs. As aforementioned in section 2.4.5, Hyland's taxonomy was adopted since it designs applicable to the domain of academic writing, the data in this pilot study was primarily academic prose. The taxonomy was, therefore, more relevant to the analysis of LBs within the sub-corpora. The taxonomy is divided into three main functional categories: 1) research-oriented (or 'referential' in Biber et al.'s (2004)), 2) text-oriented (or 'discourse organising'), and 3) participant-oriented (or 'stance') bundles, each of these contains several sub-categories. Although Hyland (2008b) include content-based bundles in his study, the study followed Chen and Baker (2010), excluded all content-based bundles from the analysis, as they did not reflect the use of general academic language. The complete adopted functional taxonomy used in this cross-sectional study will be presented in section 3.8.4 and for the longitudinal study in section 3.8.7.

In order to classify LBs in the correct sub-categories, it was important to look at the concordance line to see the bundle in its context and address the issue of multi-functionality of the target bundles. It should be noted that many of the LBs were found to serve more than one function, an issue addressed in previous studies (see section 2.4.5). Following the approach employed by Biber (2006a), LBs were categorized according to their most dominant function, based on their use in each sub-corpus. For example, the bundle *at the same time* can be categorized as a time/location bundle, but it was classified as a transition bundle, according to the function of majority of the occurrences in the target sub-corpora.

After the analysis of structural and functional of the identified LBs, chi-square (plus standardised residuals) statistical methods are used to support related arguments further, as the test used to assess the degree of difference in the use of structural types between the corpora. Levon (2010, p.74) explains that the chi-square test aims to determine whether the proportional distribution observed in the sample population is significantly different from any other population of the same size and shape. The threshold set for the chi-square test is 0.05; if the result is lower than this value, it means that chi-square value is significant and there is sufficient evidence that the difference between the corpora is significant and not due to chance. Then, a residual analysis can be carried out to identify where this significant result is coming from and which particular cells are causing the difference between the groups. A residual helps to find the difference between the observed and expected values for each cell. If the residual value is high, the contribution of the cell to the magnitude of the chi-square value that is obtained will be greater. As stated by Agresti (2018), “a cell-by-cell comparison of observed and estimated expected frequencies helps us to understand the nature of the evidence better”.

As Hyland (2008a, p. 60) states, “the study of clusters offers insights into a crucial, and often overlooked dimension of language use, providing a better understanding of the ways writers employ the resources of English in different contexts, and with the potential to inform advanced academic literacy instruction.” Therefore, the structural and functional features of the identified LBs were used to answer RQ3.

The analytical procedures required for answering RQ4 began with identifying LBs. The cut-off points and dispersion criteria of applied in the cross-sectional study were used in the longitudinal study (40 times per million words in three different texts), as shown in Table 3.6.

Table 3.6. Cut-off point and dispersion criteria of LBs.

Corpus	**Cut-off points in absolute frequency(F)**	**Range of texts**
B2	0.2 rounded up to 2 in 5007 words	3
C1	0.4 rounded up to 2 in 10597 words	3

Once the lists of the most frequent LBs in the B2, and C1 sub-corpora were ready for analysis, the two lists were compared, in order to assess the correlation between the use of LBs and learner levels to answer RQ4. If the methodology used in the pilot study to identify LBs and keywords analysis success, it will be applied in the main study.

3.8 Research findings and discussion

3.8.1 Frequency distribution of B2 and C1 sup-corpora

This section presents cross-sections study employed to examine the differences in the use of LBs across the levels. The first research question of this study asked, ‘What are the most frequently used three- and four-word bundles in B2 and C1 levels sub-corpora?

The overall number of LBs identified in B2 and C1 sub-corpora were 147 (type) and 658 (token), as shown in Figure 3.1. The B2 sub-corpus accounted for 102 types of three- and four-word LBs, which occurred 458 times, making up 9.2% of the total number of words in that sub-corpus, while the C1 essays contained 45 types of three- and four-word LBs, which occurred 204 times in that sub-corpus, making up 5% of the total number of words. What stands out in the graph below is that the low-level students used a larger stock of LBs than the high-level students. However, the low number of LBs identified in C1 sub-corpus does not support the notion that higher-level students tend to use LBs to a greater extent than low-level students. These results are in agreement with those obtained by Hyland (2008a), who found that MA writers used a wide range of LBs with greater frequency than PhD writers, who in turn used more than expert writers.

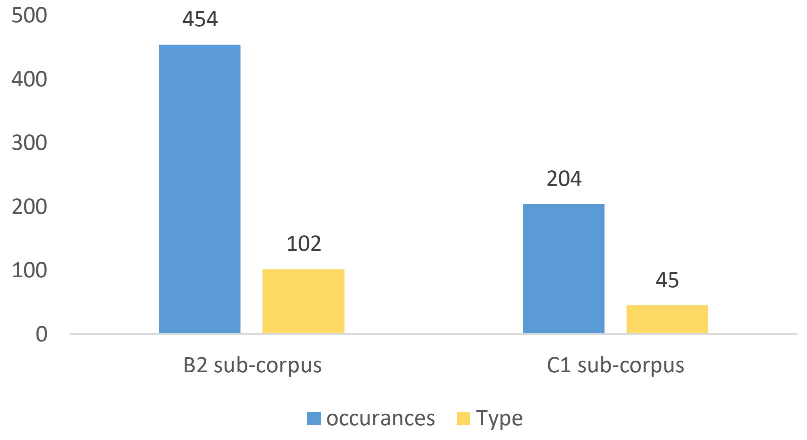


Figure 3.1. Overall LBs (types and tokens) found in B2 and C1 sub-corpora.

Turning now to the distribution of the three- and four-word bundles in the sub-corpora. Figure 3.2 shows that three-word bundles were the most common bundles in both sub-corpora, accounting for approximately 84% of total bundles at both levels. However, in regard to four-word bundles, these were less frequent than three-word bundles, accounting for approximately 16% of the B2 and C1 sub-corpora. This might be related to the complexity of their production, causing language learners to avoid using them in their writing.

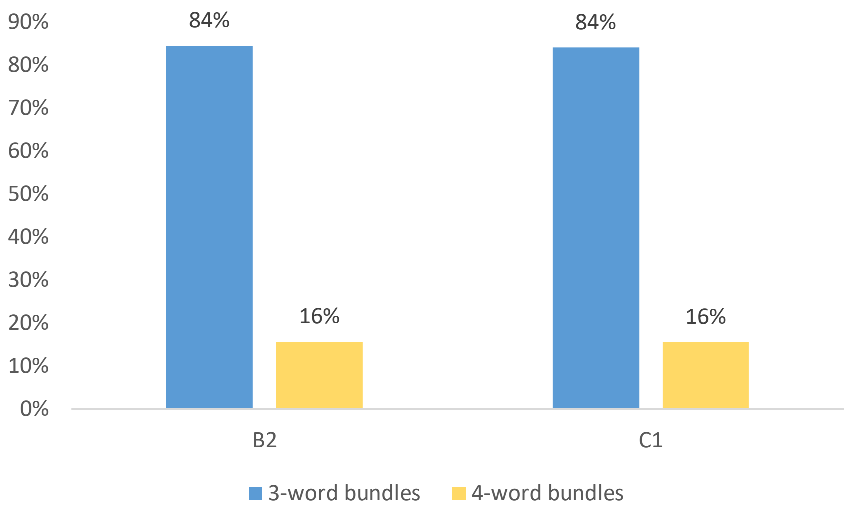


Figure 3.2. Distribution of 3 and 4-word bundles in the two sub-corpora.

Moving on to the comparison between the two groups, in total, 86 types of three-word LBs, totalling 388 LBs, and 16 types of four-word LBs, totalling 66 LBs, were identified from 15,488 words in the B2 sub-corpus. By contrast, there were fewer LBs in the C1 corpus, with only 38 types of three-word LBs, totalling 169 LBs, and seven types of four-word bundles, totalling 35 LBs, identified, which met the cut-off point.

With respect to the RC (BAWE), 328 types of the three-word bundle, totalling 4,220 LBs, were identified from 163,091 words. Thus, the LBs were constituting approximately 7.7% of the total words in the BAWE corpus. On the other hand, 37 different types of four-word LBs were identified, totalling 400 LBs. The frequency statistics of bundles (types) and frequency (occurrences) identified in the corpora are shown in Table 3.7. As can be seen, there was no difference between the three groups in the use of four-word LBs. A similarity in the frequency of occurrences was found between the three corpora, with an increase in the B2 sub-corpus. This is a surprising result as it was expected that advanced students would rely more on LBs than low-level students. The results of LBs did not reflect any gradual changes in the usage of LBs, across B2 and C1 levels. However, the low-level students tended to use more bundles than the high-level students, and, notably, there was no increase in the use of LBs found between the levels. Figure 3.3 provides an overview of the percentage of LBs identified in the three groups.

Table 3.7. Total number of bundle types and tokens across the sub-corpora. (Freq = Frequency; % of the bundles in a corpus)

Sub-corpora	Freq LBs	Type	%	(F) Per 100,000*
Three-word bundles				
B2	388	86	7.50 %	2505
C1	169	38	4.00 %	1325
BAWE	4220	328	7.70 %	2587
Four-word bundles				
B2	66	16	1.70 %	426
C1	35	7	1.00 %	274
BAWE	400	37	0.98%	245

* The raw frequency was normalised per 100,000 words to provide a basis for comparison

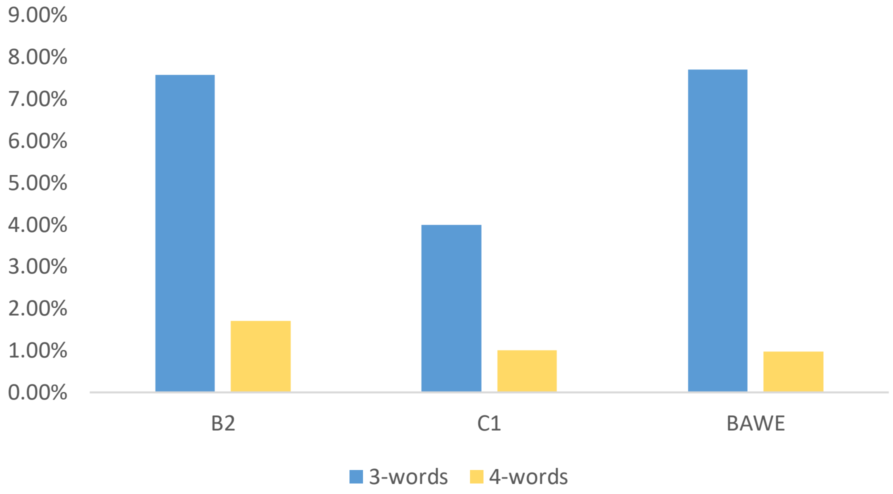


Figure 3.3. Frequency percentage of target bundles used by B2, C1 and BAWE corpora.

The degree to which the differences between the sub-corpora are significant was tested using the chi-square test (*p*<.05). The chi-square test applied to test the frequency occurrences of LBs across the sub-corpora regardless of the bundles' length. The results of the statistical analysis revealed that the difference between the groups was significant at a *P*-value= 2.9319E-145 less than .05. According to this result, B2 writers were found to use significantly more LBs in their writing than C1 and BAWE writers.

The study also enquired about whether there was a variance in the use of certain bundles across the corpora. A comparison of the two sub-corpora with the reference corpus helped to identify the shared bundles. Of the 86 three-word bundles identified in the B2 sub-corpus, 24 (28%) were also found in the BAWE corpus. In comparison, 13 (33%) three-word LBs identified in the C1 sub-corpus were found in the BAWE sub-corpus. With respect to the four-word bundles, only two of the bundles identified in the B2 and C1 sub-corpora were found in the BAWE sub-corpus, at a rate of 13% and 28%, respectively. However, as the pilot study was a small-scale test of efficacy for the main study, it compared only the top 10 LBs found in the three groups.

In regard to the ten most frequent bundles in the BAWE sub-corpus, the three-word bundles occurred between 22/100,000 and 54/100,000 times. The bundle *in order to* was the most frequent bundle, with 54/100,000 occurrences, followed by *as well as* (43/100,000 occurrences) (Table 3.8).

Closer inspection of the below table shows that almost half of the ten most frequent three-word bundles in the B2 sub-corpus were also found in the C1 sub-corpus but were not preferred by university writers in the reference sub-corpus. Furthermore, no three-word bundle was shared between all three groups, and only three bundles were shared between the reference sub-corpus and one of the sub-corpora. For example, the bundles *in order to* and *as a result* were shared by B2 sub-corpus and BAWE sub-corpus, while bundle *one of the* was shared by C1 sub-corpus and BAWE sub-corpus. Thus, there appears to be a similarity in LBs between B2 and C1 sub-corpora, although the C1 group uses far fewer bundles than B2.

Table 3.8. The ten most frequent 3-word bundles in B2, C1 and BAWE corpora. (italic = items occur in 2 sub-corpora; AF =Absolute frequency; NF= Normalised frequency)

Bundles in B2 corpus			**Bundles in C1 corpus**			**BAWE corpus**		
Type	**AF**	**NF**	**Type**	**AF**	**NF**	**Type**	**AF**	**NF**
on the other	19	122	*on the other*	14	109	*in order to*	89	54
first of all	14	90	*a lot of*	9	70	as well as	71	43
point is that	14	90	*first of all*	8	62	the fact that	56	34
some of the	10	64	*one of the*	8	62	the development of	54	33
in my opinion	9	58	it is a	7	54	in terms of	49	30
in order to	9	58	in the past	6	47	*as a result*	47	28
a lot of	8	51	be able to	5	39	such as the	42	25
as a result	8	51	it can be	5	39	due to the	41	25
to sum up	7	45	*to sum up*	5	39	*one of the*	40	24
we need to	7	45	a long time	4	31	the number of	37	22

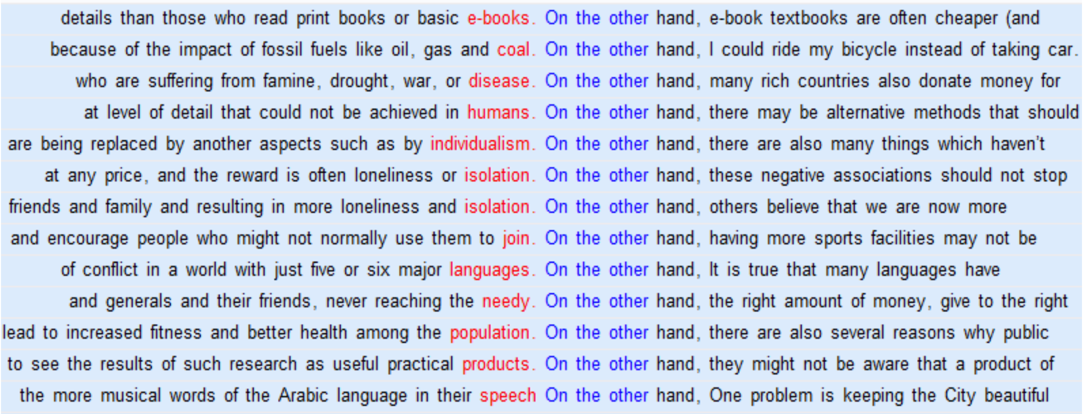
By contrast, the frequent four-word bundles, which occurred 73 times in the B2 and C1 sub-corpora, accounted for almost 3% of the total words, as displayed in Table 3.9. The top 10 four-word LBs in the B2 sub-corpora occurred between 19-109 times in 100,000 words. The most frequently occurring bundle was *on the other hand*, with 109 occurrences. However, only seven four-word bundles were identified in the C1 sub-corpora, ranging between 23-109 times in 100,000 words. The bundle *on the other hand* had the most occurrences, 109 times. Moreover, the bundle *on the other hand* was significant in that it was the most frequent bundle on both lists; the remainder of the LBs occurred at almost a similar frequency.

With respect to the reference sub-corpus, the ten most frequent four-word LBs occurred between 8-15 times in 100,000 words. The most frequent bundles were *as a result of* and *on the other hand*, occurring 15 times in 100,000 words.

Table 3.9. The ten most frequent 4-word bundles in B2, C1 and BAWE sub-corpora. (Bold = item occurs in 3 corpora; italic = items occur in 2 corpora; AF =Absolute frequency; NF= Normalised frequency)

B2			**C1**			**BAWE**		
Type	**AF**	**NF**	**Type**	**AF**	**NF**	**Type**	**AF**	**NF**
on the other hand	17	109	**on the other hand**	14	109	*as a result of*	25	15
second point is that	5	32	*is one of the*	5	39	**on the other hand**	25	15
at some of the	4	25	one of the most	4	31	the end of the	22	13
different from each other	4	25	*a lot of people*	3	23	the significance of the	17	10
a major role in	3	19	*another point is that*	3	23	as well as the	15	9
a wild range of	3	19	*is going to be*	3	23	in the development of	14	8
another point is that	3	19	third point is that	3	23	it has been shown	14	8
at the same time	3	19	x	x	x	the fact that the	14	8
first of all the	3	19	*x*	x	x	*at the same time*	13	8
I will discuss whether	3	19	x	x	x	can be used to	13	8

As can be seen in Table (3.11), it is apparent that the bundle *on the other hand* was the most frequent bundle in B2 and C1 corpora, the only bundle shared across the three groups of writers, with a similar frequency in ESL learners' sub-corpora. A closer inspection of Tables 3.10 and 3.11 suggests that low-level students used frequent and different types of LBs than high-level. Interestingly, LBs were used with broadly similar frequency in B2, C1 and BAWE writers, except for the bundles *on the other hand*, which was used far more often in the ESL learners' writing. Usually, this bundle was more frequently used in argumentative essays as a way of addressing the second part of a two-part problem, situation, or solution. Random samples of concordance lines for the bundles *on the other hand* are provided below.



```
         details than those who read print books or basic e-books. On the other hand, e-book textbooks are often cheaper (and
       because of the impact of fossil fuels like oil, gas and coal. On the other hand, I could ride my bicycle instead of taking car.
            who are suffering from famine, drought, war, or disease. On the other hand, many rich countries also donate money for
              at level of detail that could not be achieved in humans. On the other hand, there may be alternative methods that should
are being replaced by another aspects such as by individualism. On the other hand, there are also many things which haven't
       at any price, and the reward is often loneliness or isolation. On the other hand, these negative associations should not stop
friends and family and resulting in more loneliness and isolation. On the other hand, others believe that we are now more
 and encourage people who might not normally use them to join. On the other hand, having more sports facilities may not be
          of conflict in a world with just five or six major languages. On the other hand, It is true that many languages have
            and generals and their friends, never reaching the needy. On the other hand, the right amount of money, give to the right
lead to increased fitness and better health among the population. On the other hand, there are also several reasons why public
 to see the results of such research as useful practical products. On the other hand, they might not be aware that a product of
   the more musical words of the Arabic language in their speech On the other hand, One problem is keeping the City beautiful
```

Figure 3.4. Random samples of concordance lines for the bundles *on the other hand.*

The analytical insights relevant to the first research question highlighted the greatest difference between the three corpora in terms of the frequency of use of LBs. Interestingly, only the bundle *on the other hand* occurred more than 100 times per 100,000 words in the B2 and C1 sub-corpora; the remainder of the bundles on the lists occurred at a largely similar frequency. By contrast, no bundle exceeded 100 occurrences per 100,000 words in the reference sub-corpus. Thus, it is somewhat surprising that the frequency of LBs occurring in the learner corpus was slightly higher than was found in the BAWE sub-corpus.

Examining both frequency and keyness of LBs provides potential insight into the data collected. The Keyness of a bundle provides an indicator of a bundle's importance in a corpus. Thus, the next section will examine the keyness of the identified LBs in B2 and C1 sub-corpora to answer RQ2.

3.8.2 Keybundles/Keyness analysis

As mentioned in section 3.7, *WST* was used to analyse the keybundles identified in B2 and C1 sub-corpora in comparison with the RC, using the log-likelihood test and effect size metrics. In total, nine keybundles in B2 sub-corpus fulfilled the frequency and dispersion criteria applied in this study (Table 3.10). A closer inspection of the keybundles showed that 60% were types of connector expressions. The bundles with the highest keyness values were *point is that* and *first of all*. Turning now to keybundles used in C1 sub-corpus, it can be seen that the bundle *on the other (hand*) is only significantly overused by the C1 learners, as shown in Table 3.11.

Table 3.10. 3-and 4-word key LBs in B2 sub-corpus with significantly different frequency from those in BAWE corpus. (F= Frequency, RC= Reference corpus)

LBs	**F in B2**	**% in B2**	**F in RC**	**% in RC**	**Keyness**
point is that	14	0.09	0	0	68.69
first of all	14	0.09	2	0	57.00
to sum up	7	0.05	0	0	34.34
on the other	19	0.12	16	0.02	34.00
I will discuss	6	0.04	0	0	29.44
are very different	5	0.03	0	0	24.53
I believe that	5	0.03	0	0	24.53
seem to be	5	0.03	0	0	24.53
on the other hand	17	0.11	25	0.02	31.22

Significant at (*P*<0.000001). A positively keybundle occurs more often than would be expected by chance in comparison with the reference corpus.

Table 3.11. 3-and 4-word key LBs in C1 sub-corpus with significantly different frequency from those in BAWE corpus. (F= Frequency, RC= Reference sub-corpus)

LBs	**F in C1**	**% in C1**	**F in RC**	**% in RC**	**Keyness**
Can choose to	5	0.04	0	0	26.52
On the other hand	14	0.11	29	0.02	24.23

Significant at (*P*<0.000001). A positively keybundle occurs more often than would be expected by chance in comparison with the reference corpus.

The primary differences in the use of keybundles indicate that B2 and C1 learners were more likely than BAWE writers to use connecter phrases (time/order) to organise their essays. The terms ‘connectors’ and ‘transition’ in this study refer to the three- and four-word bundles that tie related sentences together, play an important role in writing academic essays. This was an expected result, as English writers use connectors to make the point clear to readers and to transition between different ideas, for example:

- A *second point is* that many foreign aid projects are unsuitable for the target country. (B2, essay76)
- *On the other hand*, there are many stronger reasons why university should be coeducational. In the first place, it is good preparation for the real world (C1, essay65)
- I think foreigners should pay more for many reasons. *First of all*, it bring more money to maintain the attraction. The higher admission fees from foreigners are important. (B2, essay 53)

3.8.3 Structural analysis

As discussed in section 3.7, the bundles identified in the B2 and C1 sub-corpora were classified according to the structural taxonomy proposed by Biber et al. (1999), which has been modified and developed in order to place the identified bundles that could not be classified under Biber original structural taxonomy. The taxonomy consists of four structural categories, each with sub-categories. The results obtained from the preliminary analysis of structural categories are displayed in Figure 3.5.

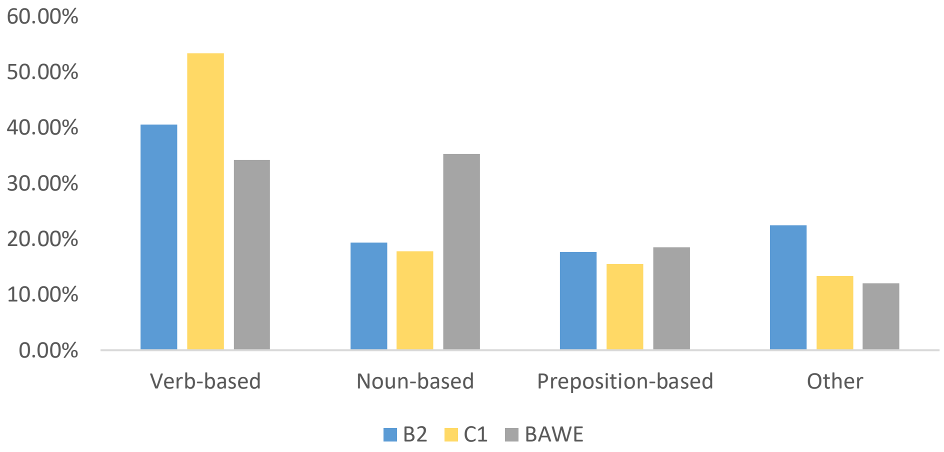


Figure 3.5. Overall distribution of the structural categories across the B2, C1 and BAWE corpora.

The results above show variations in the use of LBs within B2 and C1 sub-corpora according to the four structural categories. It is clear that there was considerable disparity in the application of structural types by B2 and C1 students and BAWE writers. Table 3.12 shows the variations in the use of different structural sub-categories of the most frequent three- and four-word bundles in the corpora.

Table 3.12. Structural types of three-and four-word LBs in the corpora. (Freq = Frequency; % = percentage within-sub-corpus)

Structural types	**Sub-types**	**B2**	**C1**	**BAWE**
		Freq (%)	**Freq (%)**	**Freq (%)**
Verb-based	Anticipatory it + verb / adjective phrase	1 (1)	4 (9)	29 (8)
	Copula be + noun /adjective phrase	3 (3)	2 (4)	9 (3)
	Pronoun/NP + be	13 (13)	8 (18)	18 (5)
	First person pronoun + dependent clause	9 (9)	1 (2)	1 (0)
	(verb/adjective +) to-clause	12 (12)	9 (20)	36 (10)
	Other verb phrases	3 (3)	-	30 (8)
	Totals Verb based bundles	41 (41)	24 (53)	123 (34)
Noun-based	NP with of-phrase	10 (10)	5 (11)	95 (27)
	NP with other post-modifier	6 (6)	3 (7)	18 (5)
	Other noun phrases	4 (4)	-	12 (3)
	Totals noun-based bundles	20 (19)	8 (18)	125 (135)
Preposition-based	Prepositional phrase with embedded of-phrase	2 (2)	-	11 (3)
	Other prepositional phrase expressions	16 (16)	7 (16)	55 (16)
	Totals preposition-based bundles	18 (18)	7 (16)	66 (19)
Other	Other structures	23 (23)	6 (13)	43 (12)
Totals	-	102	45	357

Overall, for the most frequent three- and four-word bundles, B2 and C1 writers employed a wide variety of structures to produce these LBs in their essays. What stands out in the table is that C1 students used more verb-based bundles than B2 and BAWE writers did, while B2 writers used more bundles in the 'other' category. The majority of LBs found in both sub-corpora were attributed according to phrasal types, i.e., either verb phrases (VPs) (*we have used*), prepositional phrases (PPs) (*in the presence of*), or noun phrases (NPs) (*the importance of*). This means that B2 and C1 students were drawing on phrasal bundles as professional writers.

In addition, the verb-based category was dominant in the B2 and C1 sub-corpora, accounting for approximately 40% of the identified bundles in the B2 sub-corpus and 50% in the C1 sub-corpus, while the other three categories accounted for approximately the same proportion in both sub-corpora. It can be said that the written language used by B2 and C1 learners contain a large proportion of LBs that are more frequently used in spoken language, as previous studies have found that verb-based bundles tend to be found more often in spoken English (Biber, 2006b; Hyland, 2008b). The high proportion of verb-based bundles might have resulted from the nature of argumentative essays, which required students to use a greater variety of pronouns to state their opinion on the topic and express their argumentation clearly.

The results also showed that three sub-categories were not found in the C1 data: other verb phrases, other noun phrases, and prepositional phrases with an embedded of-phrase. It seems that B2 writers used a wide variety of bundles than C1 writers.

With regard to the comparison between the RC and the ESL learners' sub-corpora, the most conspicuous similarities between the corpora in terms of the structural classification were the increased use of 'Other prepositional phrase' bundles in the B1, B2 and BAWE writers. This sub-category is the second-highest sub-category in B2 and C1 levels at 16%. The increased use of 'Other prepositional phrase' in ESL learners' sub-corpora is associated with the significant increase of connector expressions (e.g., *on the other hand, as a result*), as discussed in the analysis. Examples of these expressions in B2 and C1 sub-corpora are shown in the following examples.

- *On the other hand*, computers may have many distractions that do not effort students to achieve the learning objectives. (C1, essay 15)
- Some families are poor and they want their children to work *in order to* increase their salary. (B2, essay26)

From the structural point of view, B2, C1, and BAWE sub-corpora also showed some differences in bundle types. The extent to which these differences between the structural sub-categories may be regarded as significant was tested statistically using the chi-square test (plus the standardised residuals). As shown in Table 3.13, the chi-square result indicates significant differences between the corpora. The standardised residuals (R), used to find the particular cells causing the difference between the groups and that to make "a major contribution to the significant difference" (Chen and Baker, 2010, p.38), was applied to the cells with a value greater than +2 or -2.

Table 3.13. Standardized residuals (R) in a chi-Square contingency table for structural distribution (types). (italic = significant interaction)

	Corpora		
	B2	C1	BAWE
Verb-based			
• Count	41	24	123
• Expected	37.89	16.7	133.3
• R	0.7	*2.3*	*-2.1*
Noun-Based			
• Count	20	8	126
• Expected	31	13.6	1.9
• R	*-2.6*	-1.9	*3.5*
Preposition-Based			
• Count	18	7	67
• Expected	18.5	6.4	65.2
• R	-0.15	0.27	0.43
Other			
• Count	23	6	43
• Expected	15.5	6.4	51
• R	*2.6*	-0.18	*-2.2*
Chi-square P < 0.05	df = 2, P-value = 0.003		

If the residual is less than -2, the cell's observed frequency is less than the expected frequency. Greater than two and the observed frequency is greater than the expected frequency.

As can be seen in Table 3.13, although the 'other structural' sub-category was increasing used in B2, C1 and BAWE writers, no significant increase was found across the three corpora in the use of preposition-based bundles. However, there were significant differences between the sub-corpora in the other three categories. For example, a significant increase in verb-based bundles was found in C1 sub-corpus. Moreover, a significant increase of the 'other structural' category was found in B2 sub-corpus. In contrast, a significant increase of the noun-based bundles was found in BAWE. Thus, it can be seen that ESL learners understand little of the written language as they shared more features of the spoken language in their writing. In contrast, as there was a significant increase of noun-based bundles in the BAWE corpus, this result coincides with previous research findings (e.g., Biber et al., 1999; Hyland, 2008a; Byrd and Coxhead, 2010) that academic writing becomes "noun-centric" (Salazar and Joy, 2011).

This result can be associated with a variety of discursive functions that serve in academic discourse, which will be discussed in the following section.

3.8.4 Functional taxonomy of lexical bundles in the ESL L sub-corpora

The results present in the previous chapter have shown that LBs made up of uncomplete structural sequences that extended across verities of structural categories.

Additionally, LBs identified in B2 and C1 sub-corpora showed particular structural patterns that give insight into the nature of academic writing in these two levels. This section provides the distribution of functional categories of LBs identified in B2 and C1 sub-corpora.

As discussed in section 3.7, Hyland's classification was adopted in this study to classify the identified bundles. The following figure illustrates the proportion of different functional categories across the sub-corpora.

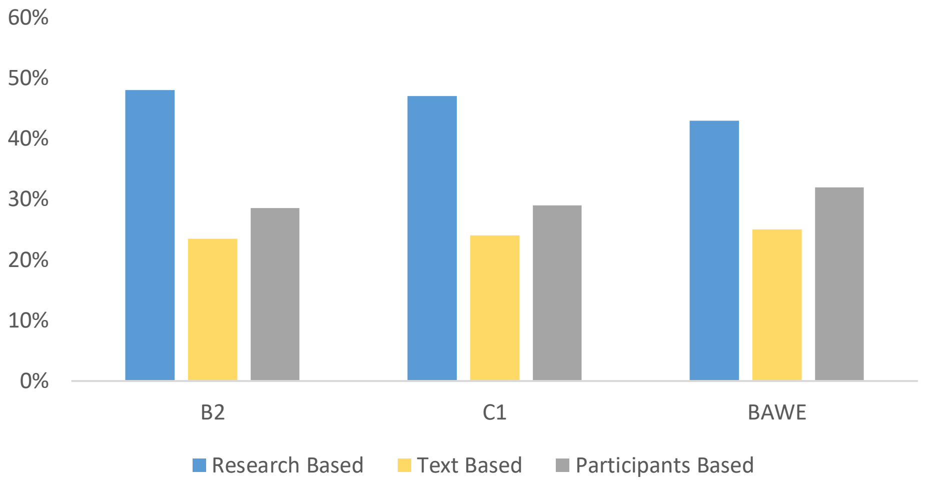


Figure 3.6. Distribution of functional categories across the corpora.

As shown in Figure 3.6, no noticeable discrepancy was found between the corpora when considering the functional distribution of the identified bundles. The results show a similarity in their usage throughout the three categories in the B2, C1 and the RC sub-corpora. However, the highest proportion of the bundles used was research-oriented in all three corpora. These bundles help writers structure their writing and experiences of the real world and are, overall, more frequent than bundles functioning to organise the text and bundles that express writers' attitude or focus.

In addition, the writers of the three corpora employed fewer text-oriented bundles in their writing. This result conflicts with previous studies by Nekrasova (2009); Chen and Baker (2010), who found that learners employed more text-based bundles or discourse organisers than research-based bundles. The following examples are parts of research-oriented bundles found in the ESL learners' sub-corpora:

- To tackle this problem the government should build flats in around the city with flat to rent at fair prices *in order to* reduce the number of homeless in the future. (B2, essay27)
- People are often quick to criticize the way other people raise their children, and there is *a lot of* pressure on parents to have perfect children. (C1, essay 47)

Closer inspection of Figure 3.6 shows indirect proportionality between the percentage of research-oriented bundles used across the three corpora. In other words, as proficiency level increased, the percentage of research-oriented bundles decreased across the three corpora. By contrast, there was a direct proportionality between the text-oriented and participant-oriented bundles across the three levels: as proficiency level increased, the use of these bundle types increased across the corpora.

Following the distribution of the functional categories of LBs, an analysis of the functional sub-categories was conducted in which the frequency of each type was calculated across the sub-corpora, as shown in Table 3.14.

Table 3.14. Functional types of LBs identified in the three groups. (Freq= Frequency; %= percentage within-sub-corpus)

	Sub-types	**B2**	**C1**	**BAWE**
		Freq (%)	Freq (%)	Freq (%)
Research-oriented	Location	7 (7)	3 (7)	18 (5)
	Procedure	5 (5)	3 (7)	18 (5)
	Quantification	18 (18)	9 (20)	29 (8)
	Description	19 (19)	6 (13)	91 (25)
	Totals Research-oriented	49 (49)	21 (47)	156 (43)
Text-oriented	Transition signals	7 (7)	5 (11)	12 (3)
	Regulative signals	2 (2)	1 (2)	13 (4)
	Structuring signals	12 (12)	5 (11)	0
	Framing signals	3 (3)	0	63 (18)
	Totals Text-oriented	24 (24)	11 (24)	88 (25)
Participant-oriented	Stance features	24 (24)	10 (22)	73 (21)
	Engagement features	5 (5)	3 (7)	40 (11)
	Totals Text-oriented	29 (29)	13 (29)	113 (32)
Total	10	102	45	357

Among the sub-categories in the research-oriented bundle, the quantification sub-category (e.g., *a lot of*) shows differences between the ESL sub-corpora and the reference sub-corpus, accounted for 18% (B2), 20% (C1) and 8% (BAWE) of bundles used. It thus appears that ESL learners used almost twice as many quantification bundles as BAWE writers did. This finding is consistent with that of Chen and Baker (2016), who found increased use of these expressions in L2 learners' writing, specifically at low-level learners.

The second highest sub-category was the description, which accounted for 18.6% (B2), 13% (C1), and 25% (BAWE) of bundles use. Many of these bundles rely on research entities or contexts, specifying aspects of models, equipment, materials, or the research environment (e.g., *the problem of, the importance of*), and are typically produced as noun phrase + of structures, for example:

- The government needs to raise awareness of *the importance of* education and even offer financial support to students to continue. This will encourage students to stay at school rather than start working. (B2, essay37)
- In conclusion, although *the problem of* drugs may seem impossible to eliminate, there are concrete steps that can be taken to weaken the hold of drugs on society. (C1, essay18)

The lowest percentage of research-oriented sub-categories across the three corpora was in the time/location and procedure sub-categories. Despite their lower frequencies in the present study, these bundles are important to the research process as they contribute to accurate documentation by identifying location/time (e.g., *at the same time*) and indicate actions, events, and methods (e.g., *is carried out*).

Moving to the participant-oriented bundles, the findings show that there were more bundles in the reference sub-corpus than the ESL sub-corpora. These expressions serve the purpose of expressing different stance meanings (e.g., *I think that*) and engagement features (e.g., *It can be seen*), which focus on the reader and writer of a text (Hyland, 2008a; Nkemleke, 2012). A depth investigation of the concordance lines showed differences in the use of these expressions across the groups. Whiles the increased use of these expressions in ESL learners was due to the heavy reliance on 'First-person pronoun + dependent clause' bundles (e.g., I believe that), the BAWE writing was full of 'Anticipatory it + verb / adjective phrase' bundles (e.g., it is important to). It can be seen that ESL learner favoured using personal pronouns to express their opinion and connect with the readers instead of the impersonal pronouns commonly used in BAWE writing. The following is a contextualised example of these bundles:

- *It is possible* to read that the tragic hero is trying to avoid later guilt for his actions and thus does not want to leave behind him evidence of the murder. (BAWE, essay3006b)
- *I believe it* will be better for work and make college more enjoyable. (B2, essay37)
- *I think it is* reasonable for foreigners to pay more for many reasons. First of all, it brings more money to maintain the attraction. (C1, essay80)

The least frequently used bundles in all corpora were text-oriented bundles, which concerned with the organisation of the text. Those bundles were the most common sub-category in the B2 and C1 sub-corpora, while framing sub-category achieved the highest percentage in the RC:

- *First of all*, older employees have an immense amount of knowledge and experience which can be lost to a business or organization if they are made to retire. (C1, essay 29)
- *The fact that the* seedlings used in this investigation were so premature was a major limiting factor and therefore we would need to repeat this investigation using more developed plants before any staining of our promoter traps could be observed. (BAWE, framing signal)

In order to test whether the difference between the functional categories across the writers is significant, the chi-square test was applied. However, unlike the structural distributions, *P*-value at 0.94 showed no significant differences in the functional distribution were found in the B2, C1, and BAWE corpora. The results support the initial observation that no differences occur between the sub-corpora as the percentage in all the three categories is quite similar.

3.8.5 lexical bundles development

To answer RQ4, which is concerned with the development of LBs across the levels, two ESL students were tracked over a period of three months across the two levels (see section 3.3). As discussed in section 3.7, *WST* was used to provide lists of the most frequently used bundles in both sub-corpora. Each sub-corpus was analysed separately, and then the lists were compared to determine whether the use of LBs correlated with the level of proficiency. The results are displayed in the form of descriptive statistics, as shown in Table 3.15.

Table 3.15. Total number of bundle types and tokens in B2 and C1 levels.

Sub-corpora	**Frequency of LBs**	**type**	**(F) Per 100,000**
Three-word bundles			
B2	75	20	1498
C1	178	34	1679
Four-word bundles			
B2	9	3	179
C1	18	5	169

From Table 3.15, it can be seen that both students in B2 level used a total of 75 bundles (1,498 per 100,000 words), consisting of 20 types of three-word bundles, and nine four-word bundles (179 per 100,000 words). By contrast, C1 level students used more three- and four-word bundles, with 178 (tokens) (1,679 per 100,000 words), consisting of 34 types of three-word bundles, and 18 (tokens) of four-word bundles (169 per 100,000 words). Figure 3.7 below illustrates the difference in target bundles in the students' sub-corpora.

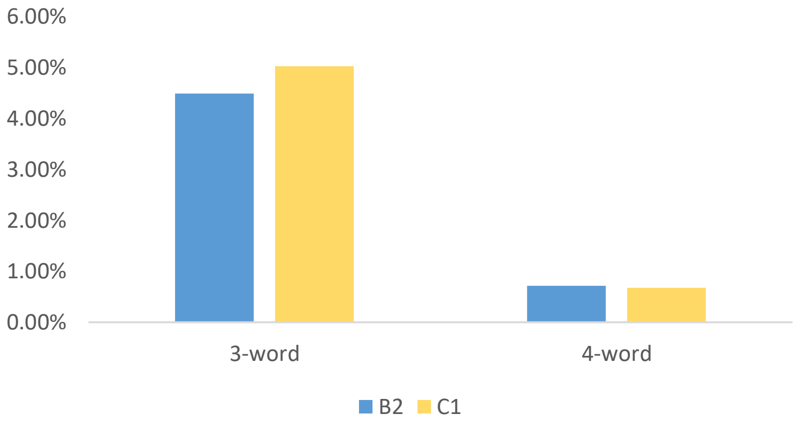


Figure 3.7. Percentage of target bundles in ESL learners' sub-corpora.

The figure above shows a notable difference between the three- and four-word bundles used by ESL learners in their academic writing. It is clear that ESL learners preferred to use shorter bundles in their essays. In addition, there is a slight increase in the use of three-word bundles as the level increases over time. The three-word bundles accounted for 5.03% in C1 and 4.49 in B2 sub-corpora, whereas the four-word bundles accounted for less than 1% in both sub-corpora. The degree to which the differences between the sub-corpora are significant was calculated using the chi-square test, as displayed below.

Table 3.16. Chi-square test distribution of the frequency of LBs.

	3-word	**4-word**
Observed value		
B2	1498	179
C1	1679	169
Expected value		
B2	1511	1635
C1	1665	182
Chi-square P < 0.05	c^2= 0.119205733	

Although there was an increase in the use of three-word bundles at C1 level and a decrease in four-word bundles over time, the chi-square value at 0.119 shows that the difference between B2 and C1 is not significant. This finding does not support the idea that there is a direct proportionality between the use of LBs and language competence over time.

Moving on to the ten most frequent used bundles across the levels. It can be seen from Table 3.17 that there was variations in the use of LBs at the two levels. The most frequent three-word bundle used at the B2 level was *in order to*, which occurred 239/100,000 words, followed by *as well as*, which occurred 119/100,000 words. By contrast, the bundle *I want to* was the most common bundle in C1 sub-corpus, with 132/100,000 words occurrences, followed by *a lot of*, with 113/100,000 words occurrences. Although the total frequency of three-word bundles increased as students' levels increased, the frequency of the ten most common bundles was higher in the B2 than the C1 level.

Table 3.17. The ten most frequent 3-word bundles in B2, C1 and BAWE. (italic = items occur in 2 sub-corpora) (AF= Absolute frequency; NF = Normalised frequency)

B2			**C1**		
Type	**AF**	**NF**	**Type**	**AF**	**NF**
in order to	12	239	I want to	14	132
as well as	6	119	a lot of	12	113
in fact the	4	79	there are many	11	103
in terms of	4	79	and city life	8	75
one of the	4	79	aim in life	7	66
the lack of	4	79	day by day	7	66
there is a	4	79	in the world	7	66
there is no	4	79	in the end	6	56
a great deal	3	59	the people of	6	56
as a result	3	59	the whole world	6	56

Turning to the most frequent four-word bundles displayed in Table 3.18. It is evident that the occurrence of four-word bundles at both levels was limited. For instance, only three bundles in B2 sub-corpus and 5 bundles in C1 sub-corpus have met the frequency cut-off point and dispersion criteria.

Table 3.18. The most frequent 4-word bundles in B2, C1 and BAWE corpora. (AF= Absolute frequency; NF = Normalised frequency).

B2			**C1**		
Type	**AF**	**NF**	**Type**	**AF**	**NF**
in the case of	3	59	my aim in life	5	47
There is no doubt	3	59	is a lot of	4	37
one of the most	3	59	I would like to	3	28
X	X	X	life and city life	3	28
X	X	X	we are living in	3	28

An interesting point that can be noted from Table 3.17 and Table 3.18 that there were no common bundles between the levels. It seems that students acquired new bundles when they reached a more advanced level.

3.8.6 Structural analysis

Following the same procedure applied in the cross-sectional study for classifying LBs structurally, the bundles identified from the sub-corpora were classified according to four main structural categories as displayed in Figure 3.8 below.

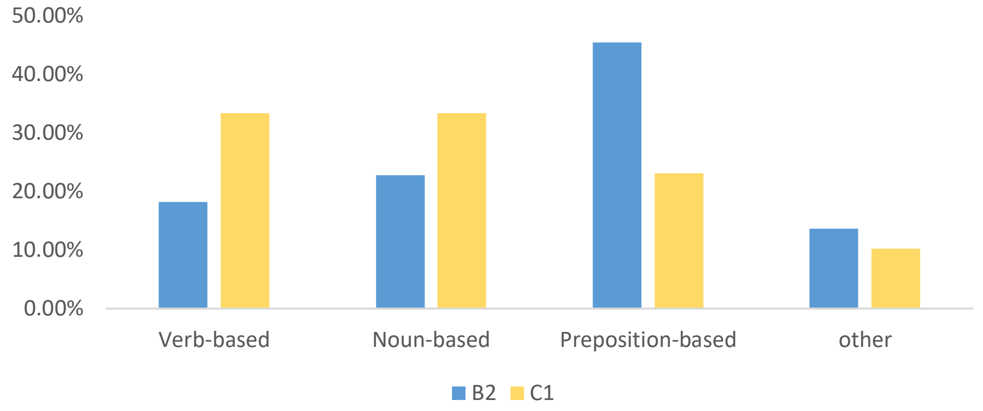


Figure 3.8. Overall distribution of the structural categories across the B2, C1. (Longitudinal study)

It is clear that there was a change in the use of LBs at each structural category over time. One of the most notable differences between the levels is that students at the B2 level relied more on 'preposition-based' bundles, which accounted for 45% of bundles used. These bundles feature either an embedded *of-phrase* to make a logical connection between the elements of an argument (16 bundles), or without an *of-phrase*, representing particular research or discourse context (17 bundles), as shown in the below examples.

- In contrast, *in terms of* using another type of fuel during the operation of producing electricity, Saudi Arabia's consumption of this type was twice as high as the UK. (B2, essay11)

- *In other words*, those people spend most of their time following this social media instead of improving their personal skills. (B2, essay19)

By contrast, students at C1 level used more 'verb-based and noun-based' bundles, accounting for 33% in both categories. Examples for the use of Noun-based bundles and verb-based bundles:

- We are indeed losing *a lot of* languages. One language expert estimates that 60%-80% of all languages will disappear in 100 years, just three generations from now. (C1, essay 63)
- *I would like* to conclude my essay by saying that no life is bad as Almighty Allah gives you this life. (C1, essay 22)

Turning to the distribution of LBs structural sub-categories, displayed in Table 3.19. LBs found in the C1 sub-corpus vary far from those in B2 sub-corpus in terms of their structure distribution.

Table 3.19. Distribution of the structural sub-categories across the groups. (Freq = Freuency; % = precentage within-sub-corpus)

Types	**Sub-types**	**B2**	**C1**
		Freq (%)	Freq (%)
Verb based	Anticipatory it + verb / adjective	0	0
	Copula be + noun /adjective phrase	1 (5)	1 (3)
	Pronoun/NP + be	3 (14)	5 (13)
	First person pronoun + dependent clause	0	5 (13)
	(verb/adjective +) to-clause	0	1 (3)
	Other verb phrases	0	1 (3)
	Totals	4 (18)	13 (33)
Noun based	NP with of-phrase	5 (23)	5 (13)
	NP with other post-modifier	0	2 (5)
	Other noun phrases	0	6 (15)
	Totals	5 (23)	13 (33)
Preposition based	Prepositional phrase of-phrase	1 (5)	0
	Other prepositional phrase expressions	9 (41)	9 (23)
	Totals	10 (45)	9 (23)
Other	Other structures	3 (14)	4 (11)
Totals	12	22	39

The most salient findings are as follows. First, 'other prepositional phrase expressions' was the most prevalent sub-category of LBs in both B2 and C1 essays, accounting for 40% and 23% of total bundles, respectively. The high proportion of 'preposition-based' bundles might also have resulted from the nature of the argumentative essays, which require students to use different prepositions to make connections between the elements of an argument. Second, when learners' level increased, ESL learners are favoured using more 'Verb-based' and 'noun-based' bundles in their writing.

Further analysis of the data reveals that ESL learners' at B2 levels underused four bundle sub-categories the 'anticipatory it + verb bundles, First-person pronoun + dependent clause, (verb/adjective +) to-clause and NP with other post-modifier. These bundles are important as they explain to the reader how the sentence or the idea should be understood. Thus, it seems that advanced ESL learners are more confident in using various LBs than lower-level learners.

To provide statistical evidence for differences between the levels, the chi-square test (and standardised residuals) was used to determine whether differences between the structural sub-categories can be regarded as significant, as shown in Table 3.20.

Table 3.20. Standardized residuals (R) in a chi-Square contingency table for structural distribution. (types).

	Corpora	
	B2	C1
Verb-based		
• Count	80	123
• Expected	110	93
• R	*-4.95*	*4.95*
Noun-Based		
• Count	100	123
• Expected	121	101
• R	*-3.36*	*3.36*
Preposition-Based		
• Count	200	85
• Expected	155	130
• R	*6.64*	*-6.64*
Other		
• Count	60	38
• Expected	53	45
• R	1.44	-1.44
Chi-square *P*< 0.05	df = 3, *P*-value = 2.53E-12	

The statistical test comparing the number of LBs from different structural categories in the B2 and C1 sub-corpora indicated a significant difference between the two groups, with a chi-squared value of 57 and df of 3, that far exceeded the value required for the highest significant P-value is < 0.00001. This result supports the earlier findings that there are differences between the two groups in terms of the grammatical structures of LBs. Further analysis using the standardised residuals (R) that compared between the observed and the expected values of each cell was applied only to the cells with a value greater than ±2.8. It can be seen that there were significantly more Verb-based and noun-based bundles in the C1 sub-corpus, and a significant increase of preposition-based bundles in B2 writing. Hence, the high-level ESL learners (C1) used LBs differently from the low proficiency levels learners (B2).

3.8.7 Functional taxonomy of lexical bundles

The results presented in the previous section showed that ESL learners favoured using particular grammatical structures in their writing. This section presents the functional distributions of the identified bundles in B2 and C1 sub-corpora using following the procedure applied in the cross-sectional design (Section 3.8.4). Figure 3.9 compares the correlations among the three functional categories across the sub-corpora.

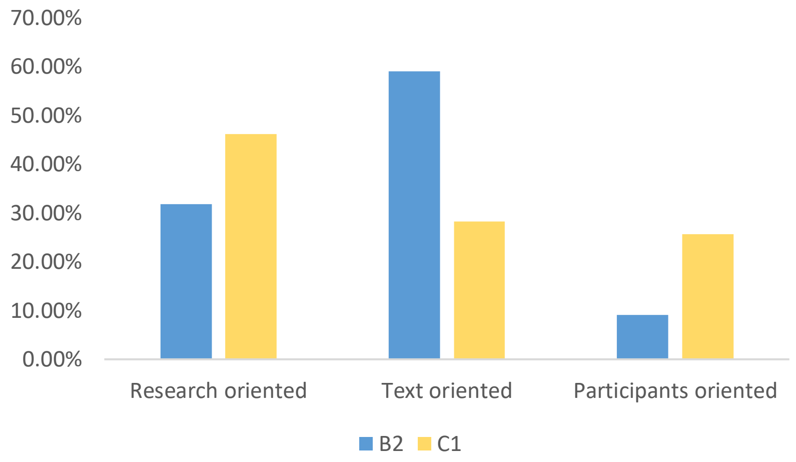


Figure 3.9. Overall distribution of the functional categories across the sub-corpora. (Longitudinal study)

Once again, the two sub-corpora varied widely in terms of the use of the three functional types. The results show that the text-oriented category had the highest value at B2 level, accounting for 59% of the total bundles. This type of bundle organises the text and is concerned with its meaning, argument, or message. On the other hand, ESL learners at C1 sub-corpus change their reliance on more research-oriented (e.g., *in the present study*, *the purpose of*) and participants-oriented bundles. The increased use of

‘Research-oriented’ bundles helps writers to structure their experience of real-world activities, thus, ESL learners at C1 levels tended to highlight the research rather than its presentation. In addition, ‘Participants oriented’ bundles are used to express author attitudes or assessments of another proposition.

- Also, the government should find solutions for their problems, such as offering many jobs opportunities *in order to* increase income. (B2, essay8, text-oriented)
- In addition to those benefits, the development of computer technology brings *a lot* of money to the country. (C1, essay116, research-oriented)
- Secondly, *it is difficult* to imagine in advance how new technology can be used, or misused. (C1, essay44, participant-oriented)

The frequency of each type was calculated across the sub-corpora, as displayed in Table 3.21.

Table 3.21. Distribution of the functional sub-categories across the groups. (Freq = Freuency; % = precentage within-sub-corpus)

Types	Sub-types	B2	C1
		Freq (%)	**Freq (%)**
Research-oriented	Location	0	6 (15)
	Procedure	0	2 (5)
	Quantification	5 (23)	4 (10)
	Description	2 (9)	5 (13)
	Topic	0	1 (3)
	Totals Research-oriented	7 (32)	18 (46)
Text-oriented	Transition signals	5 (23)	3 (8)
	Regulative signals	2 (9)	4 (10)
	Structuring signals	0	0
	Framing signals	6 (27)	4 (10)
	Totals Text-oriented	13 (59)	11 (28)
Participant-oriented	Stance features	2 (9)	10 (26)
	Engagement features	0	0
	Totals Participant-oriented	2 (9)	10 (26)
Overall Totals	10	22	39

As Table 3.21 shows, ESL learners used various functional sub-categories at both levels, with a variant use at C1 level. The change between the two sub-corpora was mainly attributable to five particular function sub-categories: location, stance, transition, framing, and quantification, with the first two being more dominant in the C1 sub-corpus, and the other three in B2. The remaining three sub-categories were of similarly negligible proportions (less than 5% in both sub-corpora).

Possibly more interesting, however, is the diversity of functional sub-categories used by the ESL writers as their level increased. It can be seen that B2 students used five out of 11 different function sub-categories; by contrast, learners at C1 level used nine sub-categories. A comparison of the two sub-corpora in terms of research-oriented bundles reveals a number of key differences between the levels. Among the sub-categories of research-oriented bundles, quantification bundles showed the greatest change, accounting for more than twice as bundles identified in C1 sub-corpus. On the other hand, the location sub-category was not found in B2 but frequently occurred in C1 sub-corpus. In regard to the text-oriented bundles, students at the B2 level were characterised by heavy use of framing sub-categories, which accounted for 27% of the total bundles in the B2 sub-corpora. Although the least frequent bundles in both sub-corpora were in the participant-oriented category, it can be see seen that C1 writers relied (25%) on stance bundles, which represented the highest proportion in all sub-categories in the C1 sub-corpus. These types of bundles convey the writers' attitudes and evaluations (Jalali et al., 2014).

The chi-square test was used to determine whether the differences between the groups were significant, as shown in Table 3.22.

Table 3.22. Standardized residuals (R) in a chi-Square contingency table for functional categories.

Functional categories	**B2 sub-corpus**	**C1 sub-corpus**
Research-oriented		
• Count	140	170
• Expected	169	141
• R	-4.18	4.18
Text oriented		
• Count	260	104
• Expected	198	166
• R	8.77	-8.77
Participants -oriented		
• Count	40	94
• Expected	73	61
• R	-6.26	6.26
Chi-square ***P*** **< 0.05**	df = 2, *P*-value = 2.35294E-19	

If the residual is less than -2.8, the cell's observed frequency is less than the expected frequency. Greater than two and the observed frequency is greater than the expected frequency.

It can be seen that a significant increase in "Research-oriented" and "Participants-oriented" bundles in C1 writing, whereas a significant increase in 'Text-oriented" bundles in B2 writing. These findings reflect on the change that occurred across the levels.

While this study focuses on the development use of LBs in relation to CEFR levels, it is not a comprehensive investigation into what factors might affect students' use of LBs. It must be noted that this study used only one genre of academic writing (i.e., ELC students' essays). However, the results of this study have demonstrated some overlap between ESL learners and BAWE writers, and considerable differences and changes in terms of the frequency, structures, and functions of LBs. The pilot study can be considered successful, as it supports the idea that students' essays can be used to measure their academic writing performance.

3.9 Summary of the study

The analysis of the forms, structures, and functions of the target LBs revealed a number of notable differences between the B2 and C1 sub-corpora. The research questions aimed to determine the degree to which the use of LBs is related to academic performance. The pilot study followed a specific procedure to answer the research questions.

RQ1 What are the most frequently used three- and four-word LBs in B2 and C1 sub-corpora?

The first question posed in the pilot study sought to determine the most frequent three- and four-word LBs in B2 and C1 academic writing. To achieve this, *WST* was used to provide lists of the most common words in both sub-corpora. The results of the pilot study revealed that three-word bundles were the most common bundles at both levels, accounting for 84% of the bundles used in both sub-corpora. Therefore, it can be concluded that ESL learners have a tendency to employ a higher number of three-word than four-word bundles, particularly amongst low-level students. This finding confirms Alipour and Zarea (2013) results in their assessment of the corpora of native and non-native English language, where they found that three-word bundles were the most frequent in both corpora. A possible explanation for the overuse of three-word bundles might be related to the complexity of their production, causing language learners to avoid using them in their writing, as it requires effort and time for students to produce longer sequences than shorter ones. However, this result was not surprising as Biber et al. (1999, p.992) state that three-word LBs are extremely common because they are “a kind of extended collocational association,” while longer bundles are “more phrasal in nature and correspondingly less common.”

Another interesting finding is that the bundle *on the other hand* was the most frequent bundle in the B2 and C1 sub-corpora. This could be because that ESL learners wanted to draw special attention to the different points of view in their argumentative essays and so the learners drew attention through emphasis placed on the LBs. This result supports evidence from previous observations that it is a common bundle that most ESL learners used in their writing.(e.g., Biber and Conrad, 1999; Hyland, 2008b; Römer, 2009; Nkemleke, 2012)

Concerning the overall comparison between the corpora, the study found that B2 level learners employed bundles more frequently than C1 level learners. Moreover, when compared individually with the LBs identified in the BAWE corpus, the results revealed that the B2 sub-corpora shared two out of the ten most frequent three-word bundles and three of the ten most frequent four-word bundles with the reference sub-corpus. In contrast, the C1 sub-corpora shared only one bundle with the reference sub-corpus in both lists. Moreover, eight out of the 50 most common LBs in the B2 and C1 sub-corpora were identified in the BAWE corpus. These results conflict with those reported by Chen and Baker (2010), who found many LBs shared in both native and non-native academic writing.

Surprisingly, few of the most frequent bundles in the BAWE corpus were found in the ESL learners’ corpora. Therefore, even if the B2 level students used more LBs than C1 students, certain bundles were new and used by only a few learners with repeated the same bundles more than once in their essays. For example, the bundle *on the other* was identified 19 times in the B2 sub-corpora (although one student used it three times in one text). The findings of the current study are consistent with those who found that learners overuse certain bundles and repeat them more than once in a single paper. A possible explanation for this is that ESL learners tend to use certain LBs as items of high frequency to reflect a high level of formality and demonstrate their language competence; or, they may still be in the process of learning additional LBs. Furthermore, even if it is assumed that ESL learners can proficiently use more bundles in their writing, they can then forget lexical coherence, which makes their writing vague and confusing to readers. Therefore, the use of different LBs in academic writing may not be an indication that meaning is conveyed.

RQ2 What distinguishing features a keyness analysis revealed about LBs identified in B2 and C1 sub-corpora in comparison with the BAWE sub-corpus?

The second research question was designed to determine the significance of the use of LBs in B2 and C1 according to the keyness values for the three- and four-word bundles identified in both sub-corpora, and compared these with reference sub-corpus using the log-likelihood statistical test. The data identified in this section provides some evidence for the common assertion in past studies that ESL learners favour particular bundles and overuse them in their writing. Therefore, it seems that low-level students are more likely to rely on the use of LBs than C1 students, and accounted for more instances in their writing. For instance, nine significant keybundles were identified in the B2 sub-corpus, whereas two keybundles were found in the C1 sub-corpus. This result might have been affected by the small dataset used in the pilot study.

Closer inspection of keyness analysis lists of the sub-corpora revealed that L2 learners overuse some connector expressions in their writing. For example, the bundle *on the other hand* had an extremely high keyness value in the B2 and C1 sub-corpora. Several studies have shown that L2 learners tend to overuse the bundle *on the other hand* in their writing (Biber and Conrad, 1999; Römer, 2009; Chen and Baker, 2010). The result also confirms previous research findings that L2 learners prefer to use connector expressions and repeat these in their writing (e.g., Hyland, 2008a; Lee and Chen, 2009; Römer, 2009).

A possible explanation for the result is the students' good understanding of functional meaning. Karabacaka and Qinb (2013) compared the LBs used by Turkish, Chinese, and American university students and found that 24 bundles were observed in American students' papers that did not occur in the Turkish or Chinese students' papers, and that 78 bundles that occurred in the Turkish and Chinese sub-corpora were not used by the American students. The first explanation put forward for this observation was the requirement of different bundles for different topics. The second reason proposed was that the native students preferred those bundles that were more practical and familiar to them. The latter reason is better able to explain the observations in the present study: That both B2 and C1 learners rely on bundles that are familiar to them, and to which they are commonly exposed in their reading. This point should be considered when teaching LBs to L2 learners.

RQ3 How do LBs in B2 sub-corpus differ from C1 in terms of structures and function?

The third research question was concerned with the structures and functions of LBs in the B2 and C1 sub-corpora. The results showed variations in the use of LBs in terms of structural classification, there were differences in the use of LBs between the ESL sub-corpora and the reference sub-corpus. It was clear that ESL learners used more phrasal bundles than clausal bundles in their writing. This result strongly supports the findings of previous studies (e.g., Biber et al., 1999; Cortes, 2002; Cortes, 2004; Jalali and Zarei, 2016) that academic writing is different from other registers, such as conversation and classroom teaching, relying heavily on phrasal rather than clausal bundles.

More specifically, verb-based bundles were the most frequent three- and four-word bundles found in the B2 and C1 sub-corpora. Among the two CEFR levels, the C1 level showed the highest proportions of verb phrase bundles, at 53.4%, while the B2 level had a lower proportion, at 40.5%. The results of the present study suggest that the language of ESL writing contains more conversational bundles; this accords with Wei and Lei (2011), who found that Chinese ESL learners preferred VP fragments. This may be because ESL learners are exposed more to listening than reading in their studies, they might repeat the language they hear from native speakers or experts to demonstrate language competence.

Comparing the identified LBs across the levels, it can be concluded that the three groups employed different proportions of most of the structural sub-categories, except for the ‘preposition-based’ category. Statistically, the chi-square test revealed a significant difference among the sub-corpora in the three structural categories. For instance, the results showed that C1 writers significantly overused “Verb-based” and “Noun-based” bundles compared to B2 which shows underuse of these bundle types, which supports the idea that C1 students rely more on spoken language in their writing. On the other hand, B2 writers significantly overused the “Preposition-based” bundles. The ‘preposition-based’ category is usually used to show a logical relationship between prepositional elements, which means that ESL learners at B2 level can use this type of LBs to link between the different ideas of the argumentation. The rationale for using these bundles could be related to L1 influence; for instance, Allen (2009) states that there could be similar bundles in students' mother tongues that led to the overuse of specific bundles. The pairwise comparison using the structural taxonomy showed that ESL students used a large proportion of verb-based bundles.

The frequency and structures of LBs provide some information about the similarities and differences between the levels. Therefore, the functional distributions of LBs were also examined to provide a clearer understanding of learners' language use at different CEFR levels. As presented in section 3.8.4, there was a similarity in the use of functional categories across the levels. The most frequent functional category was research-oriented bundles, followed by participant-oriented and text-oriented. The increased use of research-based bundles in the B2 and C1 sub-corpora might be due to the fact that, in argumentative essays, students need to describe various aspects and provide different justifications for their ideas to the readers. Although C1 writers used significantly more research-oriented bundles than B2 writers, bundles of this function accounted for more than 40% of all bundles in the corpora. This result is similar to previous studies that have found that academic writing is dominated by research-oriented bundles over other categories (Biber et al., 1999; Biber et al., 2004; O'keeffe et al., 2007; Chen and Xiao, 2015; Jalali and Zarei, 2016). The high proportion of research-oriented bundles might be a result of focusing on the procedure and describing the problems in the argumentative essay rather than its presentation. Indeed, Hyland (2008b) argues that students focus on research methods, practice, and instruments used, which enable them to emphasise demonstrable generalisation.

Comparing the functional sub-categories across the level. The result showed that ESL learners mostly used bundles for quantification and description. The former sub-category consists of expressions that describe an amount or number, while the latter details the qualities of the texts (Hyland, 2008b). In regard to the participant-oriented category, stance bundles were the most frequently used sub-category across the levels, representing approximately 24% of all bundles in the ESL sub-corpora. The high proportion of these bundles shows that ESL referring to personal ability and personal intention, as they are important in conveying their ideas. When the proportion of each category is considered, it can be concluded that B2 students relied equally as much as

C1 students on most of the functional categories. Statistically, however, the study has demonstrated significant differences in functional distributions between the levels.

RQ4 To what extent does the use of LBs correlate with learners' level of proficiency?

The fourth research question was concerned with whether there was a relationship between the learners' use of LBs and their academic performance over time. What is interesting about this result is the development in the use of three-word bundles across the levels. The result showed an increase in three-word bundles at the higher level at 1,498/100,000 words, three-word bundles were found in the B2 sub-corpus, and 1,679/100,000 words occurred in the C1 sub-corpus. However, there was no increase in four-word bundles at the high-level. The results of the pilot study provide evidence that suggests there may be developed in the use of LBs across the levels over time, but not to a statistically significant degree. This might be due to the number of essays collected to make the sub-corpus and the short period of time the learners were tracked. The result was similar to Li and Schmitt (2009), who found no correlation between learners' use of LBs and language performance, as the data did not reveal any consistent increase in the frequency and diversity of lexical phrases.

Structurally and functionally, there was much variability in terms of the structures and functions of LBs across the levels. High-level ESL learners used a greater variety of structures and functions in their writing than lower-level learners. However, the chi-square result did not show clear evidence of the relationship between the use of LBs and proficiency level.

3.10 Conclusion

The pilot study aimed to realise two major research objectives related to the use of LBs in corpora of ESL learners from different CEFR levels. The first objective was to compare the use of LBs between the B2 and C1 levels in order to identify the difference between the two levels. The second objective set out to trace the possible improvement in the use of LBs across the CEFR levels. Based on a corpus-based approach, this pilot study found that LBs are widely used in ESL learners' academic writing; their high frequency demonstrates their importance across the levels. However, the analysis shows differences and similarities between CEFR levels in terms of form, structures, and functions of LBs.

Although students at both levels did make significant use of LBs in their writing, the B2 students used LBs more frequently than C1 students. A major finding from the analysis was that, generally, ESL learners favoured signalling bundles in their writing, and three-word bundles were the most frequent in the ESL sub-corpora. Moreover, the most frequent bundle in both sub-corpora was *on the other (hand),* which is consistent with the findings of previous research (e.g. Hyland, 2008a; Römer, 2009; Nkemleke, 2012). The structural and functional analysis of the identified bundles revealed that there was much variability as the level increased, in terms of the structural and functional use of the bundles. Functionally, there was a high prevalence of research-oriented bundles in ESL learners' essays, particularly for the high-level (C1) students; these help writers to structure their activities and experiences of the real world. A large majority of these functions were fulfilled through bundles beginning with prepositional phrases. It was surprising to find that low-level students acted as professional writers from a functional perspective. Another key finding is that the language used in C1 sub-corpus is more similar to spoken language, as they overused verb phrase bundles. Finally, the longitudinal study did not show any significant development of the use of LBs across the levels though there was an increased use of three-word bundles in the high-level sub-corpus. However, significant progress was found in the variability of the structures and functions of LBs, since C1 writers were found to have used as many different structures and functions as professional writers in their academic writing.

3.11 Amendments to the methodology for the main study

This pilot study was a pre-study for the more comprehensive main study to examine the modes of identifying LBs in two selected CEFR levels. Although the analytical procedures used to answer the research questions were somewhat successful, some changes are required to achieve a better result in the main study. These are as follows:

1. The random samples of the BAWE corpus used as reference corpus proved useful in comparison with the sub-corpora. However, although the BAWE corpus is also an academic writing corpus, topic-specificity of certain expressions may lead to unexpected statistics. To be more consistent, the BAWE corpus will be used as reference corpus but with strict control over the data, and the samples for the

comparison will be extracted from one discipline (arts and humanities), which is broadly equivalent to the ESL learners' essays in terms of the language used.

2. The comparison between two CEFR levels, B2 and C1, was interesting, but there were minimal differences between the levels. It will be useful in the main study to compare the B1, B2, and C1 levels to trace the variation and the development of LBs across these levels.
3. Although the ESL learners' sub-corpora showed some significant differences between CEFR levels, the corpus size effect on the variation and development of LBs used across the levels, thus, in the main study, the data collection process will be the same, but a larger volume of data will be collected.

4 Methodology

4.1 Introduction

The pilot study chapter provided the preliminary methodology adopted for this research. Based on its results, a number of recommendations were made for methodology employed for the main study (See section 3.11). This chapter describes the final analytical procedures used to collect and analyse the data gathered to address the research questions (RQs) provided in section 1.5, and the revisions made to them following the pilot study.

The first section of this review highlights the significant role of corpus studies in formulaic language research and explores types of corpora (e.g., general, specialized, learner, and comparable) (Section 4.2).The chapter then commences by providing a description and rationale of the research design applied in this study (Section 4.3 and 4.4), followed by a detailed discussion of the data collection process employed for the English as a Second Language (ESL) learners' sub-corpora (Section 4.5). The next section concerns the identification of the lexical bundles (LBs) in the target sub-corpora, and the exploration of the keybundles (Section 4.6), followed by a description of keyword extraction (Section 4.7). What is next is a discussion of the analytical framework used for investigating these features (Section 4.8). Finally, section 4.9 discusses the corpus construction rationale, followed those limitations of research methodology, which might influence the result of the study.

4.2 Corpus Linguistics

4.2.1 An introduction to corpus linguistics

Corpus linguistics (henceforth, CL) is a method used to analyse language on the basis of computerised corpora (Mcenery and Wilson, 1996). It is not a separate branch of linguistics (e.g., sociolinguistics); rather, it is a methodology that can be applied in almost any language study using corpora. CL has enriched the fields of linguistics and learner language through the analysis of frequencies, functions, and contexts of words or phrases in learner language (Biber et al., 1999; Cortes, 2004; Hyland, 2008a; Chen and Baker, 2016; Liu and Chen, 2020). For example, it makes an empirical analysis of language possible, and so has enabled researchers to uncover patterns of language usage that had previously remained hidden from view (Breyer, 2011). On the importance of corpus linguistics over other methods of linguistic analysis, Thomas and Short (1996, p.248-259) point out, "When language is actually used, it is for communicative purposes in social situations… The reality is that by starting with real texts, corpus linguistics has the potential to develop a new kind of linguistics with a much better theoretical foundation than has hitherto been the case." CL has thus become common in many fields, such as discourse studies (e.g., Biber and Barbieri, 2007), genre analysis (e.g., Burgess and Cargill, 2013), and – most relevant to this study – learner language (e.g., Reppen and Olson, 2020), to mention but a few.

CL has developed considerably due to two important events: first, the availability of large and varied corpora; second, the development of computer software (concordancers), where the processing speed, the storage capacity of electronic data, and the structured compilation of written or spoken languages has made it possible to retrieve data easily and quickly and presented in a format ready for analysis.

The evolution of computer technology has played a central role in corpus development. Hunston (2002) states that "a corpus does not contain new information about language, but the software offers us a new perspective on the familiar." The computer software used by linguists enables practitioners and learners to process and organise large amounts of data and describe it in more detail by searching for, retrieving, sorting, and analysing linguistics data and with a degree of accuracy that would not be possible if undertaken manually. Most available computer software have two core functions; they sort the linguistic data in a corpus so that it can be analysed by practitioners and calculate statistical data about the items in the corpus. Sorting and organising the data in a corpus can be done in three ways: through wordlists, concordance, and phraseology. For example, *WordSmith Tools* (WST) (Scott, 2012), which is widely used in corpus linguistics research, allows practitioners to perform various analytical functions such as:

- Identifying wordlists and corresponding frequencies
- Sorting words in concordance lines and their collocations
- Keyword/keyness analysis

These three functions will be discussed in detail in section 4.5.8, as they are core functions used in this thesis.

Briefly, the information provided in the corpus, regardless of its size, provides a basis for making observations about language use. Corpora show how language is routinely used on a daily basis and can reveal rare or exceptional cases that are not apparent from looking at a single text. These exceptional cases can then be analysed to understand how language is used in a particular register (Granger, 1998a). For the purpose of this thesis, CL will be seen as the method of analysing and theorising language that can be done by examining amounts of real, empirical data, using advanced computerised software tools.

The advancement of computer software has enabled researchers to analyse formulaic language quantitatively and qualitatively, which is essential for theorising purposes. For example, previous research has used computer software (e.g., WordSmith, Antconc) to retrieve lists of the most frequent words or phrases within a corpus and quantitatively count how many times these items occur. This is followed by a crucial part of the corpus-based approach using concordancing to enable researchers to make a qualitative interpretation of patterns of language in context. This makes it possible to better understand the meaning(s) and the usage of a word or a multi-word unit to confirm previous hypotheses or acquire new knowledge about learner language (e.g., overuse or underuse of items in a specific text).

In short, CL is more than retrieving language data through the use of computer software; it is a method used to examine and analyse the data retrieved from a corpus. The main task of CL using computer software is not to find the data but to analyse it in use. It is considered a reliable source of authentic information about language. It provides a rapid method of processing and searching language data, as well as offer an access to and the ability to manage huge amounts of data (Kennedy, 1998). Therefore, the use of corpus linguistics has made it possible for researchers to access and discover huge amount of texts comprising millions of words in order to examine the structure and use of language (Scott and Tribble, 2006).

4.2.2 Types of corpora.

The term ‘corpus’ is a singular word rooted in Latin and meaning ‘dead body’, with the plural being ‘corpora’. It was first encountered in the 6th century when it was used to refer to a collection of legal texts, *Corpus Juris Civilis* (Francis, 1992, p.17). However, this definition is not entirely satisfactory for corpus linguists. According to one of the five definitions provided by the Oxford English Dictionary, linguists use corpus to refer to a 'body of text'(Simpson and Weinert, 2013). In applied linguistics, the word corpus refers to a collection of texts (Mcenery and Wilson, 1996), which can be read and analysed using a concordancer (computer software), to rapidly and reliably search through the collection of words. Corpora vary in size but are usually large and built to serve different purposes. Furthermore, they can be comprised language taken from written texts, transcribed speech, or both.

Sinclair, one of the pioneers of modern corpus linguistics, defines the term corpus as "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (2005, p.16). Biber et al. (1999) also used the term 'corpus' to refer to any collection of more than one natural text (written and/or spoken) put together to represent language in general, accessed by sophisticated computer tools (concordancers) which allow researchers to read, search, and manipulate the data. A corpus is a remarkable thing, due to the characteristics that it acquires if it is well-designed and carefully constructed. According to Mcenery and Wilson (1996), a modern corpus has four major features: sampling and representativeness, a finite size, machine-readable form, and a standard reference. Moreover, corpora should be designed for a specific purpose that helps researchers to explore language and provide valuable answers to research questions (Jones and Waller, 2015).

There are many different corpora, depending on the purpose of the study, written and/or spoken, small or large, modern or old, with data from one language or several languages. In terms of corpus size, corpora can be varied because they are built to serve various purposes (Mcenery and Wilson, 2001). Below is a brief description of different corpus types.

- General corpora: (e.g., British National Corpus [BNC][2]), which consist of texts of many types and comprise as wide a range of texts as possible. It may include written or spoken language, or both, and may include texts produced in one country or many" (Hunston, 2002, p. 14). Unlike specialized corpora, general corpora are intended to represent the state of language as a whole; thus, these

[2] www.natcorp.ox.ac.uk.

corpora include a wide range of texts from many different language situations. These types of corpora are made up of hundreds of millions of words from a variety of sources such as news, conversations, books, movies, journal articles, and more.

- Comparable corpora: (e.g., the International Corpus of English (Greenbaum, 1991) consists of two or more corpora of the same language or different languages. They have been built up in a similar way so that language features can be compared and contrasted across the languages/varieties.
- Parallel corpora (e.g., the Minority Language Engineering Project (Singh et al., 2000)), which contains parallel aligned Panjabi-English texts), containing two or more corpora in different languages translated from one language to another. These corpora are closely related to comparable corpora, they play an important role in examining the differences between languages.
- Monitor corpora: (e.g., the Bank of English corpus (Cobuild, 1992)), which are expandable to track changes in a language. This corpus used to trace the development of aspects of language over time" (Hunston, 2002, p.15). This type of corpus helps researchers examine changes in language in long diachrony.
- Specialised corpora: (e.g., Michigan Corpus of Spoken English (Simpson et al., 1999), a vary in their size according to the breadth of the language use domain that these corpora represent, but usually small corpora comprising texts of a specific type of language, and not intended to be representative of the language as a whole. It is often collected by researchers independently and aims to represent target language use in specific genre O'keeffe et al. (2007). This is the main corpora type used in the present study, the rationale for using a specialised corpus will be discussed in section 4.9.2.

4.3 Research design

The present study employs both quantitative and qualitative techniques to identify and analyse the frequency, structures, and functions of LBs in ESL learners B1, B2 and C1 argumentative essays (Figure 4.1).

- Three written data sets were compiled: cross-sectional data obtained from ESL learners, longitudinal data from the same population, and finally data from proficient student writers (BAWE corpus);
- *The Manual for Relating Language Examinations to the Common European Framework of Reference for Languages (Europe, 2003),* was used to determine the relevant CEFR levels, and to assign the essays to one of three ESL learner levels, B1, B2, and C1;
- A corpus-based analysis (frequency, keywords) was then employed to examine the variation in the use of LBs between ESL learners' levels;
- The structural classifications provided by Biber et al. (1999) were used to guide the structural analysis of all the data;
- The functional taxonomy provided by Hyland (2008b) was used to examine the functional distribution of all the data.
- Performing keyword analysis which serves two aims. First, bundles that are identified as dominant on all lists of keywords are doubly legitimized as significant for the data. Second, a comparison between the three lists of keybundles helps identify distinguished characteristics of the ESL learners' levels in the written discourse.

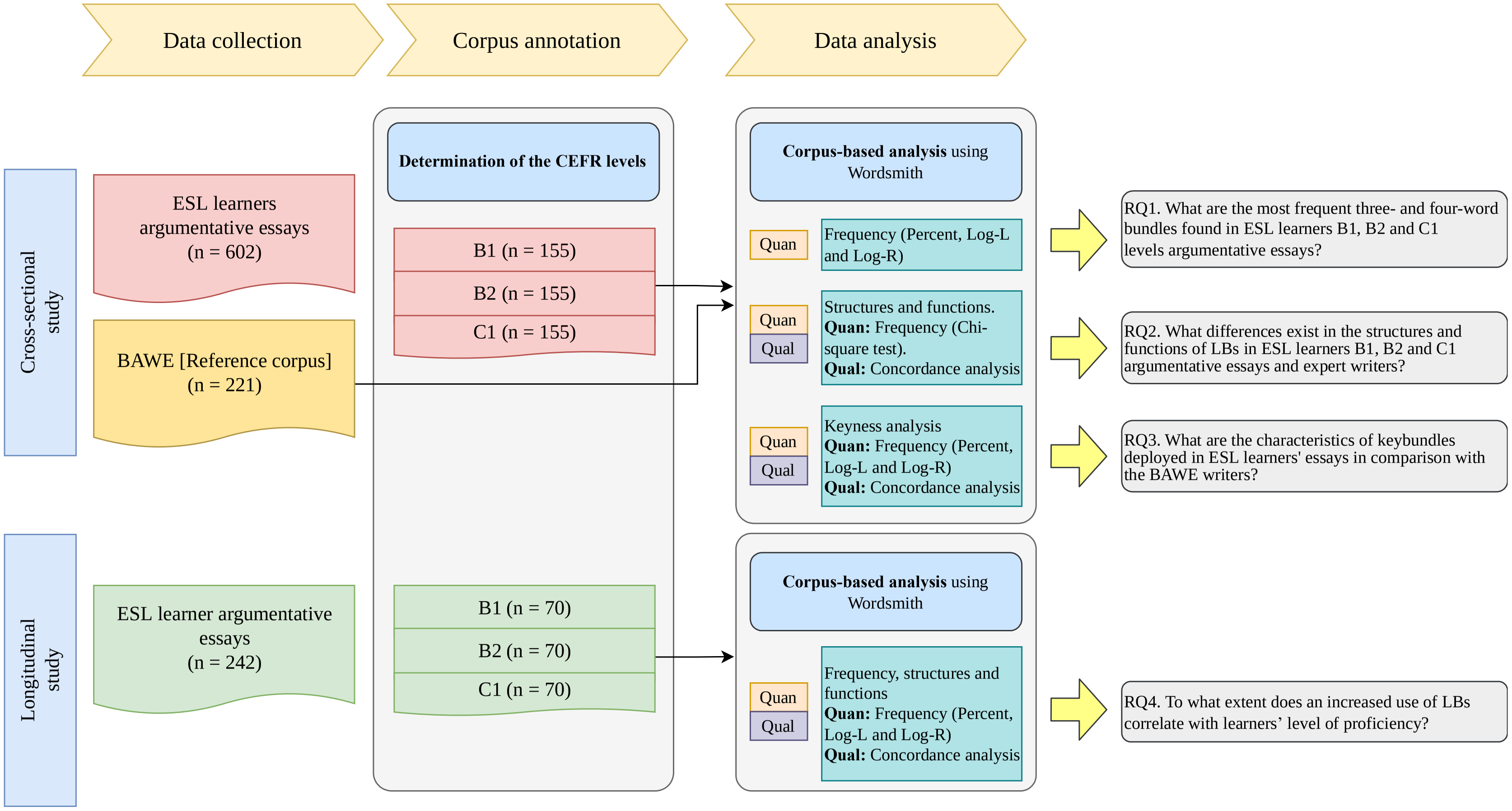

Data collection
Corpus annotation
Data analysis
Cross-sectional study
Longitudinal study
ESL learners argumentative essays (n = 602)
BAWE [Reference corpus] (n = 221)
ESL learner argumentative essays (n = 242)
Determination of the CEFR levels
B1 (n = 155)
B2 (n = 155)
C1 (n = 155)
B1 (n = 70)
B2 (n = 70)
C1 (n = 70)
Corpus-based analysis using Wordsmith
Quan
Frequency (Percent, Log-L and Log-R)
Quan
Qual
Structures and functions.
Quan: Frequency (Chi-square test).
Qual: Concordance analysis
Quan
Qual
Keyness analysis
Quan: Frequency (Percent, Log-L and Log-R)
Qual: Concordance analysis
Corpus-based analysis using Wordsmith
Quan
Qual
Frequency, structures and functions
Quan: Frequency (Percent, Log-L and Log-R)
Qual: Concordance analysis
RQ1. What are the most frequent three- and four-word bundles found in ESL learners B1, B2 and C1 levels argumentative essays?
RQ2. What differences exist in the structures and functions of LBs in ESL learners B1, B2 and C1 argumentative essays and expert writers?
RQ3. What are the characteristics of keybundles deployed in ESL learners' essays in comparison with the BAWE writers?
RQ4. To what extent does an increased use of LBs correlate with learners’ level of proficiency?

Figure 4.1. Summary of the research design employed in this study.

The data analysis was divided into two stages. The first stage was a cross-sectional study of the argumentative essays of ESL students at three levels, B1, B2 and C1. The subsequent stage was a longitudinal study based on the same population, but using students undertaking a long-term language course, progressing from intermediate to advanced levels. Following the data collection, the ESL learners' essays from both stages were classified according to the Common European Framework of Reference for Languages (CEFR) levels, using *The Manual for Relating Language Examinations to the Common European Framework of Reference for Languages* (Council of Europe, 2003). Employing the relevant determination criteria, three learner sub-corpora representing CEFR levels B1, B2, and C1 were established. The essays were then retyped into an electronic format, and then cleaned so that no titles were included, and then saved as a plain text, to be used for the analysis with the help of concordance software.

For the first stage, namely the cross-sectional study, *WordSmith* (*WST*) were used to generate the LB lists and keyword lists for the first set of data. The LBs were identified from the ESL learners' sub-corpora, and their overall frequency was calculated, then they were compared across the ESL learners' levels, and with the British Academic Written English (BAWE) sub-corpus. This was followed by the calculation of inferential statistics, namely log-likelihood, to test whether there was a statistically significant difference among the frequently found patterns across the sub-corpora. Although this data was considered to be quantitative, as it was guided by numbers and statistics, it required further qualitative examination, by assessing the sample's concordance lines, to determine how the LBs were employed in the context in which they occurred. The LB frequency lists of the ESL learners' sub-corpora were initially generated to identify the top 20 most frequent LBs. In order to address the degree of similarity and difference between the language levels, a comparison of the shared bundles was conducted. In order to compare the students' use of LBs with what may be considered the norm in ESL learners' writing, an analysis of the BAWE writing was undertaken to assess the form, structure, and function of the LBs used by proficient student writers. Each concordance line of the three sub-corpora was then investigated to identify the structural and functional distribution of the words across the sub-corpora.

For the second stage to determine the development of the use of LBs across the ESL learners' proficiency levels, *WST* was employed once again to generate lists of LBs from the longitudinal data. All the LBs identified in the ESL learners B1, B2 and C1 sub-corpora have also coded structurally and functionally. The same techniques applied to the first set of data were utilized to determine the relationship between the use of LBs and proficiency levels. A more detailed description of these analysis procedures is provided in section 4.9.

4.4 Rationale of the research design

In corpus linguistics research, both quantitative and qualitative methods are used extensively, in combination, as researchers generally commence with quantitative analysis, and proceed to qualitative analysis to explain why a specific pattern occurs; Dörnyei (2007) described this as a 'mixed methods' approach. Both qualitative and quantitative analyses have a role to play in corpus linguistics studies. The quantitative analysis identifies the salient features to be explored by a subsequent qualitative discourse. The methods can be combined usefully in a way that enhances the research process, and the results that can be obtained in corpus linguistics research.

Quantitative analysis is deductive in nature, as it includes descriptive statistics, and determines the relationship between two or more variables. It is based on numerical values, which Dörnyei (2007, p.27) explained concerns research analysis that "involves data collection procedures that result primarily in numerical data which is then analysed primarily by statistical methods". In quantitative analysis, features are classified and calculated, and more complex statistical analysis is sometimes conducted to describe what is observed in a comparison, for example between two corpora, as long as representative and valid data has been used. One example of the use of a quantitative analysis in the present study is counting the frequency of LB usage, and their patterns of use in ESL learners' essays, which was then compared with the results of the LBs extracted from the BAWE sub-corpus, with the help of *WST* (Scott, 2012).

In contrast, qualitative research "involves data collection procedures that result primarily in open-ended, non-numerical data which is then analysed primarily by non-statistical methods" (Dörnyei, p.24). It is inductive, exploratory research, that is data-driven, and analyses the data by summarizing, categorizing, and interpreting it. In the qualitative approach, subjective judgment, based on uncountable data, is employed for pursuing an in-depth investigation of linguistic phenomena. An example of the qualitative analysis employed in the present study concerned the structures and function of the LBs identified in each ESL learners' sub-corpora, as well as the interpretation of the language use by means of concordance lines. One of the advantages of the qualitative analysis of LBs lists is to discover the language patterns, such as structural or functional patterns of LBs in a specific register. That will lead to discover how LBs used in each learner's level. Therefore, we conducted a qualitative analysis of LBs to determine the distinctive features of LBs applied in each level.

To conclude, a combination of quantitative corpus linguistic analysis and qualitative discourse analysis can provide a true mixed-methods approach to understanding how ESL learners B1, B2 and C1 exhibit different characteristics of LBs in their argumentative essays. This study employed a corpus-based approach to examine the variation and the developmental use of LBs by B1, B2, and C1 level ESL learners. Its design and methods were primarily quantitative, and sought to identify the number of LB occurrences in the sub-corpora, and to compare this with a reference corpus. A qualitative approach was used to provide a more detailed analysis of the LBs in the context, and to address the issue of the multi-functionality of the target bundles by assessing the concordance lines. The next section discusses the materials and data collection used in this study.

4.5 Materials and data collection

4.5.1 Ethical clearance process

The University requires that all research projects which involve human participants or human tissues, or personal information should receive formal ethical approval before they commence, to guarantee that research will not risk causing any pain or indignity to participants. Before collecting the data, ESL learners were required to sign informed consent forms as part of the ethical process of data collection required for this research.

It requires that participants be provided with as much information as possible about a research project to make an "informed decision" about whether or not they want to take part in the research. The consent forms are information sheets, generally provided in written format but can also be spoken, are an essential part of the informed consent process. Under the ethical clearance requirements for this study, the study follows the University of Liverpool ethical procedures as detailed in the University Policy on Research Ethics (Appendix B). Thus, the researchers need to inform all the participants (the language centre directors, teachers and students) about the research, and they willingly choose whether to participate or not. To increase the number of submitted essays for the present study, I tried to collect texts from ESL learners in a range of UK language centres, and this proved to be both complicated and time-consuming. I have contacted a wide range of language centres around the UK and provided participants with a financial incentive. The ESL learner's corpus was a £ 3300 research project, which took place over six months, with funding to pay £5 per acceptable essay (compromise into the sub-corpora). These learners were chosen as they considered academic writing is the primary skill in the development of academic competence. They had volunteered to write argumentative essays equivalent to the IELTS task2. I used various ways of attracting the participants. These included consulting the teacher in taking their permission to pop in their classes and ask the students if they want to participate in the research. Then, circulating emails for the students through the language centres portal, advertising in open university fairs, hanging posters in the language centres, and handing out flyers at the language centres to attract L2 learners to participate in the study. Finally, contributors received consent applications to decide whether to take apart or not (Appendix C). The candidates were informed that it is possible to withdraw from the research at any time.

4.5.2 Methods of the analysis

The present study investigates the variations and the development use of LBs of different ESL learners’ proficiency levels. Ideally, to compare L2 different proficiency levels, participants need to answer the same writing tasks under specific sitting so that we control all the variables that could affect the quality of their writing except the proficiency. However, it is not easy to have a large number of ESL learners (such as the 632 ESL learners examined in this thesis) write the same samples under the same sitting and track them for a long period. In light of this reason, the present study decided to divide the available data into different studies. The first study is a cross-sectional study deals with writing produces in the academic context which compares ESL learners of different proficiency levels. This study aims to find the similarities and differences between the ESL learners B1, B2 and C1 levels of LBs. The second study is longitudinal which comes under second language development research. Both studies will be discussed in detail in the paragraphs below.

The first study falls within the scope of common learner's corpus research, which compares L2 learners' language. Yet, it also distinguishes itself from most of the literature in the sense that it included argumentative essays produced by ESL learners enrolled in English for Academic Purposes (EAP) courses from three different proficiency levels. Using a cross-sectional design allows the researcher to compare the student's levels of the use of LBs. In the cross-section study, the data are taken at one time and used to explore the relationship between variables and lead to finding new hypotheses for future research (Hua and David, 2008). For example, the study could examine age, gender, levels and the exposure to L2 language concerning LBs usage. Therefore, examining the use of LBs through the cross-sectional data allows for more exploration of the similarities and differences of LBs across the levels. The results might show the distinctive use of LBs in each level, this will reveal the characteristics of LBs across the ESL learners' levels, find out how language correlated to the levels of the students. However, cross-sectional may not provide researchers with specific information about the changes over a specific period as the data are taken as a snapshot in one time. Consequently, the second study uses a longitudinal design to investigate the development and change of LBs usage across the ESL learners B1, B2 and C1 levels. Longitudinal dealing with one or more groups of participants at a different point and usually use a small number of a subject as a result of the data collection nature (Hua and David, 2008). It is believed that longitudinal research provides the researcher information about the linguistics changes at an overall level Wei and Moyer (2009).

This study uses a trend longitudinal study taking the data from the same population in the first stage (cross-sectional) but from a different group of ESL learners. As a study of the way ESL learners' writers transform their proficiency knowledge through writing, this study offers insights into the connections between the use of LBs and language competence in relating to academic writing through different CEFR levels.

4.5.3 Participants and sitting (cross-sectional study)

This section provides a detailed procedure of the process used to build the ESL learners' sub-corpora that form the starting point of the study analysis. Participants of the cross-sectional study consisted of ESL learners from various nationalities who enrolled in English for an academic language course at various language centres in the UK. These learners were coming to the UK to improve their language to complete their higher education or find a job opportunity. The learner corpus is a form of specialised corpora and has been the subject of much research (Durrant, 2015; Moynie, 2018; Nekrasova-Beker and Becker, 2020; Pearson, 2021). While the pilot study data was based on learners from two selected language centres, the primary study data was taken from six language centres around the UK, therefore, applying results to the larger population from ESL learners.

A detailed profile of the nationalities for the participants is provided in Table 4.1. However, it should be noted that this study does not consider the influence of L1 on the use of LBs as it is too wide-ranging to serve as a contribution to the role of mother tongue on LBs development.

Table 4.1. Description of the participants' nationalities. (cross-sectional data)

Nationality	Number of essays	Percentage
Saudi	180	29%
Chinese	118	19%
Omani	64	10.3%
Kuwaiti	63	10.1%
Emirati	42	6.8%
Iraqi	36	5.8%
Russian	30	4.8%
Italian	30	4.8%
Spanish	16	2.6%
Turkish	14	2.3%
Peruvian	11	1.8%
Korean	9	1.4%
Mexican	8	1.3%

The participants' ages ranged from 18 to 35 years, with a mean age of 23.3, and a standard deviation (SD) of ± 3.4 years. The majority were male (75%), with only a small percentage of females (25%). Intermediate and advanced second language (L2) learners' levels were selected as the target academic registers (Table 4.2). These learners were chosen as they considered academic writing to be a primary skill in the development of academic competence. They volunteered to write argumentative essays, equivalent to the International English Language Testing System (IELTS) Task 2.

Table 4.2. Description of the cross-sectional study participants (B1, B2 and C1).

Cross-sectional study participants	
Number of participants	621
Gender	75% Male, 25% Female
Student Levels	Intermediate-advanced
Learner type	ESL
Average age	23
Average time in the UK	Minimum of two months

In the guidelines for the collection of the three sub-corpora, the students were given a list of topics and questions for writing their argumentative essays (Table 4.3). This is useful to avoid the topic influence of the type of bundles used.

Table 4.3. Control topic of the cross-sectional data.

Essay topics	**Number of essays**
Crime & punishment	70
Education	60
Environment	66
Family and children	74
Food and diet	70
Government	53
Health	52
Language	64
Technology	57
Transportation	55

The essays were required to be the student's own, although they were permitted to use reference tools, including teacher feedback, but no third-party assistance in composing the essay. The essays were required to be argumentative in nature, and must be at least 200 words in length, up to 500 words. Figure 4.2 shows an example of a student's argumentative essay at intermediate level, compiled for the purpose of the study. The essay contains 247 words, in which the student states a position, and provides an explanation for their choice.

MEN v WOMEN: COOKING

"Some people believe that women are better cooks than men. To what extent do you agree or disagree?"

Some people believe that women are better cooks than men. However, others think that women cook worse than men. The common perception that women are better at cooking comes from the past when the main task of women was manage and serve the food which men provided by hunting. Nevertheless, times have changed and now we can say we are living in a period where roles are not clearly divided.

Certainly what unites the cultures of the whole world is the reassuring figure of the mother or grandmother ready to serve the meal on the table. This cultural fact has a great importance because it means that women have handed down the "culinary arts" over the years and this explains the grater dexterity and attention they pay to cooking.

On the other hand, nowadays a marked social difference between man and women does not exist as it did in the past, and this has led to a division of domestic tasks, including cooking. For this reason a lot of men have become passionate about cooking and there are many cases where the cook in the family is the male figure. Moreover, if you consider the best chefs in starred restaurants, they are mostly men.

In conclusion, it can be seen that women have characteristics more suited to work in the kitchen, probably developed over the centuries. However, men today can stand comparison and even overcome them in the preparation of meals as can be seen in everyday life and in restaurants.

Figure 4.2. Example of an argumentative essay by an ESL learner (intermediate level).

Finally, the students were asked to submit their essay either electronically via email, or to hand them to their tutors, and were asked to include their email address, in order that they could be contacted and given the vouchers awarded for their participation in the study. It is worth to mention that some of the essays submitted were excluded from the study, as the participants declined to accept the vouchers. Table 4.4. provides a description of the cross-sectional study dataset.

Table 4.4. Description of the cross-sectional study collected data.

Data Collection	
Data collection venue(s)	Language centres in the UK
Collected essays	602 (412 intermediate, 190 advanced)
Types of texts	argumentative, descriptive essays
Total word token	187,856
Average word count per essay	312

As shown in above table, 602 ESL learners contributed 602 essays, constituting cross-sectional sub-corpora with a total of 187,856 running words, and a mean of 312 words per essay.

4.5.4 Participants and data collection (longitudinal study)

The second purpose of this study was to determine the extent to which there was a relationship between the frequency of use of LBs and proficiency level. Thus, it was essential to follow a group of ESL learners from lower level to higher levels in language study. As advocated by Cortes (2004, p.415):

> "The only reliable way to identify patterns of development in the use of LBs by students at different levels would be to conduct a longitudinal study of the same students investigating the evolution in the production of target bundles in their writing".

As Dörnyei (2007, p.82) explained, in a longitudinal study, "successive measures are taken at different points in time from the same respondents". Therefore, longitudinal data primarily provides information about what happened to a collection of research units over a specific time (Taris, 2000), enabling the measurement of linguistic changes as they occur in a particular population, thus facilitating the monitoring of change, for example of LBs usage.

According to the purpose of this study and the data collection process, it was genuinely longitudinal in nature, examining the development of different participants from the same population over a specific timeframe. The study traced nine ESL learners enrolled on English for Academic Purposes courses at UK language centres over a period of six months (same population in the cross-sectional), in order to compare the frequency of occurrence of LBs usage across three different CEFR levels, B1, B2, and C1. A profile of the participants is provided in Table 4.5 and Table 4.6.

Table 4.5. Description of the longitudinal study participants.

Data Collection	
Number of participants	9
Gender	78% Female, 22% Male
Student level	Intermediate-advanced
Learner type	ESL
Time in the UK	At least two months

Table 4.6. Nationalities of the participants in the longitudinal study.

Nationalities	Number of essays
Saudi	26
Chinese	24
Italian	26
Spanish	25
Chinese	24
Iraqi	24
Omani	24
Emirati	26
Turkish	25

As shown in the above tables, ESL learners from different nationalities are forming the longitudinal study dataset. Their age ranged from 22 to 33 years, with a mean of 25.6, and an SD of ± 3.5, seven participants were female, and two were male. The students were full-time students registered on English for Academic Purposes courses.

The data collection for the longitudinal study was undertaken from September 2018 to August 2019, at different language centres in the UK. The participants were asked to write at least two argumentative essays as a weekly homework assignment, over a six-month period. They were informed that the tasks would not be graded, and would only be used for research purposes, and thus would not affect their final evaluation. In the guidelines for the collection of the three sub-corpora, the students were given the same list of topics and questions provided in (Table 4.3) for writing their argumentative essays.

The researcher provided feedback on the students' essays, specifically on the aspects of building an argument and organizing the ideas presented, but not on their use of LBs. The students were informed that the nature of the writing required for the study was argumentative essays equivalent to the IELTS Task 2, each with a minimum of 230 words (a minimum word count recommended for achieving a good score for the task requiring the building of an argument, and which increased the chance of using more complex sentences that affected the grammatical range and the accuracy score), and no longer than 500 words, which created a comparable essay size for building the sub-corpora. Each student composed one to two assignments every week, with a total of at least 24 essays over six months. A description of the longitudinal study data is provided in Table 4.7.

Table 4.7. Longitudinal study data.

Data collections	
Data collection venue(s)	Language centres in the UK
Collected essays	242
Types of texts	Argumentative
Total words count	69777
Average word count per essay	288

In total, nine participants provided 242 texts, with a mean of 26.8 per participant, which constituted the longitudinal corpus, with a total of 69,777 running words, and a mean of 288 running words in each essay.

Once the cross-section and longitudinal study data were received, the essays were assigned to raters, who categorized them under the appropriate sub-corpora.

4.5.5 Determination of CEFR levels

This section first discusses the raters' qualifications, and their teaching experience, then provides the rating procedures employed for linking the participants' essays to the relevant CEFR level, and the evaluation of the rating performance. It explains the standardization of the judgments used in this study to categorize the essays, and the measures employed to preserve the reliability of the scoring.

4.5.5.1 Raters' profiles

An experienced and qualified English language teacher was chosen to train the raters, as he had an IELTS training certificate. Three experienced and qualified English language teachers were selected to re-rate the essays because of their experience in teaching IELTS preparation or as IELTS examiners. They had experience of examining written English ranging from five to 11 years, and held a recognized qualification in teaching for academic purposes. For example, an MA in Teaching English to Speakers of Other Language (TESOL), or a Certificate in Teaching English to Speakers of Other Languages (CELTA) or (TEFL) certificate, which can be used to teach English to non-English speakers. The three raters reported a high level of familiarity with the CEFR, and had used it for assessing students' examinations, such as in placement tests. Their education and teaching background is presented in Table 4.8.

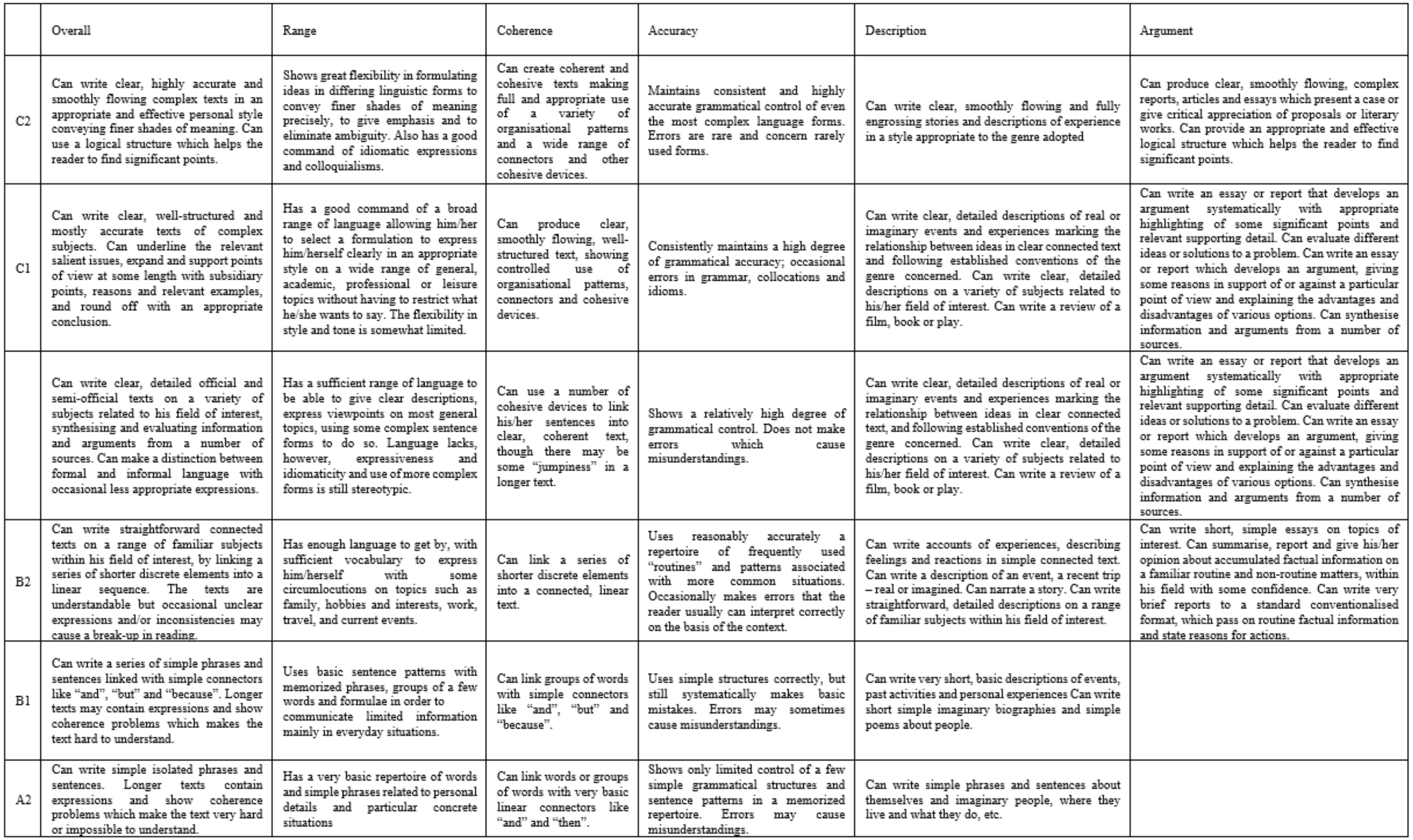
Table 4.8. Raters' profiles.

Model	**Rater 1**	**Rater 2**	**Rater 3**	**Senior rater**
First language	English	English	English	English
Current work	Teaching English of academic purposes+ IELTS preparations	Teaching English of academic purposes+ IELTS preparations	Teaching English of academic purposes+ IELTS preparations	Teaching English of academic purposes+ IELTS preparations
Qualification	MA TESOL	MA TESOL	CELTA+MA TESOL	CELTA+MA TESOL+IELTS training certificate
Teaching experience	12	10	10	20
IELTS teaching experience	5	6	4	11
Familiarity with CEFR before training	Excellent	Excellent	Excellent	Excellent
Past experience using the CEFR scales in marking written English	Yes	Yes	Yes	Yes

4.5.5.2 Procedure for rating the essays

This study sought to assess the participants' writing proficiency using the CEFR scales applied by three expert raters who taught English for academic purposes. As discussed in the previous section, one senior rater and three experienced English teachers were recruited to participate in the rating procedure. The senior rater worked as a trainer with the researcher to ensure that the three raters were familiar with the CEFR scales, and the three experienced teachers attended training in rating standardization before rating the collected essays, according to the CEFR scales.

The agreed procedure for standardizing the judgments used in this study for categorizing the essays followed that proposed by Chen and Baker (2016) originated from the manual for *Relating Language Examinations to the Common European Framework of Reference for Languages* (Council of Europe, 2003). This manual helps "the providers of examinations to develop, apply and report transparent, practical procedures in a cumulative process of continuous improvement to situate their examinations with the CEFR" (North and Jones, 2009). The framework provides six inter-related sets of procedures for distinguishing between the six CEFR levels (A1 to C2), using a Writing Assessment Scale developed by CEFR This scale consists of overall descriptors, and three writing analytical criteria: range, coherence, and accuracy, as shown in Figure 4.3.



	Overall	Range	Coherence	Accuracy	Description	Argument
C2	Can write clear, highly accurate and smoothly flowing complex texts in an appropriate and effective personal style conveying finer shades of meaning. Can use a logical structure which helps the reader to find significant points.	Shows great flexibility in formulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Can create coherent and cohesive texts making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.	Maintains consistent and highly accurate grammatical control of even the most complex language forms. Errors are rare and concern rarely used forms.	Can write clear, smoothly flowing and fully engrossing stories and descriptions of experience in a style appropriate to the genre adopted	Can produce clear, smoothly flowing, complex reports, articles and essays which present a case or give critical appreciation of proposals or literary works. Can provide an appropriate and effective logical structure which helps the reader to find significant points.
C1	Can write clear, well-structured and mostly accurate texts of complex subjects. Can underline the relevant salient issues, expand and support points of view at some length with subsidiary points, reasons and relevant examples, and round off with an appropriate conclusion.	Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say. The flexibility in style and tone is somewhat limited.	Can produce clear, smoothly flowing, well-structured text, showing controlled use of organisational patterns, connectors and cohesive devices.	Consistently maintains a high degree of grammatical accuracy; occasional errors in grammar, collocations and idioms.	Can write clear, detailed descriptions of real or imaginary events and experiences marking the relationship between ideas in clear connected text and following established conventions of the genre concerned. Can write clear, detailed descriptions on a variety of subjects related to his/her field of interest. Can write a review of a film, book or play.	Can write an essay or report that develops an argument systematically with appropriate highlighting of some significant points and relevant supporting detail. Can evaluate different ideas or solutions to a problem. Can write an essay or report which develops an argument, giving some reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options. Can synthesise information and arguments from a number of sources.
	Can write clear, detailed official and semi-official texts on a variety of subjects related to his field of interest, synthesising and evaluating information and arguments from a number of sources. Can make a distinction between formal and informal language with occasional less appropriate expressions.	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, using some complex sentence forms to do so. Language lacks, however, expressiveness and idiomaticity and use of more complex forms is still stereotypic.	Can use a number of cohesive devices to link his/her sentences into clear, coherent text, though there may be some "jumpiness" in a longer text.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstandings.	Can write clear, detailed descriptions of real or imaginary events and experiences marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned. Can write clear, detailed descriptions on a variety of subjects related to his/her field of interest. Can write a review of a film, book or play.	Can write an essay or report that develops an argument systematically with appropriate highlighting of some significant points and relevant supporting detail. Can evaluate different ideas or solutions to a problem. Can write an essay or report which develops an argument, giving some reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options. Can synthesise information and arguments from a number of sources.
B2	Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence. The texts are understandable but occasional unclear expressions and/or inconsistencies may cause a break-up in reading.	Has enough language to get by, with sufficient vocabulary to express him/herself with some circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Can link a series of shorter discrete elements into a connected, linear text.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more common situations. Occasionally makes errors that the reader usually can interpret correctly on the basis of the context.	Can write accounts of experiences, describing feelings and reactions in simple connected text. Can write a description of an event, a recent trip – real or imagined. Can narrate a story. Can write straightforward, detailed descriptions on a range of familiar subjects within his field of interest.	Can write short, simple essays on topics of interest. Can summarise, report and give his/her opinion about accumulated factual information on a familiar routine and non-routine matters, within his field with some confidence. Can write very brief reports to a standard conventionalised format, which pass on routine factual information and state reasons for actions.
B1	Can write a series of simple phrases and sentences linked with simple connectors like "and", "but" and "because". Longer texts may contain expressions and show coherence problems which makes the text hard to understand.	Uses basic sentence patterns with memorized phrases, groups of a few words and formulae in order to communicate limited information mainly in everyday situations.	Can link groups of words with simple connectors like "and", "but" and "because".	Uses simple structures correctly, but still systematically makes basic mistakes. Errors may sometimes cause misunderstandings.	Can write very short, basic descriptions of events, past activities and personal experiences Can write short simple imaginary biographies and simple poems about people.	
A2	Can write simple isolated phrases and sentences. Longer texts contain expressions and show coherence problems which make the text very hard or impossible to understand.	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations	Can link words or groups of words with very basic linear connectors like "and" and "then".	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorized repertoire. Errors may cause misunderstandings.	Can write simple phrases and sentences about themselves and imaginary people, where they live and what they do, etc.	

Figure 4.3.written assessment criteria grid (Europe, 2003).

The steps for linking the CEFR to the essays were used in the pilot study to test the instruments to be employed in the main study. Consequently, additional training tasks were added to the training course to ensure that the raters acknowledged the CEFR scales. In total, three experienced native English speakers who had taught ESL for more than five years undertook advanced training on linking the essays to the relevant CEFR levels provided by the researcher and a trainer, who was an IELTS examiner trainer. This training sought to enable the connection of the participant essays to the appropriate CEFR level of proficiency.

Before starting the training, all the raters were required to sign a consent form concerning their participation in the study and completed a raters' form detailing their qualifications and experience of teaching ESL. The procedure for linking a test to the CEFR involved five steps that needed to be completed at different stages. The first step was CEFR familiarization, a range of training tasks designed to ensure that the individuals rating the essays were fully informed about the CEFR and its illustrative descriptors. The second step was specification (Stage 2), which involved training the participants in assessing performance, in order to relate the tests to the CEFR, using standardized samples. This step can be seen as a primarily qualitative method serving a reporting function, and with a particular awareness-raising function that helped to improve the quality of the examination concerned.

The next step was benchmarking, which involved familiarizing the trainees with the CEFR via the activities provided in the previous steps, to ensure that they understood the CEFR levels, and to provide guidance for the raters (Stage 3). This stage was divided into two phases: first, the illustration, in which written samples were used to explain how performance illustrates the level described in the CEFR's overall scale in each sample. The professional teachers (raters) then proceeded to the practice phase, undertaking further standardization training of certain essays. The three raters concerned were involved in a post-standardization marking test that assigned CEFR levels to ten essays, and identified the CEFR descriptors indicated by each essay, using the CEFR overall written assessment criteria grid. After the post-standardization marking test, two of the teachers were assigned to rate all the essays independently, classifying each item under one of the six CEFR levels. If any of the texts received different scores, they were then re-rated by the third teacher. Therefore, some essays received three ratings, rather than two. However, if an essay received three different grades, it was excluded from the analysis (Stage 4). The final stage involved the use of statistical analysis to examine the inter-rater reliability, and to categorize the essays under the relevant CEFR levels. In order to compare the rating performance, the inter-rater reliability for all the ratings was calculated to determine the percentage of agreement between the raters (Stage 5). The study used Statistical Package for the Social Sciences (SPSS)- Kappa, following the method proposed by Mchugh (2012), which measures the inter-rater agreement of qualitative items, to determine the possibility of the agreement occurring by chance. The kappa ranged from −1 to +1, as with most correlation statistics, where 0 indicated the amount of agreement that could be expected from random chance, and 1 represented excellent agreement between the raters (Cohen, 1988; Mchugh, 2012). The guidelines provided by Fleiss et al. (1981) characterized kappa over 0.75 as strong agreement, 0.40 to 0.75 as fair to good, and below 0.40 as weak agreement. The assessment of the kappa provided a way of quantifying the degree of agreement between two raters regarding the rating of the ESL learners' essays, according to the CEFR levels. The results of the rating procedure established five CEFR learners' sub-corpora, as shown in Table 4.9 and Table 4.10.

Table 4.9. The results of the essay rating of the cross-sectional data.

Sub corpora	**Total number of words**	**Numbers of essays**	**Average length**
A1	6072	33	184
A2	12993	61	213
B1	50321	155	324
B2	53886	168	321
C1	64584	185	349
Total	187856	602	312

Table 4.10. The results of the essay rating of the longitudinal study data.

Sub corpora	**Total number of words**	**Numbers of essay**	**Average length**
A2	2442	10	246
B1	26795	86	311
B2	20585	76	270
C1	19955	70	285
Total	69777	242	278

4.5.5.3 Rating results (cross-sectional study)

The primary analysis in the cross-sectional study was of the ratings ascribed by the experienced raters to the learners' essays, in terms of their writing proficiency level on the CEFR scale. As discussed in the previous section, the inter-rater reliability score was required to be high to represent a strong agreement between the raters. The degree to which the two raters agreed with one other was considered. As with the observation in the study by Shechtman (1992), it was more natural to examine the overall score than to calculate each sub-band of the five competencies. The total rating score of the three raters in the present study produced a proficiency level on the overall CEFR scale for each essay, as shown in Table 4.11.

Table 4.11. Examples of the rating results (R1, R2, and R3).

Essay/ raters	**Rater1**	**Rater2**	**Rater3**
Essay 1	B1	B1	_
Essay2	C1	B2	B2
Essay3	C1	C1	_
Essay4	B1	B1	_
Essay 5	B1	B1	_

The five levels assigned to the students' essays were A1, A2, B1, B2, and C1. It was necessary to convert the ratings on the CEFR scale for use in SPSS by changing them from letters to numerical equivalents from 1 to 6, using the transform function in SPSS for the statistical analysis. The interrater reliability agreement between the raters was then examined.

As shown in Table 4.12, the result of the interrater reliability kappa = 0.865, which indicated that more than 86% of the observed values received a similar rating from the two raters concerned, with only 15% exhibiting a difference. Furthermore, the p = .000 value, actually means p<.05, showed that the kappa (κ) coefficient was statistically significant, different from zero. The Cohen's κ was run and demonstrated that the level of agreement between the raters was strong, as the kappa value showed a strong agreement of above 0.86.

Table 4.12. Symmetric measures of the cross-sectional data.

	Value	**Asymptotic Standard Error^s**	**Approximate T^b**	**Approximate Significance**
The measure of Agreement Kappa	865	.016	38.251	.000
N of Valid Cases	602	---	---	---

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

After the rating step, the total number of words in the ESL learners' corpus was 187,856 words in 602 essays. The essays were incorporated into five learners' sub-corpora, representing CEFR levels A1, A2 B1, B2, and C1, according to their rating. However, A1 and A2-rated essays were excluded from the analysis, as there was an insufficient number of samples. Level B1 was included in the main study, as the number of essays in the B1 sub-corpus (155 essays) was close to the number of B2-rated essays (168), and to the C1 sub-corpora (185 essays). Nevertheless, the number of essays in B1, B2, and C1 sub-corpora was subsequently reasonably comparable, which supported the approach of adding the B1 level to the analysis to obtain a broader view of the learners' language use across the levels. To achieve a comparable corpus size, the study used the same number of texts in each sub-corpus for the examination, as shown in Table 4.13.

Table 4.13. ESL learners' sub-corpora (B1, B2, and C1) in the cross-sectional study.

Sub corpus	**Word count**	**Number of essays**	**Average length per essay**
B1	50321	155	324
B2	49871	155	321
C1	51415	155	331
Total	168791	465	325

4.5.5.4 Rating results (longitudinal study)

In the longitudinal study, the degree to which the two raters concerned agreed with each other when rating the essays reached .84, with a *P*-value of less than 0.5, which showed a strong correlation between the raters when linking the essays to the CEFR scales (Table 4.14).

Table 4.14. Symmetric measures of the longitudinal study.

	Value	Asymptotic Standard Error^s	Approximate T^b	Approximate Significance
The measure of Agreement Kappa	.840	.029	20.173	.000
N of Valid Cases	242			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

The total number of words in the ESL learners' corpus in the longitudinal study data was 69,777 words in 242 essays. These essays were incorporated into four learners' sub-corpora, representing CEFR levels A2, B1, B2, and C1, according to their rating. However, following the same exclusion criteria employed in the cross-sectional study, A1, A2, and C2-rated essays were excluded from the analysis, as there was an insufficient number of samples. In order to achieve a comparable corpus size, all the sub-corpora included an equal number of texts (Table 4.15).

Table 4.15. ESL leaners' sub-corpora (B1, B2, and C1) in the longitudinal study.

Sub-corpora	Numbers of essay	Total number of words	Average length per essay
B1	70	21,873	312
B2	70	18899	270
C1	70	19955	285
Total	210	60727	278

Finally, three sub-corpora were established, in order to observe the development of LB usage across the CEFR levels. The next section concerns the reference sub-corpus used as a benchmark for the LBs identified in the sub-corpora.

4.5.6 The British Written English (BAWE) reference corpus

As discussed in section 3.5, the point of using a reference corpus is to have a benchmark against which you discover the main features in your corpus. This is so much related to the concept of salience. We need to know the degree of saliency of each feature in our corpus in relation to the reference one. Therefore, by considering the aim, genre and the size of the target ESL learners' sub-corpora the BAWE corpus was chosen as a reference corpus for this study. Unlike the pilot study, it was decided to use BAWE (linguistics and English) disciplines as a reference corpus. Although the chosen samples of the BAWE corpus used as reference corpus in the pilot study demonstrated useful in comparison with the ESL learners' sub-corpora, this approach was rejected as this misrepresented the sample of one discipline, deeming it unsuitable for comparison with the linguistics-dominated ESL learners' samples in this study.

However, although the BAWE corpus is also an academic writing corpus, it consists of assignments collected from 35 disciplines in four broad disciplinary groupings; thus, topic-specificity of certain expressions may lead to unexpected statistics. To be more consistent in the main study, it was decided to use BAWE (linguistics and English disciplines) as a reference corpus, to avoid topic-specificity of certain expressions. These two disciplines were large enough to be used as a reference corpus, and also included similar language to that the ESL learners employed in their academic essays. While using other disciplines, such as philosophy or biochemistry, might affect the results adversely, the linguistics and English disciplines were suited to the goal of this study, as they provided a wide range of language that was representative of that employed by ESL students writing in an authentic academic context.

The BAWE corpus was downloaded as a whole from the free online version. It included 35 main disciplines available in three formats: XML files, containing full mark-up, and categorized by discipline; a TXT version containing only a minimal number of tags, and not divided into disciplines; and a PDF version that represented the original documents. Since the reference corpus used in this study consisted of only two disciplines, English and linguistics, a computer code was generated to change the file format from XML to TXT. This code was required to divide the corpus into disciplines in TXT format for analysis using computer software (*WST*). The two subjects required were then extracted in TXT format to calculate the frequency of occurrence of three- and four-word LBs, use in comparison with the target bundles. Table 4.16 provides an overview of the size of the reference corpus used in this study.

Table 4.16. BAWE corpus overview.

Concepts	Wide range
Number of papers	221
undertaken in the period	2004-2007
Type of text	student assignments
The average length of texts	2372
types (distinct words)	22,732
Token used for the word list	524284

4.5.7 Corpus preparation

This study aimed to conduct a comprehensive search of all three- and four-word lexical bundlers in the ESL learners' sub-corpora, thus it was necessary to homogenize the number of bundles identified in each sub-corpus to compare the number of LBs used in the different sub-corpora. In order to ensure the accuracy of the extraction, it was necessary to clean all the essays of any unnecessary information, such as names and titles. However, the spelling and grammatical errors made by the learners were left unchanged in the files. The essays were then converted to TXT format for analysis using WST to determine the frequency of use of LBs, and for the analysis of keyness in academic writing. Finally, all the text files were renamed and compared manually with the actual student samples to ensure accuracy.

4.5.8 Computer software

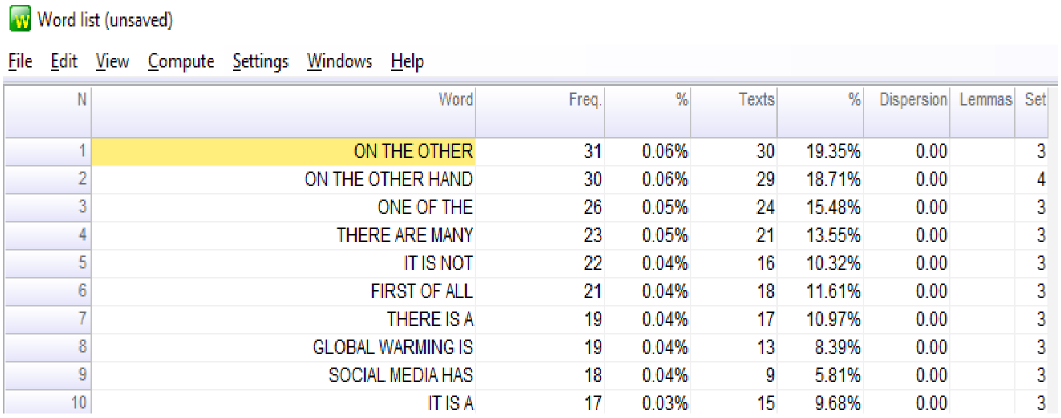
A number of computer software tools are available that useful for reaching empirical conclusions, and for analysing corpus data, such as *WST*, *AntConc, Sketch engine, etc*; however, depending on the purpose of the research, one of these programs may be more appropriate than the other. Ari (2006) reviewed three concordance software programmes and found that *WST* is the most efficient software search for LBs, as it is able to provide lists with most frequent multi-word sequences, as well as to plot distribution and professionality to show the concordance line for a specific bundle. A recent study by Fromm et al. (2020) compared between *WST* and the *Sketch engine* and found that both software are effective in analysing corpora because they provide the same functions except that *WST* allows the user knowing the lexical variation of the corpus through Type-Token Ratio.

Due in part to this review, along with its widespread use within corpus linguistic studies for the identification of LBs and for keyness analysis, *WST* 7.0 was chosen as the lexical analysis software used to search for LBs and keybundles within ESL learners' sub-corpora. The rationale for using *WST* was in part due to its wide range of programmes and functions, such as creating wordlists, and keywords and multi-word sequences, as well as the fact that it shows collocational patterns. More importantly, the ability to upload unlimited number of files, and unlimited number of words to create a corpus with no file size restriction. A brief description of the main functions in *WST* software will be discussed below.

WST allows practitioners to perform various analytical functions such as:

- Identifying wordlists and corresponding frequencies
- Sorting words in concordance lines and their collocations
- Keyword/keyness analysis

The first function corpus linguists often use when analysing a corpus is a *wordlist*. This function provides descriptive details of the corpus components in terms of the corresponding frequency of each word or phrase, types (different words in the text), and tokens (occurrence of words in texts), to name but a few, as shown in Figure 4.4. It also allows practitioners to sort the data in a corpus either alphabetically or by frequency. The user can choose to create a one-item-per-entry list or a 'bundles' wordlist, in which the list is made up of a string of words (called clusters in *WordSmith*) with a specified length.

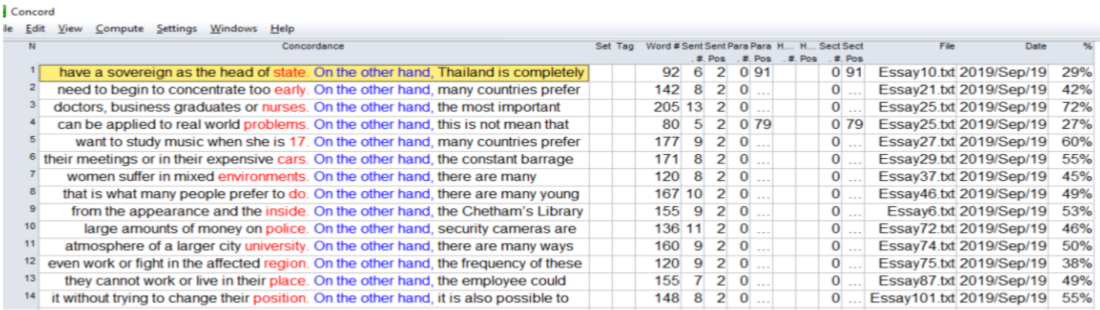


N	Word	Freq.	%	Texts	%	Dispersion	Lemmas	Set
1	ON THE OTHER	31	0.06%	30	19.35%	0.00		3
2	ON THE OTHER HAND	30	0.06%	29	18.71%	0.00		4
3	ONE OF THE	26	0.05%	24	15.48%	0.00		3
4	THERE ARE MANY	23	0.05%	21	13.55%	0.00		3
5	IT IS NOT	22	0.04%	16	10.32%	0.00		3
6	FIRST OF ALL	21	0.04%	18	11.61%	0.00		3
7	THERE IS A	19	0.04%	17	10.97%	0.00		3
8	GLOBAL WARMING IS	19	0.04%	13	8.39%	0.00		3
9	SOCIAL MEDIA HAS	18	0.04%	9	5.81%	0.00		3
10	IT IS A	17	0.03%	15	9.68%	0.00		3

Figure 4.4. A wordlist of the three- and four-word clusters ordered by frequency.

In the context of this study, the *word list* function generated a frequency list of the ESL learners' sub-corpora and the reference sub-corpus based on the plain text retrieved. After creating the wordlist, the software allows users to generate a frequency list of the most frequent LBs, by setting a bundle size and a frequency cut-off point. The frequency lists are used in this study for identifying the most frequent LBs in ESL learners' sub-corpora, which can indicate learners' language use.

The Second function provided by the software is *concordancing.* It is a process for accessing a corpus of a specific text to show how words or phrases in the content are used (Flowerdew, 1996). It presents all the instances of a target item, i.e., a word or cluster across multiple texts, and indicates its immediate context (before and after), where the word or phrase being tested is in the middle. Hence, it becomes easy to identify a pattern in language use in a specific text. The function was used in this study to confirm the decision taken to classify the LBs structurally and functionally manually from the original texts. Figure 4.5 shows an example of the LB *on the other hand* from the B1 sub-corpus.



N	Concordance	Set	Tag	Word #	Sent .#.	Sent Pos	Para .#.	Para Pos	H... .#.	H... Pos	Sect .#.	Sect Pos	File	Date	%
1	have a sovereign as the head of state. On the other hand, Thailand is completely			92	6	2	0	91			0	91	Essay10.txt	2019/Sep/19	29%
2	need to begin to concentrate too early. On the other hand, many countries prefer			142	8	2	0	...			0	...	Essay21.txt	2019/Sep/19	42%
3	doctors, business graduates or nurses. On the other hand, the most important			205	13	2	0	...			0	...	Essay25.txt	2019/Sep/19	72%
4	can be applied to real world problems. On the other hand, this is not mean that			80	5	2	0	79			0	79	Essay25.txt	2019/Sep/19	27%
5	want to study music when she is 17. On the other hand, many countries prefer			177	9	2	0	...			0	...	Essay27.txt	2019/Sep/19	60%
6	their meetings or in their expensive cars. On the other hand, the constant barrage			171	8	2	0	...			0	...	Essay29.txt	2019/Sep/19	55%
7	women suffer in mixed environments. On the other hand, there are many			120	8	2	0	...			0	...	Essay37.txt	2019/Sep/19	45%
8	that is what many people prefer to do. On the other hand, there are many young			167	10	2	0	...			0	...	Essay46.txt	2019/Sep/19	49%
9	from the appearance and the inside. On the other hand, the Chetham's Library			155	9	2	0	...			0	...	Essay6.txt	2019/Sep/19	53%
10	large amounts of money on police. On the other hand, security cameras are			136	11	2	0	...			0	...	Essay72.txt	2019/Sep/19	46%
11	atmosphere of a larger city university. On the other hand, there are many ways			160	9	2	0	...			0	...	Essay74.txt	2019/Sep/19	50%
12	even work or fight in the affected region. On the other hand, the frequency of these			120	9	2	0	...			0	...	Essay75.txt	2019/Sep/19	38%
13	they cannot work or live in their place. On the other hand, the employee could			155	7	2	0	...			0	...	Essay87.txt	2019/Sep/19	49%
14	it without trying to change their position. On the other hand, it is also possible to			148	8	2	0	...			0	...	Essay101.txt	2019/Sep/19	55%

Figure 4.5. Example of the bundle *on the other hand* from B1 sub-corpus.

The screenshot above demonstrates that the bundle *on the other hand* appeared in different texts; it met the distribution criteria, and was considered to be a LB. By clicking on the cluster *on the other hand* in the window, a larger section of the original source was displayed. This was used to confirm the structure and function of the LBs in its original text. The procedure was applied to all the sequences identified in the ESL sub-corpora and the reference sub-corpus.

Another function performed by *WST* is the "*keyword"* function. The software can also compare two wordlist files, usually one from a smaller corpus and the other from a larger corpus. The result is a list of words or cluster whose frequency is unusually high (positive keywords) or unusually low (negative keywords) in the target corpus in comparison to the reference corpus. The keyword lists provided the high-frequency tokens in order of outstandingness. This function was also used in the current study to generate individual 'key lexical bundles', namely those bundles whose frequency was found to be outstanding in the ESL learners' sub-corpora, compared with the BAWE (reference sub-corpus).

This study sought to determine how language was used in B1, B2, and C1 ESL learners' writing, compared with writing in a more general dataset. All the lists produced in the three steps were saved, cleaned and visualized manually for use in the analysis. The next section describes the identification of the LBs employed in this study.

4.6 Identification of lexical bundles

As discussed in section 2.4, the criteria under which multi-word sequences were regarded as LBs primarily concerned 1) length, 2) frequency, and 3) range. These three criteria are briefly explained below.

4.6.1 Length

The first key decision to be made in the research process was the bundle length for which to search. As discussed in section 2.4.3, although LBs can be of varying lengths, bundles of three- and four-words previously received special interest, because they were consequently the most frequent length investigated in the existing research, while longer bundles, for instance of five or six words, were less commonly examined. The present study examined three- and four-word LBs, as they are the most frequently used LBs in academic writing (Biber et al., 1999; Hyland, 2008b; Salazar and Joy, 2011; Panthong and Poonpon, 2020), while longer bundles were rare and defined according to the students' level and the text size; this approach facilitated a much easier comparison between the learners' levels. The bundles were defined empirically, rather than intuitively.

4.6.2 Frequency

The frequency threshold was the second criteria employed for identifying LBs. As argued in section 2.4.1, the frequency cut-off point is "somewhat arbitrary" (Ädel and Erman, 2012, p.82), and is based on the aim and "on researchers' evaluation of data manageability" (Chen, 2008, p.64), with no agreement in the current literature regarding the correct cut-off point. Previous research used a minimum normalised threshold of between 10 to 40 occurrences per million words, in which a group of multi-words was considered to be a bundle (Biber et al., 2004; Biber and Barbieri, 2007; Hyland, 2008a; Chen and Baker, 2010; Jalali, 2015). The high cut-off point was chosen to identify LBs, due to Biber and Barbieri (2007) discussion of the comparison between sub-corpora of over 1,000,000 and fewer than 40,000 words. As Bestgen (2018) claimed, the smaller the corpus, the higher the cut-off point must be, to ensure that the LBs are statistically significant. Thus, a high frequency of 40 times per million words was employed in the present study to identify the LBs in the ESL learners' writing, within three small sub-corpora, which were used as a basis for the calculation of LBs across all the sub-corpora, regardless of their length. The rationale for using a high cut-off-point was to skew the number of bundles included in the final analysis, to differentiate between the ESL learners' levels. It would therefore be necessary to convert the raw frequency to four times in 100,000 words, in order that it would be equivalent to 40 in a million words, to be comparable with the previous studies in the field. Thus, a bundle that occurred three times in the corpus would have a higher frequency of occurrence than 40 times per million words. This frequency threshold was applied to the B1, B2, and C1 sub-corpora in both elements of the present study.

Meanwhile, for the BAWE sub-corpus frequency threshold, this study employed a dynamic threshold that took into account the size of the corpus when choosing the cut-off points, to achieve representativeness and comparability of the bundles extracted (Chen and Baker, 2016). While the reference sub-corpus was bigger in size, this study followed Bestgen (2018), using a smaller threshold if the corpus size is at least 500,000. The study noted that for three-word bundles, a threshold of above ten is necessary for all corpus sizes, as the threshold of 10 per million words only works with four-word LBs. Thus, the frequency cut-off point applied to the reference sub-corpus in the present study was 20 times per million words (two times per 100,000 words).

Applying a cut-off point of between 20-40 times per million words fell well within the range used by previous studies.

When compiling the essays for the cross-sectional study and the longitudinal study, an effort was made to ensure that the running words for the sub-corpora were close to each other to control the number of words, to focus largely on high-frequency words only, as shown in Table 4.17 and Table 4.18. The converted frequencies of the B1, B2, and C1 sub-corpora were small, which were rounded up to two occurrences, as two was the minimum frequency employed by previous studies.

Table 4.17. Cut-off points in the cross-sectional study.

Corpus	**Cut-off points in absolute frequency(F)**
B1 sub-corpus	2 in 50321words
B2 sub-corpus	1.9 rounded up to2 in 49871 words
C1 sub-corpus	2 in 51415 words
Normalized frequency of the cut-off point	4 times per 100,000 words
Dispersion criteria	3 texts
BAWE (LING &E)	10 in 524284words
Frequency of the cut-off points per 100,000 words	Two times per 100,000 words
Dispersion criteria	Four texts

Table 4.18. Cut-off points in the longitudinal study.

Corpus	**Cut-off points in absolute frequency (F)**
B1 sub-corpus	2 in 21,873 words
B2 sub-corpus	2 in 18899 words
C1 sub-corpus	1.9 rounded up to2 in 19955 words
Normalized frequency of the cut-off point	Four times per 100,000 words
dispersion criteria	At least three texts

The initial results of LBs identified from the sub-corpora showed differences in the number of LBs across the sub-corpora in both studies, as described in Table 4.19 and Table 4.20.

Table 4.19. The initial extraction of the target bundles (cross-sectional study).

Sub-corpus	Initial extraction results
B1	7466
B2	6123
C1	8159
BAWE	1871

Table 4.20. The initial extraction of the target bundles (longitudinal study).

Sub-corpus	Initial extraction results
B1	1350
B2	2673
C1	2095

4.6.3 Shortlisting by dispersion criteria

In addition to the frequency of LB usage, the multi-text occurrences criteria was essential for identifying the LBs used, as a safeguard that prevented idiosyncratic usages, and which did not consider author bias (see section 2.4.2). Thus, the range of texts in which LBs occurred was taken into account when identifying the bundles. Range concerns the prevalence of LBs within a corpus and helps to ensure that the bundles identified are common within the corpus as a whole.

Since all of the sub-corpora were of almost identical size, using the advice of Biber and Barbieri (2007), it was decided that the LBs should be present in a minimum of three texts. The dispersion threshold was therefore counterbalanced by requiring the bundles identified to have a wide distribution throughout the texts, with a range of three texts in each ESL learners' sub-corpus, and of four texts for the reference sub-corpus. Table 4.21 shows that the initial results for the LBs identified across the sub-corpora.

Table 4.21. Result of dispersion criteria of LBs in the cross-sectional data.

Sub-corpus	Extraction results
B1	1177 bundles
B2	959 bundles
C1	2035 bundles
BAWE	1636 bundles

Table 4.22. Result of dispersion criteria of LBs in the longitudinal data.

Sub-corpus	Extraction results
B1	215 bundles
B2	254 bundles
C1	276 bundles

After extracting the items that did not meet the minimum frequency and dispersion criteria, there remained a vast number of LBs that would be useful for the investigation. The next step was the manual removal of the content-based items, to restrict the lists of the bundles to those that were considered to be used in general academic writing.

4.6.4 Manual removal for content-based bundles

As discussed in section 3.7, it was decided to discard all of the content-based bundles, such as proper nouns like *Pollution in the*, or *global warming is*, as they did not reflect the use of general academic language, following Chen and Baker (2010). In order to exclude the content-based bundles, LBs that contained words directly connected to the topic of the essay were manually discarded from the lists. For example, one of the essay's titles was 'The government has announced that it plans to build a new university. Some people think that your community would be a good place to locate the university'. To remove the content-based words, any bundles that contained keywords from this title, such as *government*, *university*, and *community*, were discarded from the lists. Examples of the content-based bundles removed from the lists are provided in Table 4.23

Table 4.23. Examples of the content-based bundles extracted from the sub-corpora.

Content-based bundles	Frequency	Texts
Of global warming	19	13
Gain more knowledge and	16	16
In the countryside	7	4
Of Saudi Arabia	4	3
Pollution in the	4	3
Of international students	3	3
The problem of overpopulation	3	3
Scientists say that people	3	3

In the cross-sectional study, a total of 1,413 content-based bundles were identified, producing 662, 641 and 1,225 content-independent bundles in B1, B2, and C1 levels, respectively, as shown in Table 4.24. For the longitudinal data, 165 content-based bundles were removed from the three sub-corpora, leaving 180,194 and 217 independent bundles for the next filtering procedures, as shown in Table 4.25.

It is evident that high-level ESL learners used more frequently content-based bundles. A comparison of the content-based bundle in ESL learners might be an important issue for future research. The study also extracted 420 content-based bundles identified from the BAWE list, producing 1,169 independent content bundles.

Table 4.24. Extraction of content-based bundles from the target bundles (cross-sectional data).

Sub-corpus	Content-based	Independent-based
B1	281	662
B2	322	641
C1	810	1225
BAWE	467	1169

Table 4.25. Extraction of the content-based bundles from the target bundles (longitudinal data).

Sub-corpus	Content-based	Independent-based
B1	35	180
B2	51	194
C1	58	217

Two further procedures remained to filter out the overlapping and other noisy bundles, as discussed in the next sections.

4.6.5 Manual exclusion of overlapping bundles

Following the same procedure applied in the pilot study (Section 3.7), the next step was filtering out the overlapping bundles that referred to those three- and four-word bundles that might be included in longer bundles, such as four-, five-, six, and seven-word bundles. Following the suggestion of Chen and Baker (2010), the overlapping LBs were also combined into one bundle, to avoid duplication in the counting of high-frequency sequences. As Chen and Baker (2010) explained, there are two types of overlapping, the first of which occurs when, for example, the three- or four-word bundles originate from longer bundles, such as five- or six-word bundles. For instance, the three-word bundle *on the other,* constructed from the four-word bundle, *on the other hand,* occurred 19 times in the B2 corpus in the present study. The second type of overlapping is where two bundles overlap, and one of the phrases subsumes the other bundle, via complete submission; for example, *first of all,* occurred 14 times in the present study, while *of all the* occurred only three times. These two bundles occurred as subsets of the four-word bundle *first of all the.* To avoid inflating the analytical results, the overlapping LBs were combined to create a single, longer unit, adding the fourth word in brackets. For example, the bundles *on the other hand* and *the other hand the* were combined to make one single bundle: *on the other hand + (the).* Table 4.26 shows the two types of overlapping LBs from the written sub-corpora to illustrate how this process worked.

Table 4.26. Examples of the two overlapping bundles.

Bundles	**Frequency**
A wide range	4
Wide range of	4
A wide range (of)	
A lot of	46
With a lot	4
A lot of	

Example 1: The two three-word bundles, *a wide range* and *wide range of*, overlap the four-word bundle *a wide range of.* To avoid duplication, the two bundles were combined, and the longer four-word bundle, *a wide range (of)*, was used by adding the fourth word in brackets.

Example 2: The bundles, *with a lot* and *a lot of,* subsumed the four-word bundle, *with a lot of.* However, the bundle *a lot of* occurred 46 times, while *with a lot* occurred only four times. To avoid inflation of the analytical results, the frequency bundle of a lower number was omitted from the analysis, as it is a subset of the more commonly occurring frequency bundle.

4.6.6 Grammatical Inflectional variants

After reviewing the lists retrieved manually, some inflectional variants of the same bundle were identified, for example, *plays an important* versus *play an important*, and *I do not* versus *I did not*. Therefore, it was necessary to conduct a further examination to consider whether the inflectional variants were separate bundles or repeated bundles. According to Sinclair (1991), each separate form is a unique lexical unit. Also, the concordance line review revealed that there were differences between the inflected forms if their frequency of occurrence and their collocational associations were taken into account (Tognini-Bonelli, 2001). Thus, the question of the grammatical inflectional variants was concerned with the notions of the LBs' type and token. The type of LBs in this study referred to the total of lemmas or lexemes in a corpus, for example, *play*, while the token referred to the total numbers of word forms in a corpus, for example, *play*, *plays*, *playing*, *played*. Therefore, as the LBs were identified by their frequency of occurrence (token), rather than by type, all the inflectional variants observed were included as a single unit.

4.6.7 Other noise bundles

The final step of the exclusion process in the creation of the target lists was the removal of certain other noise bundles that were in some way incomplete or unsuitable, such as *to sup up*, which had a spelling error. During the process of retyping the students' essays, some spelling errors identified in the ESL learners' sub-corpora. These spelling errors covered a small range, from minor and major mistakes, that might obscure what the writers' true meaning might have been. Although this method has the potential limitation on excluding some important bundles, the decision was made to exclude the small number of spelling errors identified across the essays to avoid the researcher intuition; moreover, the spelling was not the focus of this study. Table 4.27 and Table 4.28 show the final extraction of the LBs from the sub-corpora lists, which were noticeably smaller following the removal of the bundles that were not desirable for the investigation.

Table 4.27. Final extraction of the target bundles (cross-sectional data).

Sub-corpus	Overlapping and other noisy bundles	Bundles for investigation
B1	159	499
B2	167	374
C1	535	690

Table 4.28. Final extraction of the target bundles (longitudinal data).

Sub-corpus	Overlapping and other noisy bundles	Bundles for investigation
B1	27	153
B2	30	164
C1	29	188

4.7 Keybundles' extraction

One of the main approaches used to compare the ESL learners' use of LBs in this study was the exploration of keybundles, those items whose frequency is unusually high or low in a given text or corpus, when compared with a reference corpus, regardless of their importance in that corpus (Scott, 1997; Baker, 2004). This will help to explore the variability and distinguished characteristics of the learners' levels in the written discourse (see section 2.5).

Two computer software are commonly used for keybundles analysis: *WST* and Sketch engine. These software are effective in their purpose, because, for example, they process the keybundles analysis and allow configuring the language according to the study *corpus* and calculate the number of type/tokens. I have applied the keyword function on a small dataset and I found that both made a similar result of keybundles; however, depending on the purpose of the research, *WST* tool was selected over other tools because its ability to upload unlimited number of files, and unlimited number of words for building a corpus, with no file size restriction (see section 4.5.8).

For the keyness calculation, a number of intermediary steps must be undertaken. First, choosing the reference corpus. As Scott and Tribble (2006, p.58) recommended, the reference corpus "should be an appropriate sample of the language which the text we are studying (the 'node-text') is written in". An appropriate sample is a "large one, preferably many thousands of words long and possibly much more" (ibid.). …What constitutes ‘large enough’ for examination, and what constitutes an acceptable size for a reference corpus is debatable and discussed in section 3.5. The present study used the linguistics and English disciplines from the BAWE corpus, as a reference sub-corpus as these two disciplines were of sufficient size, included relative language that the ESL learner participants used in their academic essays, and provided comprehensive information about the language concerned (see section 4.5.6).

The second step to consider when to conduct keyness analysis is the generation of the frequency lists for both the target corpus and the reference corpus, with the help of appropriate software, such as *WST*. The concordance software usually enables users to establish multiple parameters. The first parameter is the minimum frequency threshold and dispersion criteria. In this step in the present study, minimum frequency cut-off points and dispersion criteria were applied to the target corpus lists, to determine the bundles included in the statistical analysis, following the same procedure applied in the frequency analysis of the LBs. A frequency threshold helps to exclude bundles that are unusual, but infrequent, in order to minimize spurious hits, while the dispersion criteria allows the researcher to exclude bundles that are not found across the texts concerned.

The third parameter is the statistical measures. Keyness value can be conducted using standard statistical measures, such as log-likelihood (LL) (De Schryver, 2012), chi-square (χ2) (Leone, 2010), to name but a few. These tests calculate the difference between the two frequencies of key items, enabling users to confirm the strength of the difference between the corpora concerned. This aids the specification of language patterns to a particular genre or domain, and the distinguishing of patterns of communicative style in various contexts (Adolphs et al., 2004). Both the log-likelihood and the chi-square statistical tests were commonly used in previous keyword research. According to researchers such as Chujo and Utiyama (2006) and Culpeper (2009), both statistical tests provide similar KWs results, thus the selection of either the chi-square test or the LL test does not affect the KWs analysis. However, Rayson and Garside (2000); Rayson et al. (2004) argued that the chi-square test becomes inaccurate when the expected frequency is less than five; thus, the LL test is the more reliable statistical test (Dunning, 1993). Therefore, the LL test was selected for the keyness analyses in the present study.

The LL test is similar to many other statistical tests that are a “null hypothesis significance test that has associated probability value (*P*-values) showing the probability of observing the data under the null-hypothesis” (Pojanapunya and Todd, 2018, p.145). In a keyness analysis, the *P*-value indicates whether or not the difference between the two corpora is due to chance. As Baker (2006, p.125) explained, “the smaller the *P*-value, the more likely that the word’s strong presence in one of the sub-corpora isn’t due to chance but a result of the author’s (conscious or subconscious) choice to use that word repeatedly”.

The actual *P*-value thresholds range from 0 to 1. It is worth noting that the *P*-value thresholds are arbitrary. There is no agreement on when the probability value score result in a unit is identified as a key, and previous research used various *P*-value levels for their results to be statistically significant (Hoffmann et al., 2008). In terms of the variation between disciplines, most social sciences use a *P*-value of 0.05 to indicate that the results are significant (Wilson, 2013), whereas corpus linguistics research generally employs a *P*-value of 0.01. However, as Luab et al. (2017, p.57) suggested, in the case of a keyness analysis, the results provide too many keybundles for researchers to examine, therefore it is preferable to set a low *P*-value, such as 0.000001. As Scott Scott (2012, p.145) noted, “with keywords where the notion of risk is less important than that of selectivity, you may wish to set a comparatively low *P*-value threshold such as 0.000001 (1 in a million) (1E-6 in scientific notation) so as to obtain fewer keywords” for researchers to explore manually. This provides only keybundles with a high keyness value. Moreover, Scott (2012) explained that the choice of reference corpus size does not make a difference if a relatively small *P*-value, such as (0.000001), is proposed. Therefore, by adopting a small statistical significance score as the indication of keyness, *WST* conforms with widespread contemporary practice in disciplines that employ quantitative analyses.

In addition to LL statistic, *WST* 7.0 also computed BIC score where the KWs are identified on the basis of the log likelihood score which indicate the level of confidence and effect size metrics which indicate the extent of the frequency difference of a word in a study corpus and a reference corpus (Gabrielatos & Marchi, 2011). It is possible that the BIC score would produce slightly different rankings for KWs than the Log likelihood statistic. In this study, I have considered both statistical measures when identifying keybundles to see the possible differences for ranking keybundles. If there is no direct correspondence between the two ranking, the extent to which they did would provide useful indications regarding their similarity in identifying keyness. However, if the BIC and the LL return the same ranking order, the study will use the BIC score as new common metric for ranking keybundles taking into account both frequency differences and statistical significance.

After establishing all the relevant metrics in the present study, the *keyword* function in *WST* obtained the keybundles in the ESL learners' sub-corpora, with the top ones being more key than those at the bottom of the list, in terms of their keyness results. Finally, the frequency cut-off point in stage 3 was applied to all the keybundles identified.

Keyness was measured in this study firstly to identify the LBs with significantly different densities in the ESL learners' writing, compared with the reference corpus. Second, to explore the variability and distinguished characteristics of LBs across the learners' levels. Based on the findings, insight was gained regarding why the ESL learners exhibited different uses of LBs. The aim was that the present study would aid writers who have difficulty writing academic essays to improve their language level.

4.8 Corpus analysis

Having established the research design, and the method used to compile the sub-corpora, this section discusses the procedures by which the sub-corpora and research tools were examined. The analysis was guided by the RQs, and proceeded in a number of stages to address them.

Following the refinement of the data in section 4.6, all the LBs identified were normalized to facilitate parity of comparison, and to reduce the effect of the random corpus size, and therefore to estimate the frequency of occurrence of the bundles identified in each sub-corpus, on the basis of the given normalized frequency threshold. Thus, the normalized frequency of a bundle was calculated with its raw frequency-time, 100,000, then divided by the corpus size. For example, a bundle that occurred three times in the B1 sub-corpus (50,321 words) would have a normed rate of six per 100,000 words, and 60 per million words, as follows:

(3/50321*100,000) = 6 per 100,000 words or
(3/50321*1,000,000) = 60 per million words

In order to address the RQ1, which concerned the most frequent LBs in the B1, B2, and C1 sub-corpora, the lists of the most frequent bundles were compared across the ESL learners' levels. The present study investigated the top 20 in each of the sub-corpora (B1, B2, C1, and BAWE). The rationale for focusing only on these bundles was because the sub-corpora size used in this study were small; the top 20 bundles represented the most frequent and widest range, reflecting the actual academic language use of the ESL learners. The lists of the 20 most frequent LBs were compared across the three sub-corpora and were then checked manually to identify the bundles that were shared across the levels.

After identifying the shared bundles, the overall frequency (type/token) of the LBs was examined across the ESL learners' sub-corpora, in order to compare the three levels, and also to compare them with the reference sub-corpus. The degree to which the differences between the B1, B2, and C1 levels were statically significant was tested using log-likelihood, via the *University Centre for Computer Corpus Research on Language (UCREL) Significance Test System*, which is one of the most accurate tests for comparing the relative frequency of phrases among corpora, and can compare more than two corpora (Rayson and Garside, 2000) (http://ucrel.lancs.ac.uk/llwizard.html).

The results obtained from the statistical analysis demonstrated remarkable variation in the use of three- and four-word LBs across the sub-corpora. Next, the frequency range of the LBs across the sub-corpora was assessed, in order to compare the distribution pattern, in terms of the normalized frequency of the bundles identified, according to the following five frequency bands: below 10, 10-30, 30-50, 50-70, and over 70. This comparison highlighted certain differences between the higher frequency use of LBs across the levels and showed how frequently the LBs were used in each sub-corpus.

The comparison was based on the LBs extracted from the ESL learners' sub-corpora and the BAWE sub-corpus. As discussed in section 3.5, the LBs present in the BAWE were earmarked as those that ESL learners should emulate, in order to achieve a high-level writing style. Both the BAWE and the ESL learners' sub-corpora were examined for the similarities and differences in LB use, in terms of their frequency, structure, and function. Following the identification of the LBs within the sub-corpora, according to their frequency and distribution, they were classified both structurally and functionally, with the help of the concordance line, in order to address RQ2, as discussed in detail below.

Similar to the pilot study (Section 3.7), this study adopted the structural taxonomy proposed by Biber et al. (1999), which has been modified and developed for the purpose of the present study to incorporate all the target bundles, including those that were not present in the classification of Biber et al. These modifications included the addition of sub-categories adapted from Biber et al. (1999) classification. Two bundles of sub-categories were added to the 'Verb-based' category, and another six sub-categories that did not fit any of the three main categories were assigned to the category 'Other'. These sub-categories were 'Wh-clause', adjective phrase, adverbial phrase, conjunction clause, model + verb, and personal pronoun. The complete adopted taxonomy used in this study is presented below.

Table 4.29. Structural classification adopted for this study.

Structural types	Sub-types	Example
Noun-based	Noun phrase with other post-modifier fragments	*an important role in*
	Noun phrase with of-phrase fragment	a combination of
Preposition-based	Prepositional phrase with embedded of-phrase	at some of
	Other prepositional phrase expressions	on the other hand
Verb-based	1st/2nd person pronoun + VP fragment	I believe that
	Be + noun/adjective phrase	is consistent with the
	Passive verb + prepositional phrase fragment	are shown in figure
	Anticipatory it + verb/adjective phrase	it is possible that
	(Verb phrase) + that-clause fragment	our data suggest that
	(Verb/adjective) + to-clause fragment	is likely to be
	Pronoun/noun phrase + be (+…)	there was no difference
	Other verb phrase	do not get
Other expressions	Adjectival phrase	and many other
	Adverbial clause fragment	as shown in figure
	Personal pronouns	his or her
	Model + verb phrase	Should be able to

The additional sub-categories enabled the analysis of the bundles identified to be refined from more general categories to a specific one.

The LBs extracted from each sub-corpus were analysed in terms of the grammatical structures discussed previously to assess the similarities and differences between the B1, B2, and C1 levels. The structural distribution of the LBs was conducted in two ways: the type and the token distribution of the target bundles, as examining them in this way across the sub-corpora clearly showed the differences between the levels, as a corpus could present a small range of LBs, but with a very high frequency of the same bundles, especially in a small corpus.

A similar analysis was conducted for the pragmatic functions of the bundles identified in the three sub-corpora. In the same way as in the pilot study (Section 3.7), Hyland's (2008b) functional taxonomy was applied to classify the LBs identified in the ESL learners' sub-corpora, and to compare them with the BAWE sub-corpus, in order to identify the differences between the ESL learners of different CEFR levels, in terms of the variety and accuracy of use of the LBs. After the categorization criteria were applied, the study compared the function of the bundles within the B1, B2, and C1 sub-corpora, and also compared their usage with proficient student writers in the BAWE sub-corpus. The analytical structures and functions described here were used as the basis for the analysis of the LBs in all three sub-corpora, to address RQ2 (Section 5.3).

Based on the pilot study results, the study also found a strong relationship between the structural and functional categories. For instance, most participant-oriented bundles consisted of clausal bundles, whereas the phrasal expressions were composed of noun-phrase and prepositional-phrase bundles (Pan et al., 2016; Wijitsopon, 2019). In this study, the Chi-square test of independence was employed to determine the possible correlation between the two categorical variables in this study, namely the structural and functional categories.

In order to address RQ3, which concerned keyness analysis, *WST* generated lists of the keybundles for the three sub-corpora, comparing each sub-corpus with the reference sub-corpus, individually (see section 4.7). The keyness analysis procedure followed in this study is summarized below:

1. Set up the parameters in *WST,* including frequency threshold (40 times), length of bundles (three- and four words), dispersion criteria (three texts), and p-value (0,0000001);

2. Upload the frequency target corpus lists (B1, B2, and C1) one at a time, and the reference corpus frequency list (BAWE sub-corpus), to obtain the keybundles list;

3. Check the concordance lines manually to exclude keybundles that do not meet the frequency threshold, and the dispersion criteria of the three lists;
4. Exclude the overlapping and content-based keybundles;
5. Search for patterns, or significant use of particular keybundles, such as connectors.

The keyness analysis explained above provided a comparison and revealed the differences between the ESL learners. This analysis was followed by an examination of the developmental use of LBs and proficiency level to address RQ4. The analytical procedures required to address this RQ commenced with tracking nine students over six months, in order to determine the relationship between the frequency of use of LBs and language competence, to assess how the frequency of use of LBs in the ESL learners changed across the learners' levels over time. The study hypothesized that as the ESL learners' level increased, they would begin to produce more LBs.

Once the lists of the most frequent LBs in the B1, B2, and C1 sub-corpora were ready for analysis, the three lists were compared, in order to assess the correlation between the use of LBs and learners' levels. The LBs and their frequency were first examined quantitatively, then their structure and function were examined using the same taxonomy applied in the cross-sectional study. The study then quantitatively compared the shared bundles between the ESL learners, to explore whether the bundles identified were commonly associated with written discourse.

This section discussed the methods employed in this study, describing in detail the process of analysis, which was guided by the RQs. The next section presents the rationale for the planning stages used to develop the sub-corpora.

4.9 Corpus construction rationale

The planning stages for building a corpus are essential and complex and include determining the corpus size, type, and target population to supply the appropriate texts that form the corpus. These stages require careful consideration to provide an appropriate basis for the investigation of language use (Biber et al., 1998b). The decisions concerning the construction of a corpus should be determined by the purpose of the study, and constitute the guidelines for collecting the data concerned. As Sinclair (2004, p.81) observed, researchers "must make the best corpus they can in the circumstances". Thus, if there is no extant corpus suitable for the purpose of a research study, a new corpus must be created, taking care to consider several factors, including representativeness, balance, size, and sampling, together with additional elements, all of which are discussed in the context of the present study in the next sections.

4.9.1 The rationale for using argumentative essays

An argument is the core of argumentative writing, typically occurs in the context of a discussion, and is used to persuade, defeat and negotiate, and consult and debate between different opinions (Wang and Bakken, 2004). In other words, it is a discussion between people with different opinions about the debatable issue that seeks to convince or resolve the difference in view. It requires the provision of evidence from multiple sources to support an idea logically. Therefore, argumentative writing is a complex task that requires a certain level of cognitive and linguistic skill (Nippold and Ward-Lonergan, 2010). For learners to improve their language ability, they must master all these skills, including the ability to develop and logically defend a position (Campbell and Filimon, 2018).

As discussed in section 2.7, it is one of the most common genres in academic setting that L2 learners are required to write, since the development of an argument is regarded as a key feature of successful academic writing. It is an important skill for students who have to write persuasively to make other people accept their point of view on a particular topic. Mastering argumentative essay composition is key for achieving writing success, and provides significant preparation for students' transition into higher education in a variety of disciplines, from hard to soft sciences. It helps students to acquire critical thinking, research skills, and develops their ability to defend their ideas. For test-takers, an argumentative essay is a required genre since the English language level is often tested by many internationally recognized admission tests, such as CEFR tests, the IELTS test, etc. Therefore, they constitute "the gatekeeping mechanism within individual courses as well as at critical stages of passage through secondary schools and into college" (Heath, 1993, p.105). Thus, since it is the most common educational tool used to determine language proficiency levels in L2 learners, the use of argumentative essays in the present study was an appropriate way of measuring the variation and development of ESL learners' academic writing.

Nevertheless, argumentative writing has been confirmed by researchers to be the hardest model in writing, L2 learners lack preparation for writing in the English language, which causes a poor performance in their papers (Ferretti et al., 2007; Neff-Van Aertselaer and Dafouz-Milne, 2008; Qin and Karabacak, 2010). Moreover, Wang and Bakken (2004, p.184) observed that many ESL, and English as a Foreign Language (EFL) learners and researchers “lack adequate writing experience and a basic understanding of academic writing”. According to Ramage et al. (2018) students do not understand the structure of the argumentative essay and the function of each part of the essay, so they write in an unclear and unstructured manner.

As discussed in section 2.2, one way to succeed in academic writing is using formulaic language and particularly LBs, which enables students to create natural and fluent spoken and written texts. These expressions repeatedly occur within the same register, demonstrating that language is “register specific and performs a variety of discourse functions” (Allen, 2009; Biber and Barbieri, 2007).

Therefore, being aware of students’ difficulties in writing argumentative essays, this study investigated the use of LBs in ESL learners’ argumentative essays, seeking to aid understanding of how the frequency of use of lexical features, particularly LBs, supported the ESL learners concerned in their argumentative essays. Although the nature of the essays varies considerably across and even within disciplines, the development of an argument is regarded as a key feature of successful writing by academics across disciplines (Lea and Street, 1998). Comparisons between L2 learners of different levels can unintentionally cause “a monolithic conception of good writing based on practices” of ESL learners (Hartse and Kubota, 2014, p.73). Moreover, considering the similarities and differences between students’ ability levels promotes a better understanding of ESL learners’ use of LBs in academic writing at the various stages of language learning, and provides important insights into ESL writing pedagogy.

4.9.2 Corpus types

Corpora are machine-readable compilations of authentic texts that often aim “to be representative of a particular kind of language” (Hunston, 2002, p.28). There are various types of corpora, depending on the purpose of the study (see section 2.3). For example, general corpora such as BNC, can consist of hundreds of millions of words, and can include various types of texts, either written or spoken, on a variety of subjects. Meanwhile, monitor corpora, such as the *Bank of English* corpus, are also large in size, and can be expanded to track changes in a language. In contrast, specialised corpus, such as that used in the present study, tend to be smaller, but vary in size. There is no word count limit for building a specialized corpus, and such corpora vary in size, according to the genre they represent and the restrictions in the text types and timeframe involved in a specific study. For example, the *Cambridge and Nottingham Corpus of Discourse in English* (CANCODE) has five million words, and the *Michigan Corpus of Academic Spoken English* (MICASE) has approximately 1.8 million words. While most specialised corpora are small in size, and are between 20,000 to 200,000 words (Aston, 1997), they can be larger due to restrictions in text type, as they are dictated by topic, genre, or both. In addition, they are often collected independently by researchers to examine particular language use in a specific genre. Nevertheless, a corpus is usually small as it does not seek to represent the language as a whole; instead, it attempts "to be representative of a given type of text" (Hunston, 2002, p.14).

Arguably, there are some reasons why general corpora do not represent the aims of the present study. Firstly, those general corpora include a variety of writing genres that are beyond the aim of the study. Secondly, because analysing the LBs used in argumentative essays is the main aim of the study, those corpora are not considered as representative of this particular writing genre. Although some large-scale corpora (e.g., BAWE corpus) considered representative of student writing, the data still includes different writing genres. Therefore, in my opinion, there is a need to compile a specialised corpus so that the comparisons between argumentative essays written by ESL learners from different proficiency levels can be made. With the compilation of specialized corpora, a specific genre, argumentative essays, can be analysed.

Accordingly, the present study fell under the category of a smaller, specialized collection of texts. It was compiled specifically for the purpose of this study, and sought to describe language use in English language centres (ELCs). It was representative, as far as possible, of the language variety of different CEFR levels, although was not generalizable to the entire range of language variety. Choosing to compile small corpora, based on specific instructional needs, entails a considerable amount of work, but if the tasks involved are clearly defined, it is achievable. Such corpora are able to identify, for example, specific patterns, phraseology, and the frequency of use of relevant words, and therefore align with the definition provided as they make it possible to "identify in what respects learners differ from each other" (Hunston, 2002, p.15), in the case of the present study at different ESL learner levels.

To conclude, the sub-corpora used in this study were learner sub-corpora that were specialized in nature, and contained samples of written language composed by EFL learners at different CEFR levels (B1, B2, and C1), which satisfied the definition of a learner corpus, and controlled for a wide range of variables.

4.9.3 Corpus size

Corpus size was a crucial feature for this study for determining whether the subset samples were representative. However, the concept of representativeness is vague, and primarily concerns the need to consider the main features that differentiate a corpus from any other corpora. As Biber et al. (1998a, p.246) explained, "a corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language", noting the difficulties associated with corpus compilation and the need for it to be "representative". Corpus size remains a contentious issue in terms of ensuring the validity of corpus analysis. Sinclair (2004 cited in Koester, 2010, p.66) argued that "small is not beautiful; it is merely a limitation". However, other researchers, claimed that a small corpus is valuable in linguistics research for certain purposes (Howarth, 1996; Chen, 2008). For instance, a relatively small, specialised corpus can be useful for studying grammatical items, such as model verbs or pronouns that are very frequent. Baker (2006, p.28-29) stressed the importance of the quality of a corpus over the quantity of the sample or words collected, arguing:

> One consideration when building a specialised corpus in order to investigate the discursive construction of a particular subject is perhaps not so much the size of the corpus, but how often we would expect to find the subject mentioned within it … Therefore, when building a specialised corpus for the purposes of investigating a particular subject or set of subjects, we may want to be more selective in choosing our texts, meaning that the quality or content of the data takes equal or more precedence over issues of quantity.

In other words, a corpus is generally limited, as it examines a particular community's actual language use. It is also essential to ensure that only appropriate texts are included; the samples collected should be chosen carefully before building the corpus, in order to be able to draw reliable conclusions from the finding. As Hunston (2002, p.15) noted, general corpora "will include as wide of a spread of texts as possible", and will comprise "texts of many types". Flowerdew (2004) argued that the size of a corpus should correlate with the occurrences being studied. In other words, the size of a corpus depends on the features being examined. If it is a common, frequently used feature, a small corpus can be used, whereas if the feature is less frequently-used, a sufficiently large corpus is necessary for a fruitful analysis, and to obtain generalizable findings. Accordingly, given the ubiquity of LBs, it would be sufficient to use a small sub-corpus for the study of ESL learners in the present thesis.

4.9.4 Representativeness of the learner sub-corpora

Regardless of the corpus type, it should be designed to constitute specific types of discourse. According to Tognini-Bonelli (2001, p.57), there is general agreement among scholars who work on corpora that they should be representative of a certain population, and that the statements derived from the analysis of the corpus should be largely applicable to a larger sample of the language as a whole. Otherwise, "without representativeness, whatever is found to be true of a corpus is simply true of that corpus—and cannot be extended to anything else" (Leech, 2007, p.135). It is important to understand "what it is meant by representing for a corpus" Biber et al. (1998a, p.246) to achieve the representative of the target sub-corpus, which is determined by the purpose of the study. Therefore, a decision must be made in advance of what exactly will be included in the corpus (e.g., type, number of texts, and number of words in each text). According to Granger (2012, p.1), "'mixed bag' collections of L2 data present little interest".

Consequently, Tono (2003) advised the following design considerations for constructing learner corpora (Table 4.30). As the table shows, the design is divided into three categories: language-related, task-related, and learner-related, and was the design employed in the construction of ESL learners' corpora used in the present study.

Table 4.30. Corpus design considerations for learner corpora (Source: Tono, 2003, p.800).

Tono's (2003) consideration corpus design			**ESL learner corpora of this study**		
Language-related	Task-related	Learner-related	Language-related	Task-related	Learner-related
Mode (written / spoken)	Method of collection (e.g., cross-sectional /longitudinal)	Internal – cognitive (age /cognitive style)	Mode: Written	Method of collection: cross-sectional /longitudinal)	Internal – cognitive: 18 - 40
Genre (e.g., fiction / essay)	Method of elicitation (e.g., spontaneous / prepared)	Internal-affective (motivation / attitude)	Genre: Argumentative essay	Method of elicitation: spontaneous	Internal-affective: motivation
Style (e.g., narration / argumentation)	Use of references (e.g., access to dictionaries, source texts)	L1 background L2 proficiency	Style (e.g., argumentation	Use of references: N/A	L1 background: Varying proficiency: B1, B2, C1
Topic: general	Time limitation (e.g., fixed / free / homework)	L2 environment ESL/EFL / level of school)	Topic: varying	Time limitation: N/A	L2 environment: ESL, intermediate and advanced levels

As shown in Table 4.30, the learners' sub-corpora in the present study were designed to be representative of B1, B2, and C1 ESL learners' academic writing. Therefore, three sub-corpora were compiled in order to identify and compare the target bundles.

4.9.5 Balance

A learners' corpus is a "collection of authentic machine-readable texts (written data) which is sampled to be representative of a particular language or language variety" (Mcenery et al., 2006, p.5). After ensuring that the sub-corpora used in this study were as representative as possible for the purpose of the study, namely the populations of

ESL learners' academic writing, it was necessary to address the concept of the corpus' 'balance'. The concept of balance concerns the percentage of varying texts or topic types that seek to represent a specific type of language. As Leech (2002, p.5) noted, "the subsamples or the sub-corpora of different language varieties must in some sense be proportionate to their importance in the language-importance is the difficult word there".

The importance of considering the balance of a corpus includes the requirement that the sub-samples of a corpus are "proportional in size to the parts of the variety/register/genre the corpus represents" (Scheepers, 2014), including texts that have been compiled in a "natural communicative setting" (Gilquin and Gries, 2009). In order to achieve a balanced corpus, the principle should be applied of collecting samples for the corpus that represents one register. As Hyland and Tse (2007) explained, examining a specific type of text from one genre can be more valuable pedagogically than focusing on and analysing general academic English.

The sub-corpora in the present study included texts that were representative of ESL learners' academic writing from B1, B2, and C1 level students' argumentative essays. These texts were chosen as they represented the academic writing genre of ESL learners that requires students to investigate a topic, provide evidence, and briefly establish a position on the topic. In order to obtain a sample of responses in typical argumentative essays, the participants in the cross-sectional study element of this project were required to select an essay question from given topics. In total, 621 essays composed by ESL learners studying at various language centres around the UK in 2018-2019 were collected to create the sub-corpora. The essays were composed on a range of different topics to increase the generalizability of the topics in building the sub-corpora (See section 4.5.3).

All the essays were re-rated, to ensure that they were incorporated under the correct CEFR levels. Therefore, it was possible to create a sub-corpus comprised an equal number of essays, to achieve balance in the corpus size. However, it would have been difficult to control the total number of words in each sub-corpus, as the number of words in each essay ranged between 200-400 words. In contrast to the pilot study, the decision was made to include an equal number of essays under each CEFR level. As far as possible, the approach adhered to situational and linguistic criteria to achieve balance in the sub-corpora.

4.10 Chapter conclusion

This chapter documented the methodology that formed the framework of this research, in order to compare the use of LBs by ESL learners at levels B1, B2, and C1, especially regarding whether there were noticeable variations and developments between the learners' levels in their written language. This study design was cross-sectional and also longitudinal in nature, and sought to analyse the LBs identified in three ESL learner groups' written English production quantitatively and qualitatively. The next chapter presents the findings, and discusses the production of LBs of the B1, B2, and C1 level ESL learners, by comparing the quantity and variety of use of LBs in the written production between the sub-corpora.

5 Results

5.1 Introduction

This chapter reports the findings of this study regarding the forms, structures, and functions of lexical bundles (LBs). The study investigated the variations and the development in the use of LBs in ESL learners B1, B2 and C1 academic writing. Employing the analytical procedures discussed in 4.8, the characteristics of the LBs used by B1, B2, and C1 ESL learners were examined quantitatively and qualitatively. The results are presented and discussed in this chapter, according to the research questions RQs provided in section 1.5.

This chapter is subdivided into four main sections according to the RQs. The first section discusses cross-section study employed to explore the differences in the use of LBs between the B1, B2, and C1 ESL learners' written production. The first section 5.2, examines the frequency of LB usage in the ESL learners' sub-corpora, including an assessment of the most frequent LBs, shared bundles, overall frequency, frequency range, and finally a comparison of the target bundles in the ESL learners' sub-corpora with a reference corpus (RC), namely the British Academic Written English (BAWE) sub-corpus. The subsequent section 5.3 discusses the grammatical variations of the bundles identified in the three ESL sub-corpora, followed by an investigation of the relationship between the structural and functional categories. Next, section 5.4 examines whether a specific bundle, overused or underused, acts as a key, by comparing the target sub-corpora with the RC using *WordSmith* tool (WST). Finally, section 5.5 employs a longitudinal study to explore the change in the frequency of LB usage across the proficiency levels over time. To aid discussion, each section will contain results and discussions guided by the main research questions; a summary for each finding will also be presented as bullet points before another section is introduced.

5.2 Frequency analysis (Cross-sectional Study)

5.2.1 Frequency of lexical bundles in the ESL learners' sub-corpora

The first research question of this study asked, 'What are the most frequent three- and four-word bundles found in ESL learners B1, B2 and C1 levels argumentative essays?'. As discussed earlier, the study sought to compare the three sub-corpora, and also to compare them with an RC that was used as a benchmark. It should be noted that because the three ESL learners' sub-corpora differed in size, all the raw frequencies generated from each sub-corpus were normalised to a frequency in 100,000 words, for the sake of comparison, as discussed in section 4.8. The top 20 three- and four-word LBs most frequently used by the ESL learners in their academic essays are listed in descending order in Table 5.1 and Table 5.2, along with their normalised frequency per 100,000 words (Complete lists of the identified bundles in Appendix D).

Table 5.1. The 20 most frequent three-word bundles in B1, B2 and C1 sub-corpora. (Freq = normalised frequency)

B1		B2		C1	
Bundle	**Freq**	**Bundle**	**Freq**	**Bundle**	**Freq**
a lot of	91	a lot of	98	I think that	399
on the other	62	on the other	84	first of all	206
one of the	52	there are many	62	second of all	173
there are many	46	they do not	58	I believe that	150
it is not	44	first of all	54	I think it	121
first of all	42	point is that	48	it is a	113
I want to	42	I want to	48	to sum up	97
there is a	38	in this essay	42	on the other	93
in this essay	36	in the world	40	in order to	84
the use of	36	in order to	38	his or her	72
as well as	34	one of the	36	the opportunity to	64
it is a	34	some of the	34	in addition to	56
as a result	32	there is no	34	do not have	54
in order to	32)	do not have	32	I do not	53
day by day	30	this essay will	32	I will give	53
most of the	30	we need to	32	to support my	53
of the world	30	in this world	30	I want to	51
that it is	30	there is a	30	in the following	51
I do not	28	as a result	28	a lot of	47
be able to	28	it is a	28	be able to	47

Table 5.2. Comparison of the most 20 frequent 4-word bundles in ESL learners' sub-corpora. (Freq = normalised frequency)

B1		B2		C1	
Bundle	**Freq**	**Bundle**	**Freq**	**Bundle**	**Freq**
on the other hand	60	on the other hand	78	I think it is	103
in this essay I	32	in this essay I	42	to sum up I	98
in the field of	24	I would like to	16	on the other hand	80
first of all it	18	is one of the	16	I think that the	51
I am going to	18	one of the most	16	in the following paragraphs	51
one of the most	16	there are a lot	16	reasons to support my	45
a second point is	12	a second point is	14	in conclusion I think	41
anywhere in the world	12	different from each other	14	on the one hand	39
as a result of	12	they are aware of	14	reasons which I will	39
I do not think	12	this essay will examine	14	when I was a	37
increasing day by day	12	a third point is	12	first of all I	35
a negative effect on	10	another point is that	12	I am sure that	33
a wide range of	10	all over the world	10	second of all I	33
all over the world	10	at the same time	10	as a result of	29
and as a result	10	first of all I	10	aspect of this is	29
can be used to	10	if you want to	10	in this essay I	29
for the betterment of	10	in order to reduce	10	have the opportunity to	27
I want to become	10	in the form of	10	I think that every	27
I would like to	10	to go to a	10	I will give my	27
is one of the	10	we can say that	10	is one of the	27

As Table 5.1 and Table 5.2 show, the usage frequency of the 20 most frequent three-word LBs ranged between 28-91/100,000 at B1 level, 28-98/100,000 at B2 level, and 47-399/100,000 words at level C1. Meanwhile, the usage frequency of the most commonly used four-word bundles ranged between 10-60/100,000 at B1 level, 10-78/100,000 at B2, and 27-103/100,000 at C1 level.

The most frequent three-word bundles in the B1 and B2 sub-corpora was *a lot of*, which accounted for 91/100,000 and 98/100,000, respectively whereas, the bundle *on the other hand* was present in the top four-word lists of both levels. It accounted for 60/100,000 in B1 and 78/100,000 at B2, followed by the bundle *in this essay I*, at 32/100,000 at B1, and 42/100,000 at B2 levels. Meanwhile, the most frequent three-word bundles in the C1 sub-corpus were *I think that*, which accounted for 399/100,000, followed by *first of all*, with 206/100,000, followed by *second of all, I believe that, I* *think it,* and *it is a*, which all occurred more than 100/100,000 words. In the four-word bundle list, the bundle *I think it is* took the lead in the C1 sub-corpus, with 103/100,000, followed by *to sum up I*, which occurred 98/100,000 times. Thus, it can be seen that C1 writers favoured different bundles than B1 and B2 writers.

The growth of the frequency use of LBs across the proficiency levels was also noteworthy in these lists, as there was a wide difference between the learners at B levels and C1 level. Again, strong evidence of the increased use of LBs was identified, as the C1 writers employed more bundles than the other writers. This may infer that higher ESL writers show a higher preference for the use of LBs in their academic writing. For instance, the usage frequency of the 20 most common three-word bundles at B1 sub-corpus was 797/100,000 words, 888/100,000 at B2 level, and 2,037/100,000 at C1 sub-corpus. The following examples (1–6) illustrate the use of such bundles in each sub-corpus:

- As a result, *a lot of* people will stop buying from them and they will look for another company. (B1 sub-corpus, essay 7)
- *A lot* of gyms, parks, and sports centres can be a good thing for people health. (B2 sub-corpus, essay 178)
- *I think* that these days, it is very important to know how to use computer technology. (C1 sub-corpus, essay 18)

Finally, strong evidence of the increasing use of LBs in the C1 sub-corpus was identified when assessing the frequency occurrence of the three- and four-word bundles. For example, the frequency of the top six three-word bundles in the C1 list exceeded 100/100,000 times, including one that accounted for 399/100,000 words, and another that accounted for 206/100,000 words. Meanwhile, the top 20 most frequent four-word bundles in C1 sub-corpus accounted for 103/100,000 words.

A comparison of the most frequent bundles also showed that the bundle *a lot of* ranked first place in B1 and B2 sub-corpora, occurred 91/100,000 times at B1 sub-corpus; and in 98/100,000 times in the B2 sub-corpus. However, this bundle ranked 19th place in the C1 sub-corpus, only accounted for 47/100,000. This meant that the bundle *a lot of* exhibited a high-frequency occurrence at B1 and B2 levels, compared with C1 level. In contrast, the bundle *I think that* was foremost at C1 level, at 399/100,000 words. This bundle was not found in the top 20 most frequently used three-word bundles in the B1 and B2 sub-corpora. This meant that C1 writers exhibited different use of LBs than B1 and B2 levels.

In terms of four-word LBs, the bundle *on the other hand* showed a relatively similar frequency across the levels, and came at the top of B1 and B2 lists. This bundle was usually more frequently used in argumentative essays as a way of addressing the second part of a two-part problem, situation, or solution. In the C1 sub-corpus, the bundle *I think it is* had the highest frequency, with above 100/100,000 words, appearing in the highest number of texts (42), occurring in about 27% of all 155 texts. This bundle was not found in the top 20 lists of the other levels.

The analytical insights relevant to the first research question highlighted great similarities of the use of LBs at B1 and B2 essays, and with great difference than C1 essays. Judging from the components making up the bundles, it can be seen that B1 and B2 lists contained quantifier bundles to indicate the quantity of something (*a lot of, there are many*). These expressions are important because they let us express the quantity of something. However, the increased use of quantifier expressions by B1 and B2 levels can be attributed to overgeneralisation in B1 and B2 writing, which showed an informal style in learner writing. On the other hand, C1 writing list, mainly composed of self-mentioning bundles (*I think that, I think it is*), which provide interpersonal information. This can be probably attributed to the type of essay that required the writers to convey opinions to the readers.

The findings of this section could contribute to support the difference in the use of LBs across the levels. The next section compares the overall usage frequency between the sub-corpora.

5.2.2 Overall frequency of lexical bundles

In total, there were 465 written essays (16,8791 words) produced by ESL learners B1, B2, and C1 sub-corpora, in which 1,667 types of target bundles were identified, representing 10,429 tokens (instances) of total bundle usage by all ESL learners across the three levels. The final lists were composed primarily of three-word expressions, accounting for 85% (B1), 83% (B2), and 71% (C1) of the whole bundles sub-corpora. As might be expected, the length of the bundles was inversely related; the three sub-corpora were comprised a large number of three-word bundles.

A total, 1,667 bundles were identified in the ESL learners' sub-corpora used for this investigation, representing 10,429 tokens, as presented in Table 5.3. Column 1 shows the ESL learner groups, according to their level, and the LB lengths, while column 2 provides the number of LBs identified in each sub-corpus. Column 3 shows the percentage of LB types that occurred in the learners' sub-corpora, calculated as the number of bundles counted in column 2, multiplied by three for three-word bundles, or four-word bundles, then divided by the number of words in each sub-corpus. The 'Frequently representative' in column 4 refers to the total occurrences of all types of LBs in each sub-corpus, after applying the removal criteria (See section 4.6) and being normalised per 100,000 words, followed by the percentage of the target bundles in each sub-corpus.

Table 5.3. Total number of bundle types and tokens in the ESL learners' sub-corpora. TBs= refers to total bundles)

Length/Corpus	Total number of words	Bundle type	Frequency representative	Percentage of TBs in sub-corpus
3-word B1	50321	427	4457	13%
4-word B1		76	696	3%
3-word B2	49871	394	4391	13%
4-word B2		80	740	3%
3-word C1	51415	487	7196	23%
4-word C1		203	2544	10%

When assessing the overall count of the LBs, it was apparent that the C1 writers' essays contained more types and more tokens of three- and four-word LBs than those of the learners at levels B1 and B2. Interestingly, there were also gradual increases in the use of four-word bundles across the sub-corpora, with a sharp rise in the C1 level. Examining the number of target bundle types and normalised frequency of the corpora in Figure 5.1 and Figure 5.2 showed an increased use of LBs by the higher proficiency students (C1). Although the amount of three-word bundles used by the B1 writers dropped slightly in the B2 writers' essays, it grew in the C1 level essays. Meanwhile, the increase in the number of four-word bundles used by the B1 level participants, compared with the B2 level, gradually increased, with a far steeper growth curve in the number of four-word bundles used at C1 level.

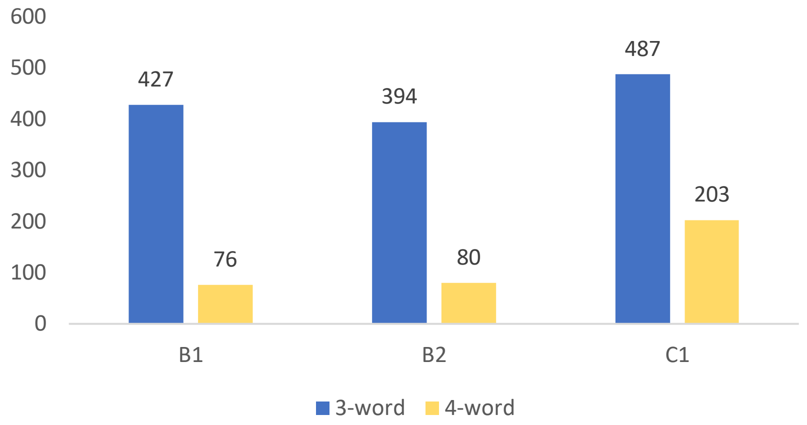


Figure 5.1. Target bundle types used by the ESL learners in this study.

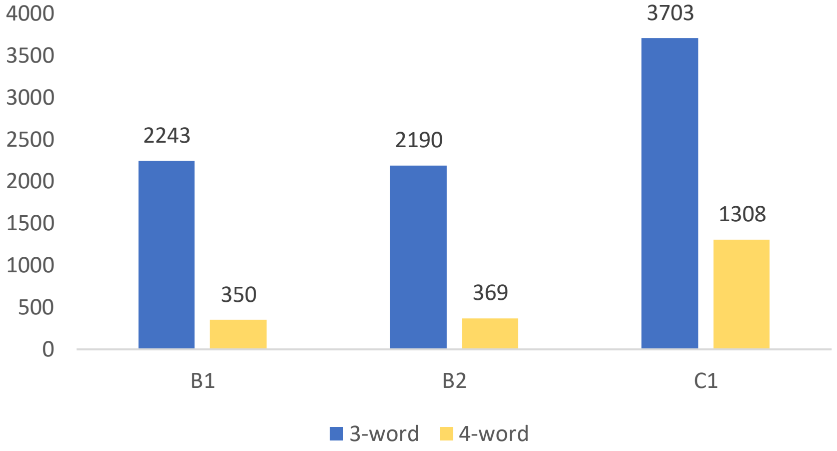


Figure 5.2. Normalised tokens of the LBs used by the ESL learners in this study.

As the results show, the usage frequency of the bundles varied, therefore, in order to determine the degree to which these differences between B1, B2 and C1 levels were statistically significant, a log-likelihood (LL) analysis was conducted. The LL statistical test took into account the frequencies weighted over two different corpora, and the LL value reflected how much more likely it was that the word frequency between the corpora was different than it was the same. The initial LL analysis was conducted between the B1and B2, sub-corpora to test the overall frequency difference accounted between them, as displayed in Table 5.4.

Table 5.4. LL ratio of the B1 and B2 sub-corpora. Sig = significant value

	O1 B1 data	**%1**	**O2 B2 data**	**%2**	**LL ratio**	**Sig.**
LBs	503	1	474	0.95+	5.17	p < 0.0001

In Table 5.4, O1 and O2 refer to the overall frequency of LBs in the B1 and B2 sub-corpora. The 1% and 2% values show the relative frequencies in the texts of each sub-corpus; the relative frequency of 1 indicates that there was around 1 LB in each 100 words in the B1 sub-corpus. Similarly, there were 0.95 LB in every 100 words in the B2 sub-corpus. The results of the LL ratio revealed a slight overuse of LBs in the B1 level essays, with an LL value of 0.62, indicating the difference between the B1 and B2 sub-corpora, in terms of the LBs' frequency (*P* < 0.05), and their overuse in the B1 level writing, compared with the B2 writing. The next comparison was between the B levels and the C1 level, as presented in Table 5.5.

Table 5.5. LL ratio of the B levels and the C1 level sub-corpora. Sig = significant value

	O1 B1 data	**%1**	**O2 C1 data**	**%2**	**LL ratio**	**Sig.**
LBs	503	1	690	1.34-	26	p < 0.0001
	O1 B2 data	**%1**	**O2 C1 data**	**%2**	**LL ratio**	**Sig.**
	474	0.95	690	1.34-	34	p < 0.0001

As expected from the absolute frequency differences between the sub-corpora, the LL value between the two sub-corpora revealed a significant difference in LB usage, as the C1 writers used 1.34 LBs in every 100 words. The LL ratio between the B1 and C1 levels found +26, and a LL value +34 between levels B2 and C1, revealing the differences between the ESL learners' use of LBs. Although the difference shrank between the B1 and B2 levels, the difference between these groups increased significantly from the B levels to the C1 level. These results supported Hypothesis one, as there was a clear difference between the groups, with evidence of a relationship between academic performance and the use of LBs.

The next section explores the frequency range of the bundles identified, highlighting the different use of LBs across the CEFR levels.

5.2.3 The frequency range of lexical bundles across the sub-corpora

An initial assessment of the bundles identified in the three sub-corpora indicated that their frequency occurrences ranged from six to 399 times per 100,000 words. The frequency of these bundles was classified under five normalised bands: below 10, 10-30, 30-50, 50-70, and over 70, to determine which band included the majority of the bundles. This could be useful to compare the distributional pattern between the sub-corpora, in terms of the normalised frequency of the bundles identified. Figure 5.3 summarises the overall count of the LBs used by the groups.

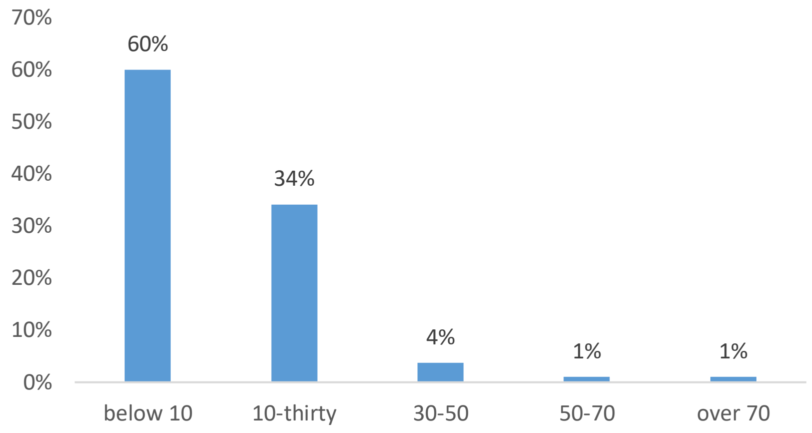


Figure 5.3. Range of frequency occurrences of the target bundles extracted from the ESL learners' sub-corpora.

As shown in the figure above, most of the bundles identified in the ESL learners' sub-corpora were those with a frequency of below 10, which accounted for 60% of the total bundles. This is followed by bundles with a frequency between 10 - 30 occurrences per 100,000 words, which accounted for 34% of the corpora. Finally, there were only a few bundles with a frequency of over 30 times per 100,000 words, accounting for 6%, with the highest frequency of 399, which was a bundle used by the ESL learners at C1 level. In other words, bundles used by ESL learners in written argumentative essays are mostly low-frequency ones. While most of the LBs occurred fewer than 10/100,000 words, there was a significant increase in extremely high-frequency LBs in the C1 level essays. Table 5.6 below presents the LBs identified across the sub-corpora.

Table 5.6. Normalised frequency percentage of the LBs detected in the sub-corpora.

Frequency band	**Three-word bundles**			**Four-word bundles**		
	B1	B2	C1	B1	B2	C1
below 10 occurrences	27%	24%	26%	5%	6%	12%
10-30 occurrences	25%	24%	32%	4%	3%	13%
30-50 occurrences	24%	21%	40%	2%	2%	13%
50-70 occurrences	12%	18%	53%	6%	0%	12%
over 70 occurrences	6%	12%	59%	0%	6%	18%

As the table above shows, the percentage of LB usage below 10, when compared between the sub-corpora, was similar across the levels, but as the usage frequency increased, the three-word LBs in the C1 sub-corpus took the lead with up to 59%, while the other lists decreased the higher they went. This may be an indication that the C1 learners used more high-frequency LBs than the other students, as seven bundles occurred more than 100/100,000 words. These results reflected those of previous studies that found that lower-level learners are more likely to use a narrower range of LBs than those at higher proficiency levels (Chen and Baker, 2010; Ädel and Erman, 2012; Novita and Kwary, 2018). In other words, it can be argued that the demands of C1 level require writers to invest more rhetorical effort in building their arguments, which in turn causes the conventionalisation of certain bundles at this level, over time.

In short, it can be assumed that there is a direct relationship between the use of three-word LBs and academic performance, as the level increases, students employ LBs with increasingly high frequency. In order to understand the difference between the ESL learners' levels, it was necessary to compare the shared bundles used in their writing, as discussed in the next section.

5.2.4 Shared bundles

There were noticeable variations in the use of LBs across the sub-corpora. Out of the 1,667 bundles identified in the three sub-corpora, 513 (39%) three-word bundles and 77 (21%) four-word bundles were shared by the sub-corpora. In total, 160 bundles were shared by the B1 and B2 sub-corpora, the B1 and C1 sub-corpora shared 133, and 115 were shared by the B2 and C1 sub-corpora. In addition, 82 bundles were noteworthy, as they occurred in all three sub-corpora, as shown in Appendix E

Of the 82 bundles shared across the sub-corpora, 40 were used increasingly in the C1 writing, 21 bundles were used more in the B2 writing, 20 bundles were used more by B1 level learners, and one bundle was used equally across the levels. For example, the bundles *first of all*, which was used 206 times in the C1 level essays, and fewer than 60 times in the B1 and B2 level essays. Meanwhile, the bundle *a lot of* showed an increased use in the B1 and B2 sub-corpora at 91/100.000 and 98/100,000 times, respectively, and occurred only 47/100,000 times in C1 sub-corpus. Despite certain differences between the bundles used, this large number of common bundle types across the three sub-corpora emphasised the high reliance of the ESL learners on LBs in particular bundles to convey their messages. Of the top 20 most frequent three-word bundles, six bundles were shared between the three levels (*a lot of, on the other, first of all, I want to, in order to*, *it is a*), as shown in Table 5.7. The bundle with the greater difference in usage of shared bundles across the levels was the bundle *first of all*, accounted for 42/100,000 in B1, 54/100,000 words in B2, 206/100,000 words in C1. Moving to the most frequent 4-word bundles shown in Table 5.8. It can be seen that the sub-corpora shared three LBs (*on the other hand*, *in this essay I*, and *is one of the*).

Table 5.7. Shared bundles in the most frequent 20 three-word bundles in ESL sub-corpora. (Bold = Shared in all sub-corpora, italic = Shared in two sub-corpora, Freq= Normalised frequency per 100,00 words)

B1		**B2**		**C1**	
Bundle	**Freq**	**Bundle**	**Freq**	**Bundle**	**Freq**
a lot of	91	**a lot of**	98	I think that	399
on the other	62	**on the other**	84	**first of all**	206
one of the	52	*there are many*	62	second of all	173
there are many	46	they do not	58	I believe that	150
it is not	44	**first of all,**	54	I think it	121
first of all,	42	point is that	48	**it is a**	113
I want to	42	**I want to**	48	to sum up	97
there is a	38	*in this essay*	42	**on the other**	93
in this essay	36	in the world	40	**in order to**	84
the use of	36	**in order to**	38	his or her	72
as well as	34	*one of the*	36	the opportunity to	64
it is a	34	some of the	34	in addition to	56
as a result	32	there is no	34	*do not have*	54
in order to	32	*do not have*	32	*I do not*	53
day by day	30	this essay will	32	I will give	53
of the world	30	in this world	30	**I want to**	51
that it is	30	*there is a*	30	in the following	51
I do not	28	*as a result*	28	**a lot of**	47
be able to	28	**it is a**	28	*be able to*	47

Table 5.8. Shared bundles in the 20 most frequent four-word LBs in the ESL sub-corpora. (Freq = Normalised frequency; bold = shared in the three sub-corpora; italics = shared in two sub-corpora)

B1		B2		C1	
Bundle	**Freq**	**Bundle**	**Freq**	**Bundle**	**Freq**
on the other hand	60	**on the other hand**	78	I think it is	103
in this essay I	32	**in this essay I**	42	to sum up I	89
in the field of	24	*I would like to*	16	**on the other hand**	80
first of all it	18	**is one of the**	16	I think that the	51
I am going to	18	*one of the most*	16	in the following paragraphs	51
one of the most	16	there are a lot	16	reasons to support my	45
a second point is	12	*a second point is*	14	in conclusion I think	41
anywhere in the world	12	different from each other	14	on the one hand	39
as a result of	12	they are aware of	14	reasons which I will	39
I do not think	12	this essay will examine	14	when I was a	37
increasing day by day	12	a third point is (that)	12	*first of all I*	35
a negative effect on	10	another point is that	12	I am sure that	33
a wide range of	10	*all over the world*	10	second of all I	33
all over the world	10	at the same time	10	*as a result of*	29
and as a result	10	*first of all I*	10	aspect of this is	29
can be used to	10	if you want to	10	**in this essay I**	29
for the betterment of	10	in order to reduce	10	have the opportunity to	27
I want to become	10	in the form of	10	I think that every	27
I would like to	10	to go to a	10	I will give my	27
is one of the	10	we can say that	10	**is one of the**	27

To examine the importance of the shared bundles identified across the ESL learners' levels in academic writing. The shared bundles' lists compared with the findings of similar previous studies that examined the use of LBs (L2) learners regardless of the writing genre, or of their first language (L1). One of these studies were similar to the present study; that conducted by Chen and Baker (2016), who used the CEFR for determining the proficiency levels. They address argumentative essays restricted by L1 Chinese learners retrieved from the Longman Learner Corpus (LLC) publishes between 1990 and 2002 to examine L2 English data across B1, B2 and C1 levels. The second study conducted by Appel (2011a), who used a corpus composed of argumentative essays written by test takers of the Canadian Academic English Language (CAEL) assessment. The CAEL tests were divided into two broad bands, those scoring 40 and below (Low-Level Corpus - LLC), and those scoring 50 and above (High-Level Corpus - HLC). The final study was conducted by Du (2013), who compared corpora of timed essays (TEM) composed by Chinese EFL learners at two different university levels with the Academic Formulas List created by Simpson-Vlach and Ellis (2010). In order to conduct a fair comparison, the frequency occurrences of the shared bundles shown in Table 5.9 were normalised to 100,000 words.

Table 5.9. Comparison of the shared bundles with previous studies.

Shared LBs	**Present thesis**			**Du (2013)**		**Appel (2011a)**		**Chen and Baker (2016)**		
	B1	B2	C1	TEM4	TEM8	LLC	HLC	B1	B2	C1
a lot of	91	98	47	90	62	97	31	30	18	-
first of all	42	54	206	33	24	48	-	-	-	-
I want to	42	48	51	-	-	-	-	-	-	-
in order to	32	38	84	55	62	12	35	-	-	6
it is a	34	28	113	55	47	-	-	-	5	-
is one of the	10	16	27	24	26	33	33	17	20	16
in this essay I	32	42	29	-	-	-	-	-	-	-
on the other hand	60	78	80	33	20.3	38	-	14	2	3.2

As shown in the table above, the LBs *is one of the*, *on the other hand, a lot of, first of all*, and *in order to* were shared by all the studies, albeit with a different normalised frequency in the various learner groups. This result shows that some bundles used by ESL learners B1, B2 and C1 may constitute important building blocks in the participants' discourse. A possible explanation for the similarities between the shared bundles may be related to the fact that in academia, a common group of LBs is generally usually used, and can be seen as a necessary element, thus such sequences appear widely in the language use, regardless of the genre or the L1 of the users.

Besides the found similarities, each sub-corpus is characterised by a number of specific bundles. B1 and B2 writers do not tend to use 336 of the C1 target bundles which seem to be exclusively used by C1 writers in their essays (e.g., *it is necessary, a variety of*). On the other hand, B1 and B2 writers used 237 and 222 LBs which are different from those found in the other groups. One could assume that ESL learners in each level have specific bundles that seem to be exclusively used in their essays. In other words, ESL learners of different proficiency levels might overuse certain bundles that they are exposed to in their learning.

One way to realize the variations in the language use of different levels divisions is their selection of LBs. The fact that ESL learners in each level rely on a specific set of word combinations (which were not found in the other sub-corpus) would support the idea that they have their own peculiarities and ways of organizing the discourse to deliver their messages. The next section considers the use of LBs across the ESL learners' levels in comparison with the BAWE sub-corpus.

5.2.5 Comparison of the target bundles in ESL learners' sub-corpora with the BAWE sub-corpus

This section compares the usage frequency of the LBs in the ESL learners' sub-corpora with the BAWE, in order to distinguish specific features in the ESL learners' writing. The bundles derived from the ESL sub-corpora were searched for in the RC list to identify their respective frequency of usage. The frequency statistics of bundles (types) and frequency (Tokens) identified in the corpora are shown in Table 5.10.

Table 5.10. Bundles (types) and frequency (token) identified in the corpora. (TBs = total bundles; Freq = Normalized frequency)

Length/Corpus	Bundle types	Percentage of TBs in corpus	Freq
3-word B1	427	13%	4457
4-word B1	76	3%	696
3-word B2	394	13%	4391
4-word B2	80	3%	740
3-word C1	487	23%	7196
4-word C1	203	10%	2544
3-word BAWE	827	11%	3643
4-word BAWE	127	2%	481

After applying the exclusion criteria to the RC lists, a total of 827 types of three-word bundles, totalling 3,634/100,000 words, were identified in 524,284 words, constituting approximately 11% of the total words in the BAWE corpus. In addition, 127 different four-word LBs were extracted, accounting for 481/100,000 words, constituting approximately 2% of the total words in the corpus. Interestingly, as shown in column 4 of Table 5.10, the BAWE writers employed fewer and less varied LBs than the ESL learners.

When the bundle types identified in the ESL learners' sub-corpora were compared with those of the BAWE sub-corpus, of the 827 three-word bundles identified in the BAWE sub-corpus, 239 bundles were located in the ESL learners' sub-corpora, and 53 bundles occurred in the four lists Appendix F . In fact, less than 30% of the bundles used by the ESL learners' writing were located in the BAWE writing. Separately, 119 bundles (14.3%), 138 bundles (16.6%), and 146 bundles (17.4%) were found in B1, B2, and C1 writing, respectively. In terms of the four-word LBs, out of the 127 bundles extracted from the BAWE corpus, 23 LBs were found in ESL learners' sub-corpora. In total, 11 bundles (8.6%), 12 bundles (9.4%), and 16 bundles (12.5%) were shared by the B1, B2, and C1 sub-corpora, respectively. The higher use of distinctive bundles by the ESL learners in comparison with the BAWE can be attributed to their lower writing competence (Chen and Baker, 2010; Staples et al., 2013).

The result of the most ten frequent three- and four-word LBs in the ESL learners' sub-corpora and the BAWE sub-corpus were given in Table 5.11 and Table 5.12.

Table 5.11. The ten most LBs whose normalised frequency increased in the ESL learners' sub-corpora.

Bundles	**Normalised frequency**			
	BAWE	**B1**	**B2**	**C1**
I think that	3	18	20	399
a lot of	5	91	98	47
on the other	19	62	84	93
there are many	7	46	62	10
one of the	26	52	36	47
they do not	6	14	58	41
it is not	19	44	24	37
be able to	14	28	22	47
there is a	26	38	30	27
in addition to	3	8	8	56

Table 5.12. Four-word bundles who normalised frequency increased in the ESL learners' sub-corpora.

Bundles	**Normalised frequency**			
	BAWE	B1	B2	C1
on the other hand	19	60	78	80
in this essay I	3	32	42	29
is one of the	3	10	16	27
as a result of	7	12	8	29
I would like to	2	10	16	14
one of the most	3	16	16	8
at the same time	7	6	10	23

As Table 5.11 and Table 5.12 show, the C1 writers used LBs more frequently than the other writers, although there was a fluctuation at the in-between levels. The results supported the previous finding that the C1 writers not only used a higher number of LBs, but also used a greater variety in their academic essays. Meanwhile, a closer inspection of the concordance lines showed that most of the shared bundles had a similar pattern of use in the ESL learners' sub-corpora and the BAWE sub-corpus. For example, all the writers generally used the bundle *in this essay I* at the beginning of the sentence to quantify what the essay concerned, and as a guide to the reader regarding its content, as in the examples below:

- *In this essay*, I will suggest some step each of us can take and some ways to motivate others to do the same. (B1, essay 45)
- *In this essay* I will discuss some of the reasons for this and give some suggestion about how to tackles this problem. (B2, essay 115)
- *In this essay* I will first focus on the reasons why I agree with this statement and then list a few points why from my opinion in some cases is not true. (C1, essay 267)
- *In this essay* I will carry out a close textual analysis of the ball scene in Joseph L. Mankiewicz's film in relation to the theory of mise-en-scène and its adherence to the techniques of narration in Classical Hollywood Cinema. (BAWE, essay 2160B)

However, despite the similarity in the use of the shared bundles, there were certain qualitative distinctions in their use across the sub-corpora. For example, the bundle *I think that* was used across the levels, but in different ways. The ESL learners tended to link this bundle with a word or phrase to summarise an argument, or with a conjunction adverb to show how a sentence was related to what had already been stated; this was the case in over a third of the occurrences. Examples of the use of this bundle are as follows:

- To sum up, *I think that* all young people should have the chance to get a higher education. To take or not this chance must be up to them. (C1, essay93)
- In summary, *I think that* money still a good way to helps other people. Of course, it cannot be thrown at the problem, but every other form of aid is useless without cash. (B2, essay61)
- Finally, *I think that* it will be better for students to pick the subjects that they like and to only study them in school. (B1, essay142)

Conversely, the BAWE university students showed little variation in terms of the word or phrase connected to the bundle *I think that*, tending to use it more frequently at the beginning of the sentence to provide a transition, and to explain or express their thoughts, as shown in the following example:

- *I think that* there is no definite answer as to whether females are indeed more class conscious through their language use than men, but indeed there are some situations which suggest this. *I think that* it is important to look at other social variables to be able to fully assess whether gender and class alone affect speech. (BAWE, essay2145b)

Another distinctive usage difference across the learners was in the bundle *'is one of the'*. In the B1 and B2 learners' writing, the bundle was used primarily as the head of a noun phrase linked to a pre-modifier superlative adjective, such as most, biggest, greatest, whereas the C1 and the BAWE writers employed this bundle in a more complex structure to connect the ideas in the text. Examples of the use of this bundle are as follows:

- In my opinion, the Internet *is one of the* greatest people created. You have an access to libraries from your computer, you can speak with your friends around the world. (B1, essay167)

- This is probably why *is one of the* most popular tourist attractions in the world, and why I would love to visit this city someday. (B2, essay 252)
- Unemployment *is one of the* major causes for a country to be a third world country. More and more people graduate every year but not even 50% of those get the job that they were trained for (C1, essay368).
- In opposition to the neutral report of the dialogue, the rewrite introduces subjectivity through the presence of a narrator: this *is one of the* astronauts (from the original poem) who is telling his own view of the scene relating it to his personal story. (BAWE, essay 3057b)

Finally, the bundle *on the one hand* was associated with the bundle *on the other hand*, and used together when comparing two different facts, or two opposite ways of thinking about a situation. A careful examination of the concordance lines indicated that the bundle did not appear with its counterpart in ESL learners' writing, but was presented in the BAWE writing, as in the following example.

- *On the one hand* there is the solidarity created through the shared knowledge of the limerick's linguistic features and typical content composition, and especially through the shared enjoyment of its humorous effects, given that communal laughter is one of the major interpersonal binders of human culture. *On the other hand* the limerick creates a feeling of inclusion or exclusion through its degree of vulgarity (BAWE, essay6064).
- Village is a small area where there is lack of facilities. *On the other hand*, City is the area where every kind of facility is available. People who live in villages are quite simple but people of cities are modern. (B2, essay343)
- *On the other hand*, situation is totally opposite in villages. The people of the cities are simple and caring. Their lifestyle is very ordinary and simple. they are not provided with all facilities in village. (C1 essay397)

In summary, it was evident that the language in the ESL learners' writing was more descriptive, and a higher percentage of LBs were used by the three levels compared to the BAWE sub-corpus, with a significant increase in the C1 writing. As Foster et al. (2000, p.355) observed, "the more proficient speaker will be the person who can keep track of more complex micro-units", which is to say a person able to memorise and retrieve LBs quickly, to use them when necessary. Therefore, it could be claimed that using LBs when learning a language at high level not only renders L2 learners' language usage more 'native-like' (Pawley and Syder, 1983), but the ability to produce longer LBs quickly also helps them to achieve language competence. When the B1 and B2 writers sought to make their argument more comprehensible for a reader, this may necessitate the use of a higher proportion of LBs, as in the C1 writing. For example, *I think that, I believe that*, and *I think it* occurred more frequently in the C1 sub-corpus than the B1 and B2 sub-corpora. This finding relates to an issue discussed by Römer (2009, p.89), who asked "does nativeness matter when we are dealing with English in academia or are there other, perhaps more important aspects to consider that influence our performance in English academic settings?" While this research concerned the use of LBs across proficiency levels, the results of the present study suggested that students' level may play a vital role in language fluency. This result was supported by the appropriate and increased use of LBs in higher proficiency learners' academic essays.

5.2.6 RQ1 discussion

RQ1 What are the most frequent three- and four-word LBs found in ESL learners B1, B2 and C1 sub-corpora?

As discussed in sub-section 4.6.2, the initial search of the three- and four-word LBs in the ESL learners' sub-corpora, using a frequency threshold and dispersion criteria, yielded 499 LBs at B1, 374 at B2, and 690 at C1. This demonstrated that the ESL learners at C1 level increasingly employed LBs in their writing more than other levels. The most frequent three- and four-word bundles in B1 and B2 sub-corpora were the bundles *a lot of*, and *on the other hand* in contrast, The most frequent bundles in the C1 sub-corpus were *I think that* and *I think it is*. While there was a great similarity of the use of LBs between B1 and B2 levels, there was a big difference between those two levels and C1 level. Strong evidence of the different use of LBs was evident in the C1 writers' use of frequent and varies bundles than the other writers, indicating that the C1 writers showed a higher usage of LBs to build their arguments.

Further investigation of the log-likelihood analysis revealed significant differences in the use of LBs among the levels. Although these differences reduced between B1 and B2 levels, the difference between the groups increased significantly from the B2 level to C1 level. Therefore, the correlation between the variations of LBs and raising proficiency could be established. The difference between the ESL learners' levels was noticeable only between B2 and C1 levels. This means that ESL learners below B2 level shared similar language features. However, C1 writers include varied types with different bundle frequencies compared to other levels, showing that they are more aware of the use of these expressions in their writing.

These results supported the argument that the competent use of LBs reflects a strong degree of integration into the academic community (Chen and Baker, 2010; Chen and Baker, 2016). This finding partially addressed the question posed by Chen and Baker (2004, p.44) regarding the possibility of "a relationship between proficiency and the number of formulaic expressions used", by Cortes (2006) concerning whether the appropriate use of LBs contributes to the perception of good writing, and by Li and Schmitt (2009, p.98-99) regarding whether "the appropriate and diverse use of lexical phrases affects the evaluation of academic writing". While the latter's research showed no obvious connection between students' use of lexical phrases and the evaluators' assessment of their writing, the relationship between the use of bundles and performance in the present study suggested that the increased use of LBs may play a role in the assessment of student writing, particularly in advanced L2 learners C1 level. Further evidence of this was examined in the analysis of the longitudinal study.

It should be noted that these above studies are not directly comparable with the present studies due to the different L1 backgrounds of the participants, research environments; therefore, the comparisons of LBs frequencies should be treated with caution. For example, Chen and Baker (2016) compared essays produced by L1 Chinese students of L2 English, whereas the present study used essays produced by ESL learners from different L1 backgrounds. Thus, the result of the present study can be generalised to different L1 backgrounds more than the other ones.

Apart from the variation of LBs among the ESL learners B1, B2 and C1 sub-corpora, the findings confirmed the results of (Biber et al., 1999; Alipour and Zarea, 2013; Heng et al., 2014), who found that three-word LBs were the most prevalent bundle length used by L2 learners. A possible explanation for the overuse of three-word bundles may be related to the complexity of the production of longer bundles that causes language learners to avoid using them in their writing, as it requires effort and time for students to produce longer sequences. Indeed, B1, B2 and C1 writers in the present study produced twice as many four-word bundles. This result was not surprising, as Biber et al. (1999, p.992) noted that three-word LBs are extremely common, because they are "a kind of extended collocational association," while longer bundles are "more phrasal in nature and correspondingly less common".

In terms of the comparison of the target bundles' usage by expert writers, the comparison of the three- and four-word LBs in the present study in both the three ESL learners' sub-corpora and the BAWE sub-corpus identified certain differences in overall bundle prevalence, showing that the ESL learners exhibited a noticeably higher usage frequency than the writers in the reference corpus (RC), regardless of proficiency level. This indicated that the ESL learners employed more LBs that satisfied the specific purposes of argumentative writing. A possible explanation for this finding was that every mode (e.g., written); genre (e.g., argumentative essay); register (e.g., formal); and discipline (e.g., linguistics) engenders the "employment of] a distinct set of LBs, associated with its typical communicative purposes" (Biber and Barbieri, 2007, p.265). For example, argumentative essays, such as IELTS essays, undertaken in controlled conditions involve more aspects such as formulaic language and two-clause sentences than other genres of university academic writing (Demetriou, 2019). In contrast, university writing requires the writer to demonstrate their familiarity with the subject, and their ability to synthesise research (Turabian, 2013). This may increase the range of LBs used, but not their frequency.

Therefore, the greater use of LBs in the ESL learners' sub-corpora may be due to the way that argumentative essays are structured, requiring the use of LBs to enable the reader to follow an argument. For example, the use of certain bundles, such as *on the other hand* to introduce alternative points, or to indicate a conclusion using a bundle such as *to sum up*, is common in argumentative essays, help writer to state their argument and to deliver it clearly, as shown in the following examples, as shown in the examples below.

- *On the other hand,* parents have a great influence on children' success in school too. (C1, essay67)

- *To sum up,* helping reduce the effects of global warming is not only good for yourself but everyone and our home planet Earth. (B2, essay244)

- *In conclusion, I* think that parents are the best teachers for their children because they give their knowledge that cannot be learn from books. (B1, essay 145)

This assumption was also made by Cooper (2013); Pearson (2021) who noticed a similar increase in the proportion of bundles drawn from corpora of argumentative essays produced for IELTS Task 2 than in university students. As Nippold and Ward-Lonergan (2010) explained, writers of the argumentative essay seek to convince their readers to adopt a stance regarding a debatable issue. This can increase the use of certain bundles, such as first person pronoun, to state their position in relation to others or other common bundle to achieve essay coherence.

The results of the present study reflected those of certain previous studies that found a greater increase in the use of LBs, in terms of type and token, by L2 learners than by native or professional writers (Hyland, 2008a; Bal, 2010; Peromingo, 2012; Alipour and Zarea, 2013; Öztürk, 2014; Pan et al., 2016; Güngör and Uysal, 2016; Pourmusa, 2014). For example, Öztürk (2014) examined the use of LBs in English academic texts of non-native English speakers in a particular academic discipline and found that the non-native English writers used LBs two times more than the native writers. Similarly, Peromingo (2012) assessed the use of LBs by EFL learners from different L1 backgrounds, compared with that of native writers, and found that the former used significantly more LBs in their argumentative essays than the latter. Therefore, it is not the case that ESL learners studying in the UK use LBs more than others, rather it might be a common characteristic of L2 learners' academic writing; or indeed, it might be related to the writing genre, which is support by the following claim.

The analysis of the most frequent bundles in the ESL learners' sub-corpora and the BAWE also found that many LBs that were used frequently by the BAWE writers were never used by the ESL learners. For example, certain bundles, such as *the way in which*, *the idea of* and *it is possible*, were not acknowledged, or were never used by the ESL learners. In contrast, some frequently used bundles in the ESL learners' sub-corpora were also rarely, or never found in the BAWE. For example, the most frequent LBs in the ESL learners' sub-corpora, such as *I think that*, *a lot of* and *first of all it*, were rarely used by the BAWE writers.

An assumption of the increased use of certain LBs may be due to the impact of the L1 on multi-word sequences, as previous studies found (Granger, 1998b; De Cock, 2003; Paquot, 2008; Peromingo, 2012). For example, Peromingo (2012) examined the grammatical collocations in the written production of Spanish university students. The study found that the overuse of some multi-word sequences may be due to the transfer factor, as learners use collocations in their L1 language and transfer it to the L2, especially if using them in the same structure. This was confirmed by Granger (1998b), who found that French learners used collocations that had a direct translation in their mother language. Meanwhile, Paquot (2008) examined the potential influence of the L1 on learners' production of multi-word units, and also found that French learners overused units that had a direct translation in their mother tongue, such as the multi-word unit *let's take the example of* is, which is a direct translation of the French unit *prenons l'exemple de*. Moreover, Nesselhauf (2003); Nesselhauf (2005) also claimed the influence of the L1 on the native-like use of multi-word expressions. Since the present study investigated ESL learners from different L1s, further investigation of their use of LBs according to their L1 language is required as multi-word sequences in the English language are also multi-word sequences in other languages are widely used in academic writing. Further investigation of use of the most frequent LBs in the ESL learners' sub-corpora, and their corresponding sequences in other languages, would be a valuable topic for future research, since to the best of the my knowledge, only a few previous studies addressed the concept of transfer and interference in the use of LBs from the L1 to the L2, and how this affects the increased use of LB production in academic writing (e.g., Paquot, 2013).

Another possible explanation for the increased use of LBs in ESL writing is what Hasselgren (1994, p.237) described as how, in an L2, we "regularly clutch for the words we feel safe with: our 'lexical teddy bears' ". L2 learners tend to rely on high-frequency occurrence, and on words with which they are familiar to express ideas and to avoid grammatical mistakes. However, as Pearson (2021) noted, this principle should be used with caution to avoid the reliance on particular LBs causing a repetition of the bundle used.

It is important to note that learners' increased use of certain bundles can cause repetitiveness of these bundles in the same piece of writing. For example, a manual check of the concordance line for the high frequency LBs in the ESL learners' sub-corpora, such as *I think that*, *a lot of*, *on the other hand*, and *I think it is* revealed that some learners used them repetitively in their essay, which might make their writing tedious and wordy, particularly in short essays. This was a common feature of non-native academic writing identified in previous studies (Cortes, 2004; Hyland, 2008b; Wei and Lei, 2011; Ozturk and Kose, 2016). For example, Cortes (2004) identified the LBs in academic prose published in two disciplines, history and biology, reporting that the biology students tended to repeat specific bundles in a single short paper, making their writing wordy, and including unnecessary words. Biber et al. (1999) claimed that while learners may be familiar with these common bundles, and know how to use them structurally and functionally, it may also engender a form of overuse or "using them when it is not necessary" (Cortes, 2004, p.412).

Another possible explanation for the repetitive use of LBs is that ESL learners tend to use certain LBs as items of high frequency to reflect a high level of formality, and to demonstrate their language competence, or they may still be in the process of learning additional LBs. Furthermore, even if it is assumed that ESL learners can use a range of bundles proficiently in their writing, they can also forget lexical coherence, which makes their writing vague and confusing. Therefore, a high frequency of LBs in ESL learners' academic writing may not be an indication of language competence when they are used repeatedly in their writing.

Finally, the increased use of certain bundles can also be enhanced by instructions that learners are exposed to in their teaching materials, and which tend to focus on the requirement to use specific multi-word sequences. As Wray (2002, p.183) noted, "collocations can only be learned if they are present in the input learners are exposed to". Granger (2004) also observed that the overuse of connectors by non-native writers might be due to, the direct consequence of the long lists of connectors found in most ELT textbooks, which classify connectors in broad semantic categories (contrast, addition, result, etc.) but fail to provide guidelines on their precise semantic, syntactic and stylistic properties, thereby giving learners the erroneous impression that they are interchangeable. To conclude, although ESL learners show an increased use of LBs in their writing, many of these bundles were rarely used by proficient student writers in the BAWE sub-corpus. This means that ESL learners exposed to different bundles than university students.

To conclude, the frequency analysis of the LBs revealed the following:

- The results of the frequency analysis were conclusive regarding whether or not there was a direct relationship between the number of LBs used and proficiency level, since there were significant variations between B2 levels and C1 levels;

- It was evident that, while there is a significant difference in terms of the number of LBs identified across the levels, this discrepancy of the most frequent bundles was less pronounced between B1and B2 levels;
- The most frequently used LBs, *a lot of* and *on the other hand*, were at the top of the B1 and B2 lists; the bundles *I think that* and *I think it is* were the most frequent bundles in the C1 writing;
- According to the information presented in this section, three-word bundles make up more than 85 % of the whole identified bundles.
- LBs were used more often by the ESL learners at all three levels than by the proficient student writers in the RC, many LBs were rarely or never used by the ESL learners, compared with the proficient student writers.

The next section turns the attention to the nature of the target bundles produced to see whether this had an effect on the ESL learner B1, B2, C1 language proficiency.

5.3 Grammatical and pragmatic variations of lexical bundles in the target sub-corpora (B1, B2 and C1)

The second research question asked' What differences exist in the structures and functions of LBs in ESL learners B1, B2 and C1 argumentative essays and proficient student writers?'. The section sought to examine the structures and functions of LBs in ESL learners' writing. This will allow to discover the possible variations of LBs used in ESL learners' writings, as addressed in the following sections.

5.3.1 Structural analysis of lexical bundles

The examination of the type and the token of the LBs across the sub-corpora aimed to clearly show the differences between the levels, as a corpus, especially a small corpus, can only present a limited range of LBs, although with very high frequencies of the same bundles. As discussed in section 4.8, this study developed the framework to cover all the target bundles that did not fall into the classification that divided the bundles into four structural types: noun-based, verb-based, preposition-based and Other. Thus, two bundle sub-categories were added to the verb-based categories, and the six sub-categories that did not fit any of the three main categories were assigned to the category 'Other'. These sub-categories were Wh-clause (*when I was a*), adjective phrase (*available in the*), adverbial phrase (*as well as*), conjunction clause (*and this is*), model

+ verb (*will look at*), and personal pronoun (e.g., *his* or *her*). These sub-categories together accounted for 12.7% in the B1 sub-corpus, 11.4% in the B2 sub-corpus, 10.8% in the C1 sub-corpus, and 7.8% in the BAWE corpus. However, they were excluded from the structural analysis, as some bundles occurred less than five times in the sub-corpora; as a rule of thumb, the chi-squared is invalid if any cell has fewer than five instances (David and Sutton, 2004). A detailed examination of the type of distribution of the target bundles is discussed in the next section.

5.3.1.1 Type distribution

The result of types distribution of the identified bundles given in Table 5.13 and Figure 5.4 show that LBs in the ESL learners' sub-corpora included all three main structural categories in the taxonomy developed. The verb-based bundles, such as *not be able to*, and *are based on*, acted as a leading category in the ESL learners' sub-corpora and the RC, with more than half of the bundles consisting of this category, a greater proportion at B1 level (60%), B2 level (61%), and C1 level (60%) than in the BAWE corpus (52%). However, the overuse may be related to the author's desire to express their thoughts and opinions in certain parts of their essays. For example, a good conclusion in an argumentative essay expresses the writer's personal views to explain how the topic affects them personally (Horkoff and Mclean, 2015), as shown in the example below.

- *I think that* with the help of the contemporary technologies people can do many things that were even difficult to imagine a century ago (C1, essay101)

While these views must be linked to facts and data to support the argument, the overuse of verb-based structures can be interpreted as a sign of a strong argument in academic writing. Interestingly, the B1 and B2 levels writers used almost the same proportion of both noun-based and preposition-based bundles, such as *in the field,* and *in favour of*, whereas the C1 and BAWE writers used more noun-based bundles, such as *according to their,* and *a wide range of,* than preposition-based bundles.

Table 5.13. Type distribution of the target bundles across the structural taxonomy. (χ2 = Chi-square value; Freq = Absolute freuency; % = precentage within-sub-corpus)

χ2 = 23.9; df=6, p < 0.05	**B1**	**B2**	**C1**	**BAWE**
Structural types	**Freq (%)**	**Freq (%)**	**Freq (%)**	**Freq (%)**
Noun-based	86 (20)	79 (19)	142 (24)	244 (28)
Preposition-based	86 (20)	80 (20)	95 (16)	172 (20)
Verb-based	256 (60)	251 (61)	354 (60)	444 (52)

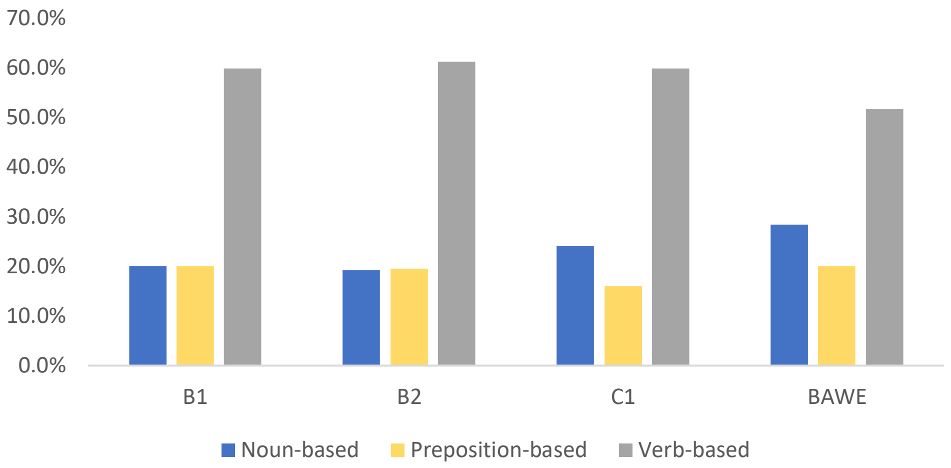


Figure 5.4. Overall distribution of the structural types across the sub-corpora.

In order to provide statistical evidence of the differences in the structural categories between the sub-corpora that were significant, it was necessary to determine the differences by conducting the chi-squared test (and standardised residuals). The test was used to evaluate whether or not the differences between the sub-corpora were random, as shown in Table 5.14.

Table 5.14. Standardised residuals (R) in a chi-square contingency table for the structural distribution. (italic = significant interaction)

χ2 = 23.9; df=6, p < 0.05	**Noun-based**	**Preposition-based**	**Verb-based**
B1	-2.13	0.68	1.29
B2	-2.51	0.33	1.89
C1	-0.02	-2.04	1.64
BAWE	*3.73*	1.02	*-4.03*

If the residual is less than -3, the cell's observed frequency is less than the expected frequency. Greater than 3 and the observed frequency is greater than the expected frequency

It can be seen that the chi-squared test found a significant difference between the sub-corpora, in terms of their use of the structural pattern, with a chi-squared value of 23.9, and a df of 6, a significant *P*-value at 0.05. Further analysis using the standardised residuals (R) that compared between the observed and the expected values of each cell was applied only to the cells with a value greater than ±3.

The results showed a significant increased use of the noun-based and a significant decrease of the verb-based bundles in the BAWE. Meanwhile, no structural types contributed to the significant difference in the ESL learners' sub-corpora. Following from the argument by Biber et al. (2004) that noun phrases and prepositional phrases bundles are dominantly used in academic writing, this difference suggests that even advanced ESL learners may be not closely approximating the academic prose typical of university writing. The extent of these differences will be examined according to their token distribution in the following section.

5.3.1.2 Token distribution

As mentioned in section 2.4.1, frequency occurrence is the primary factor used to identify LBs. Therefore, examining the difference in the use of LBs' structural pattern in the ESL learners' sub-corpora and the RC, according to their frequency, showed the difference between the groups. As explained in section 4.8, the frequent occurrence of the target bundles was normalised to 100,000 words, as the sub-corpora differed in size. The results are summarised in Table 5.15 and Figure 5.5 below.

Table 5.15. Distribution of the target bundles, according to their structural taxonomy. (χ2 = Chi-square value; Freq = Normalised freuency; % = precentage within-sub-corpus)

χ2 = 23.9; df=6, p < 0.05	**B1**	**B2**	**C1**	**BAWE**
Structural types	Freq (%)	Freq (%)	Freq (%)	Freq (%)
Noun-based	918 (20)	908 (20)	2217 (26)	1204 (32)
Preposition-based	1025 (23)	1057 (23)	1521 (18)	839 (22)
Verb-based	2554 (57)	2579 (57)	4956 (56)	1761 (46)

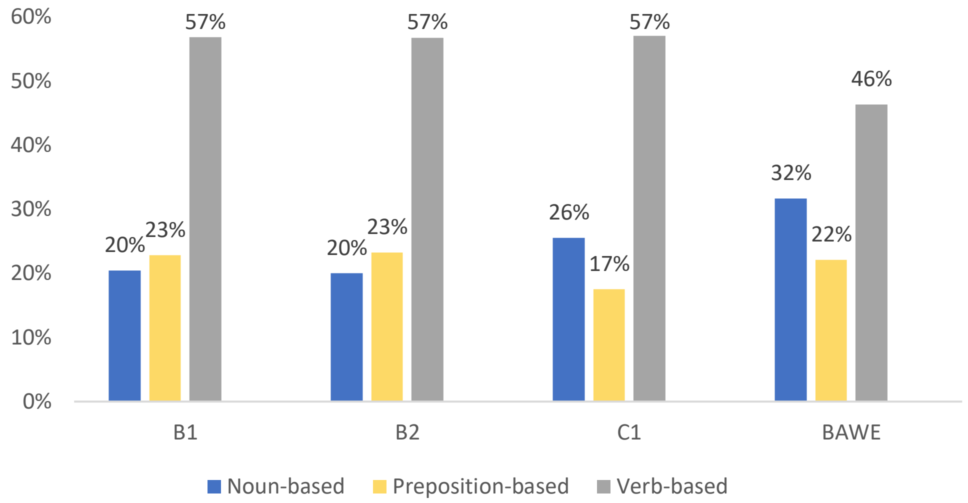


Figure 5.5. Overall distribution of the structural types across the sub-corpora.

The result shows that although the structural distribution of the LBs across the four sub-corpora was partially different from that indicated by the structural type, the difference in the proportion of the structural categories among the groups did not exceed 10%. A comparison of each category's normalised frequency in the ESL learners' sub-corpora showed that the verb-based category remained the most prevalent structure of LBs in both these sub-corpora and the RC, with the frequency of this structure category accounting for almost half of the bundles in each sub-corpus. The above result shows that the B1 and B2 sub-corpora had a similar frequency for each structural category, while there was a slight difference in the C1 and BAWE corpora. Interestingly, the structural type that came second in the B1 and B2 sub-corpora was preposition-based bundles, such as *in the field* and *in favour of*, accounting for around 23% of the total bundle types in both sub-corpora. These bundles featured either an embedded of-phrase to make a logical connection between the elements of an argument, or were without an of-phrase, representing particular research or discourse context (Hyland, 2008a). Meanwhile, the noun-based structure category, (e.g., *according to their* and *a wide range of*) came second in both the C1 and the BAWE sub-corpora. According to (Biber et al., 1999; Marco, 2000), these bundles are used mainly for quantifying, categorising, and quality. This difference was interesting, as the more proficient learners in the C1 and BAWE sub-corpora relied on noun-based bundles in their writing, at 26% and 32%, respectively.

In order to provide statistical evidence of the differences in the structural categories between the sub-corpora that were significant, the chi-squared test (and standardised residuals) conducted on the use of the three main structural categories across the ESL learners B1, B2 and C1' sub-corpora and the BAWE sub-corpus, as presented in Table 5.16.

Table 5.16. Standardised residuals (R) in a chi-squared contingency table for structural distribution. (italic = significant interaction)

Chi-square P < 0.05	**df = 9, P-value 7.58941E, χ2 = 286.73**			
	B1	**B2**	**C1**	**BAWE**
Noun-based				
• **Count**	918	908	2217	1204
• **Expected**	1906	1107	2118	927
• **R**	*-6.93*	*-7.73*	*3.21*	*11.52*
Preposition-Based				
• **Count**	1025	1057	1521	839
• **Expected**	927	937	1793	785
• **R**	*4.06*	*4.94*	*-9.34*	2.41
Verb-based				
• **Count**	2554	2579	4956	1761
• **Expected**	2474	2500	4783	2093
• **R**	2.68	2.65	*4.82*	*-11.90*

If the residual is less than -3, the cell's observed frequency is less than the expected frequency. Greater than 3 and the observed frequency is greater than the expected frequency.

The table above showed that there was a significant difference between the sub-corpora in terms of the three structural categories, with a chi-squared value of 286.73 and a df of 9 that far exceeded the value required for the highest significant *P*-value at 0.0001. Further analysis employed the standardised residuals (R), showed that the noun-based category contributed to the significant difference across the sub-corpora. Specifically, there were significantly more noun-based bundles in the C1 and BAWE sub-corpora, and significantly fewer of the same bundles in the B1 and B2 sub-corpora. In addition, the result confirmed the significant increase of preposition-based bundles in the B1 and B2 writing, and their decreased use in the C1 level writing. Hence, the high-level ESL learners (C1) used LBs differently from the low proficiency levels learners (B1 and B2), and similarly to the university students (BAWE).

The differences between the sub-corpora, supported by the frequency usage of the structural sub-categories when the above broader structural categories were broken down into 13 minor sub-categories, is discussed in the next section.

5.3.1.3 Preliminary analysis of structural sub-categories across the groups

Based on the target bundles, the LBs identified in ESL learners' sub-corpora and the BAWE sub-corpus were classified structurally using the structural taxonomy of Biber et al. (1999). The taxonomy was divided into four categories and 13 sub-categories. A preliminary analysis of the structural categories, with the frequency of each sub-category of the complete framework used in this study, is presented in Table 5.17.

Table 5.17. Distribution of the structural sub-categories across the groups. (Freq = Normalised frequency; % = relative proprtion within-sub-corpus)

Sub-corpus	**B1**	**B2**	**C1**	**BAWE**
Sub-categories	**Freq (%)**	**Freq (%)**	**Freq (%)**	**Freq (%)**
Noun-phrase with other post-modifier fragment	320 (6)	445 (9)	1007 (10)	357 (9)
Noun phrase with of-phrase fragment	598 (12)	463 (9)	1210 (12)	846 (21)
Total Noun-based	**918 (18)**	**908 (18)**	**2217 (23)**	**1204 (29)**
Prepositional-phrase with embedded of-phrase	244 (5)	194 (4)	194 (2)	206 (5)
Other prepositional phrase expressions	781 (15)	862 (17)	1326 (14)	633 (15)
Total preposition-based	**1025 (20)**	**1057 (21)**	**1521 (16)**	**839 (20)**
Be + noun/adjective phrase	163 (3)	219 (4)	309 (3)	192 (5)
Passive verb + prepositional phrase fragment	60 (1)	92 (2)	58 (1)	218 (5)
Anticipatory it + verb/adjective phrase	292 (6)	317 (7)	478 (5)	301 (7)
(Verb phrase) + that-clause fragment	256 (5)	225 (4)	222 (2)	237 (6)
(Verb/adjective) + to-clause fragment*	634 (12)	493 (10)	1352 (14)	469 (11)
Pronoun/noun phrase + be (+…)	306 (6)	602 (12)	329 (3)	214 (5)
1st/2nd person pronoun + VP fragment*	739 (14)	463 (9)	1809 (19)	80 (2)
Other verb phrase	103 (2)	168 (3)	399 (4)	50 (1)
Total verb-based	**2554 (50)**	**2579 (50)**	**4956 (51)**	**1761 (43)**
Other expressions	656 (13)	588 (11)	1053 (11)	320 (8)
Total bundles frequency	**5153**	**5131**	**9746**	**4124**

The table above shows the proportion of bundles (tokens) across the structural sub-categories of the target bundles in the B1, B2, C1, and BAWE sub-corpora, showing that the distribution of LBs across the structural sub-categories differed dramatically. The most salient findings of the normalised frequency across the sub-corpora are as follows. First, the distribution of LBs across the structural sub-categories in B1 and B2 sub-corpora was similar, but was dramatically different from those in C1 and BAWE sub-corpora. For example, the sub-category ‘Other preposition phrase expressions’ was the most frequently used in the B1 sub-corpora, with 15%, and in B2 sub-corpora, with 17% of the LBs tokens. Example included *in a productive, in the form,* and *at the same time*. This sub-category accounted for the second-highest occurrences in the BAWE sub-corpus (15%). In contrast, the sub-category ‘1st/2nd person pronoun + VP fragment’ was the most used bundle type in the C1 sub-corpora, with examples including *I think it is* and *I think that both*, accounting for 19% of the total bundle tokens. Meanwhile, the sub-category ‘noun phrase with of-phrase fragment’, such as *the nature of the,* and *the influence of,* took the lead in the BAWE corpus with 21% of the total bundle tokens.

The next section involves a more qualitative inspection, in which concordance lines of the target bundles are further examined.

5.3.2 Structural categories analysis of the target lexical bundles

This section examines the qualitative differences and similarities between the ESL learners’ and the BAWE sub-corpora from structural categorisation. These differences best are discussed by analysing the most frequent bundles in each structure category, with examples extracted from the corpus to convey their functions, as addressed in the following sections.

5.3.2.1 Comparison of verb-based bundles

Consistent with previous studies (e.g., Chen and Baker, 2016; Pan et al., 2016; Bychkovska and Lee, 2017; Lu and Deng, 2019), the ESL learners in the present study used significantly more verb-based bundles, with nearly twice as many types, and over twice as many tokens, than the other categories. Table 5.18 presents the proportion of the verb-based bundles in the ESL learners’ sub-corpora and the BAWE corpus, the former of which contained more verb-based bundles than the latter.

Table 5.18. Distribution of the verb-based bundles. (Freq = Normalised freuency; % = Relevant ptoportion within-sub-corpus)

Sub-corpus	**B1**	**B2**	**C1**	**BAWE**
Sub-categories/	Freq (%)	Freq (%)	Freq (%)	Freq (%)
Passive verb + prepositional phrase fragment	60 (1)	92 (1)	58 (1)	218 (5)
Anticipatory it + verb/adjective phrase	292 (6)	317 (7)	478 (5)	301 (7)
(Verb phrase) + that-clause fragment	256 (5)	225 (4)	222 (2)	237 (6)
(Verb/adjective) + to-clause fragment	634 (12)	493 (10)	1352 (14)	469 (11)
Pronoun/noun phrase + be (+…)	306 (6)	602 (12)	329 (3)	214 (5)
1st/2nd person pronoun + VP fragment	739 (14)	463 (9)	1809 (18)	80 (2)
Other verb phrase	103 (2)	168 (3)	399 (4)	50 (1.2)
Total verb-based	2554 (50)	2579 (50)	4956 (51)	1761 (43)

As shown in the table above, there were only small variations in terms of the overall usage of verb-based bundles across all the sub-corpora, in total constituting more than 40% of the total bundles identified. However, exanimation of the sub-categories revealed major differences across the sub-corpora. The first sub-category in which the ESL learners and the BAWE writers showed the greatest differences was the '1st/2nd person pronoun + VP fragment' category (e.g., *I think that*, and *I would like to*), indicating that the ESL learners in general employed verb-based bundles beginning with personal pronouns, particularly 1st and 2nd person pronouns (*I, you, we*), in the initial position of verb-phrases in their academic essays.

As Table 5.18 shows, this usage was more prominent in ESL learners (9%-18% of occurrences) than the BAWE writers (2%). Examples in ESL learners' sub-corpora included, *I think it is, we need to, I would like to, I hope I, I believe that, we must use, I could not, I will list,* and *I think this*, which were also found in the BAWE sub-corpus. The following examples illustrate the concordance lines in which some of these bundles appeared in ESL learners' sub-corpora:

- Finally, *I want to* say that cities and villages are good. It is up to you if we want to change yourself or not. We can enjoy both city and village life. (B1, essay141)
- It has become popular in the world today to punish smoking. However, although *I feel that* smoking can be dangerous, I do not think it should be banned completely. (B2, essay,196)

- So, *I think it* is a great experience that makes people stronger, more self-confident. They gain more knowledge and experience that will be very helpful and valuable in the future. (C1, essay, 186)

This confirmed a common pattern found in L2 students' EFL argumentative writing (Hong, 2013; Kim, 2013; Yoon and Choi, 2015a). Indeed, the ESL learners in the present study exhibited an overall higher proportion in the sub-categories that are more representative of spoken language. In the English language, these bundles are those most commonly used to indicate doubt or uncertainty, but have a great subjectivity, and are not commonly used in academic writing.

Another interesting difference across the sub-corpora was the rare use of 'passive verb + prepositional phrase fragment' in the ESL learners' writing (e.g., *can be used to*, *is based on the*). The use of a passive verb followed by a prepositional phrase fragment indicates the impersonalised voice for a locative or logical relationship (Hyland, 2008b). The result was surprising, because the passive voice is usually employed in formal writing, such as academic papers, in which actions themselves are often considered to be more important than the person or object that performs the action. As shown above, less than 2% of this structure was used in ESL learners' sub-corpora, whereas more than 5% in the BAWE sub-corpus. A close examination of the concordance lines of this structure determined that not only was it found increasingly in the BAWE writing, but it was also found with a great variation, as shown in the examples below:

- Therefore, it *can be argued* that Coleridge found it difficult to connect with his audience, and thus used his narrative form to instruct and condemn. (BAWE essay,3008b)

- In many studies the L1 *student is shown* to have a great deal more difficulties in achieving an appropriate linguistic level. (BAWE, essay3118)

The use of passive voice in the BAWE writing demonstrated the general tendency of academic writing to contain more passive voice, which is also associated with stance expression (Biber et al., 1999). Since the BAWE writers were more experienced writers, they were more likely to use more academic expression than the ESL learners. Therefore, the rare use of this structure by the ESL learners was an indication of misuse. Although it was beyond the scope of this study to examine this further, it was noteworthy that this precisely imitated the distinguished characteristics of the ESL learners and the BAWE writers. Therefore, it could be concluded that ESL learners misuse some features of written discourse, which contradicted the conclusion of Wei and Lei (2011), who found that advanced L2 learners used the passive voice far more frequently than professional writers.

Although there were differences between ESL learners and the BAWE writers, they exhibited a similar use of 'anticipatory it + verb/adjective phrase', such as *it is important to, it is often,* and *it is difficult*. This bundles type is commonly used in academic writing (Biber et al., 1999), and typically serves as a linguistic resource that provides impersonalised evaluations (Ädel and Erman, 2012), and conveys a range of epistemic, evaluative, and attitudinal meanings (Jalali et al., 2009). Examples of the use of this sub-category are as follows:

- Secondly, *it is difficult* to think how technology can be used, or misused. (B1, essay35)
- However, *it is important* to know that there are many arguments regarding sending children to schools. (B2, essay108)
- However, I think *it is a* controversial question whether the building of a new university will bring only benefits to our community. (C1, essay43)
- However, *it is interesting* to note that religion alienates certain characters within the novel. Much like Hardy's Tess it offers Hetty no comfort in her time of need. (BAWE, essay3001)

The similar frequency of these bundles across the sub-corpora not only indicated their importance as basic concepts, but also their necessity for presenting impersonalised assessments, as most use of 'anticipatory it' also reflects the speaker or writer's evaluation (Hewings and Hewings, 2002).

Theoretically, there were two relationship patterns involving the ESL learners' language data, in terms of the distribution use of LBs. Firstly, they were shared by the ESL learners' data and the BAWE data, hence the former included aspects of the characteristics of academic language, because LBs of this kind shape the characteristics of this form of language. The second pattern was the fact that the ESL learners' data shared a form of LBs that probably did not profile the written register, as it was found only rarely in the BAWE data.

5.3.2.2 Comparison of preposition-based bundles

Turning to the comparison of preposition-based bundles across the groups. The results in Table 5.19 shows that the most frequently used sub-category was 'other prepositional phrase expressions', namely bundles introduced by a preposition.

Table 5.19. Distribution of the preposition-based bundles. (Freq = Normalised freuency; % = Relevant ptoportion within-sub-corpus)

Sub-corpus	**B1**	**B2**	**C1**	**BAWE**
Sub-categories	**Freq (%)**	**Freq (%)**	**Freq (%)**	**Freq (%)**
Prepositional phrase with embedded of-phrase	244 (5)	194 (4)	194 (2)	206 (5)
Other prepositional phrase expressions	781 (15)	862 (17)	1326 (14)	633 (15)
Total preposition-based	1025 (20)	1057 (21)	1521 (16)	839 (20)

Examples of such bundles in the ESL learners' sub-corpora included *on the other hand*, *in this essay, as a result,* and *in order to*, for instance:

- *On the other hand*, there are many disadvantages such as streets are well covered, light system is available. (B1, essay467)
- *On the other hand*, there are also several reasons why public health might not improve.
- *As a result of* this, the amount of traffic congestions will increase, as well as contamination of the air. So, all these obviously will not make one's life happier and healthier in my community. (C1, essay67)
- Here, he insisted that a healthy material base and equality of opportunity for all was necessary *in order to* form a liberal democratic society. (BAWE, essay3005)

Although LBs are largely incomplete structural units, the above examples show that they act as connector expressions when used as a complete unit. The 'other prepositional phrase expressions' sub-category is one of the two, together with 'noun phrase with of-phrase fragment' that can integrate these complete structural units. For instance, *on the other hand, as a result of,* and *in this essay.*

When comparing between the sub-corpora, six prepositional phrase expressions without-*of* bundles were found to be shared: *in addition to, in the future, in the world, as they are, on the other hand, in this essay I*, and *at the same time*. Of these bundles, only *on the other hand* was used far more frequently by the ESL learners than the

BAWE writers, suggesting that the ESL learners at all levels used this bundle more than professional writers. A close examination of the concordance lines of *on the other hand* also revealed that the ESL learners primarily used it in the same way as the BAWE writers, to contrast two ideas, concept, or activities as in the below examples:

- *On the other hand*, the life in village is easier. People have good and pure food, and They have good relationships with each other. (B1, essay 483)
- *On the other hand*, it can bring freedom to deny the speaker's intent easily because it seems to be very indirect and speaker's intension is not clear. (BAWE, 3125h)

The increased use of this expression in each group reflected its prevalence in academic writing as a useful phrase with which to organise a text.

5.3.2.3 Comparison of noun-based bundles

As reported by previous research, 70% of the most common bundles are usually part of a 'noun phrase with an of-phrase fragment' (Biber et al., 1999; Hyland, 2008b; Chen and Baker, 2010; Johnston, 2017). The bundles in this sub-category are mainly used to describe events or processes (e.g., *the way of*), to identify abstract qualities (e.g., *the nature of*), and to describe (e.g., *the end of*) place, size, and amount. The assessment of the noun phrase-based bundles in all the sub-corpora revealed that the 'NP with of-phrase fragment' pattern constituted the majority of this structural type, with a higher usage frequency in the BAWE and the C1 writers, as shown in Table 5.20.

Table 5.20. Distribution of Noun-based bundles. (Freq = Normalised freuency; % = Relevant ptoportion within-sub-corpus)

Sub-corpus	**B1**	**B2**	**C1**	**BAWE**
Sub-categories	**Freq (%)**	**Freq (%)**	**Freq (%)**	**Freq (%)**
Noun phrase with other post-modifier fragment	320 (6.2)	445 (8.7)	1007 (10.3)	357 (8.7)
Noun phrase with of-phrase fragment	598 (11.6)	463 (9)	1210 (12.4)	846 (20.5)
Total Noun-based	918 (17.8)	908 (17.7)	2217 (22.7)	1204 (29.2)

The occurrence of these 'noun phrase with of-phrase' fragment bundles in the ESL sub-corpora suggested that the ESL learners' writing also contained phrases that are often associated with highly formal academic writing style, thereby illustrating the nature of the argumentative essay. According to Biber et al. (1999), the NP expressions *with* and *of* are the largest proportional type in academic prose.

Further examination of the use of the sub-category 'noun phrase with other post-modifier fragment' in each sub-corpus identified significant differences between the groups. As might be expected, given its high raw frequency, the bundle *the use of* took the lead in this category in the BAWE lists, whereas *a lot of* was the most frequent bundle in the ESL learners' sub-corpora, with a higher presence in the B1 and B2 writing. This bundle is used to specify quantity, usually followed by a noun phrase, though no particular words or phrase that frequently followed it in the ESL learners' sub-corpora. Although this bundle can be found at the beginning of a sentence or clause, or in the middle, the concordance lines showed that the B1 writers used it only in the middle, whereas the other writers used it in various position, as shown in the examples below. Therefore, it appeared that the higher proficiency ESL learners were more confident in using such LBs.

- But now a days there are *a lot of* common ways of communication we can get information about anything in seconds. (B1, essay373)
- *A lot* of gyms, parks, and sports centres can be a good thing for people health. (B2 sub-corpus, essay 178)
- Unemployment is also causing *a lot of* mental issues within the people especially youth. (B2, essay355)
- Unemployment is also caused by companies who require *a lot of* experience for a job (C1, essay386)
- *A lot of* primitive jobs are done by machines nowadays. (C1, essay313)
- The reader is given *a lot of* description regarding the personalities of the main characters. (BAWE, essay31)
- *A lot of* useful information regarding this subject is extracted from spontaneous speech.

The salient overuse of the quantifier bundle *a lot of* in the ESL learners' writing may be related to the tendency for overstatement in L2 writing (Chen, 2009), which is not found in BAWE writing. Hinkel (2005) observed that many L2 learners' writing is overstated, due to the excessive use of many quantifiers when presenting their argument. It may be the case that this stylistic overuse is a characteristic of ESL learners' writing. It would be better to raise the learners' awareness that the meaning of such phrases is nonspecific and subjective, but that there often are generally accepted ideas for what forms "a lot". This knowledge could better help learners to use these bundles effectively.

The second high frequency bundle in this subcategory across the ESL learners' writing was *first of all*. This bundle is a type of listing expressions generally used when there is a list or subsequent events that help organise written text (see section 5.4.1). Listing expressions lend integral structure to a text to increase its readability (Geva, 1992; Heino, 2010). This bundle was used increasingly by the ESL learners, and ranked in the top 20 most frequent bundles in all the sub-corpora. In contrast, the BAWE writers preferred to use a single word, *firstly*, as a linking adverb to organise their text, as in the examples below:

- *First of all*, childcare centres help children in their development. (B1, essay28)
- *First of all*, I believe we should waste things as little as possible, for example people should use electrical appliances when it is really necessary. (B2, essay273)
- *First of all*, television helps a child to extent his or her range of interests. Children can find out many new things and make many exiting discoveries for themselves. (C1, essay62)
- *Firstly*, one of the most noticeable aspects of both texts is that they are mainly in past tense. (BAWE, essay3066F)

The examples above from both the EFL learners and the BAWE groups show the use of the listing expressions to organise ideas in a text, but using different expressions, all of which can be employed interchangeably when there is a list or subsequent events. However, while there is a considerable number of listing expressions in BAWE writing, the number is much more frequent in the ESL learners B1, B2 and C1 sub-corpora, and thus is over-used by ESL learners. As Gilquin and Paquot (2007) argue that *first of all* is more representative of speech than of academic writing, and their overuse in written argumentative essays by ESL learners B1, B2 and C1 may thus be characterised as somewhat problematic. Further examination of the sequential expressions is discussed in the keyness analysis.

5.3.3 Functional analysis of lexical bundles across the sub-corpora

As discussed in section 2.8, the functional analysis of LBs was the subject of previous studies (e.g. Biber et al., 1999; Hyland, 2008a; Salazar and Joy, 2011; Jalali et al., 2014; Durrant, 2015), in which, with the exception of the unusual frequency of LBs in both L1 and L2 English, these bundles were found to serve a variety of discourse functions. This section discusses the distribution of the functional categories and sub-categories of the target bundles in the B1, B2, C1, and RC sub-corpora.

As discussed in section 3.7, the first step of the analysis examined the concordance lines, in order to allocate the target bundles to a corresponding functional category adopted from that proposed by Hyland (2008b), according to its discourse function in the text. Contrary to the approach employed for the structural analysis, all the target bundles were categorised under Hyland's taxonomy. A detailed examination of the type of distribution of the target bundles is discussed in the next section.

5.3.3.1 Type distribution

This section presents an overview of the type distribution of target bundles according to the functional taxonomy applied in this study, as displayed below.

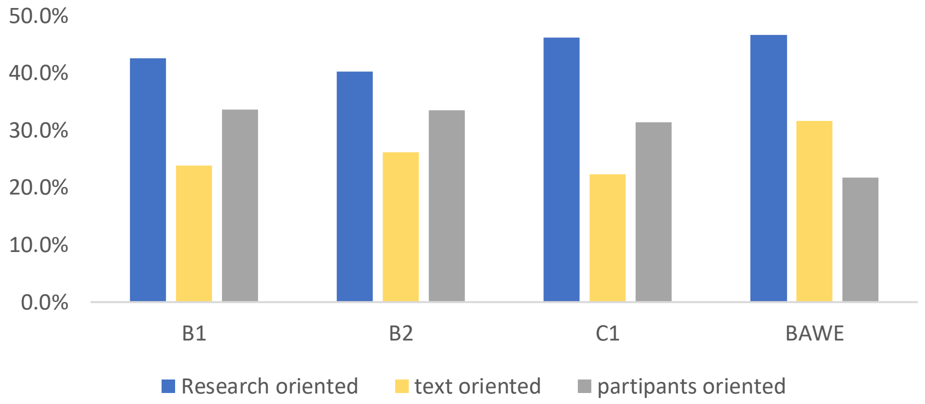


Figure 5.6. Overall type distribution of the target bundles, according to their functional taxonomy across the sub-corpora.

Table 5.21. Overall type distribution of the functional types across the sub-corpora. (Freq = Normalised freuency; % = Relevant ptoportion within-sub-corpus)

	B1	**B2**	**C1**	**BAWE**
Functional types	Freq (%)	Freq (%)	Freq (%)	Freq (%)
Research-oriented	214 (42)	191 (40)	319 (47)	445 (47)
Text-oriented	120 (24)	124 (26)	154 (22)	302 (32)
Participant-oriented	169 (34)	159 (34)	217 (31)	207 (21)

As can be seen, research-oriented bundles (e.g., *a lot of*, and *the end of the*) were the most frequent category, accounting for 42% (B1), 40% (B2), 47% (C1), and 47% (BAWE). This concurred with the main features of academic writing, which focuses on the subject of the research. The finding agreed with that of other studies, in which research-oriented bundles, which corresponded with the referential bundles in Biber et al. (1999) taxonomy, were proven to be the most predominant functional category in academic writing (e.g.,Chen and Baker, 2010; Salazar, 2012; Güngör and Uysal, 2016).

Meanwhile, the participant-oriented bundles (e.g., *it is necessary, it is probably,* and *I believe that*) came second in the ESL learners' sub-corpora, representing almost a third of the total bundles in each sub-corpus. These bundles were used to convey the writer's attitudes and evaluations. This reflected the findings of the studies by (Staples et al., 2013; Cooper, 2013), in which the learners used more participant-oriented than text-oriented bundles.

The same procedures employed for the structural analysis were used to conduct a chi-squared statistical analysis to examine the difference between the sub-corpora, in terms of the usage frequency of the functional categories. The standardised residuals were then calculated to locate the cells that contributed to the significance, as displayed in Table 5.22.

Table 5.22. Standardised residuals (R) in a chi-squared contingency table for functional distribution. (italic = significant interaction)

Chi-square P < 0.05	**df = 4, P-value 4.14979E-08, χ2 = 98.67**			
	B1	**B2**	**C1**	**BAWE**
Research-oriented				
• **Count**	214	191	319	445
• **Expected**	224	211	308	425
• **R**	-1.03	-2.08	1.00	1.59
Text-oriented				
• **Count**	120	124	154	302
• **Expected**	134	127	184	255
• **R**	-1.60	-0.29	-3.03	*4.33*
Participant-oriented				
• **Count**	169	159	217	207
• **Expected**	144	136	198	274
• **R**	2.70	2.58	1.86	*-5.98*

If the residual is less than -3.07, the cell's observed frequency is less than the expected frequency. Greater than 3.07 and the observed frequency is greater than the expected frequency.

The result showed a significant difference in the functional distribution of the target bundle types between the four sub-corpora, with a chi-squared value of 45.26 and df of 6, far beyond the value required for the highest significant *P*-value at 0.0001. However, the results of the R-value calculation were the same as those for the structural categories, in terms of type distribution, as the BAWE writers were the only ones to exhibit a significant result in the use of text-oriented and participant-oriented bundles, indicating that they used more text-oriented bundles, and fewer participant-oriented bundles than expected in their writing. No further investigation of the type distribution was conducted, as no significant differences emerged between the sub-corpora. The extent of the difference between the sub-corpora was examined further, according to the token distribution.

5.3.3.2 Normalised token distribution

A comparison of each functional category's normalised frequency among the sub-corpora demonstrated the differences between the groups, as presented in Table 5.23 and Figure 5.7. The results of LBs in the ESL learners' sub-corpora (B1, B2, and C1) demonstrated the same results presented in the type distribution of the functional sub-categories. It was noteworthy that each category included an increased usage frequency in the C1 level sub-corpora than B1, B2, and BAWE corpora. Among the increases observed, the most considerable growth was in the normalised frequency of the participant-oriented bundles, while the smallest was in the text-oriented bundles.

Table 5.23. Overall token distribution of the functional tokens across the sub-corpora. (Freq = Normalised freuency; % = Relevant ptoportion within-sub-corpus)

	B1	B2	C1	BAWE
Functional types	Freq (%)	Freq (%)	Freq (%)	Freq (%)
Research-oriented	2214(43)	2007 (39)	3534 (36)	1938 (47)
Text-oriented	1284 (25)	1486 (29)	2727 (28)	1345 (33)
Participant-oriented	1655 (32)	1938 (32)	3485 (36)	841 (20)

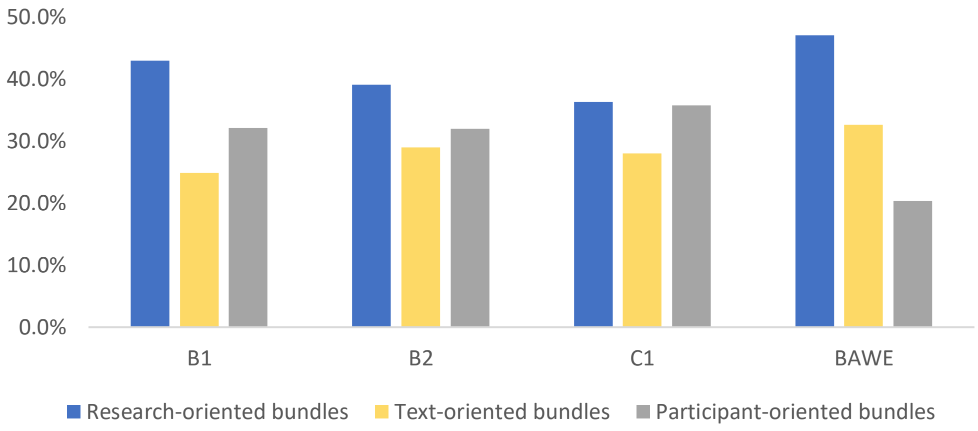


Figure 5.7. Overall distribution of the functional types across the sub-corpora.

The preliminary analysis of the functional categories, with the detailed frequency of each sub-category of the complete framework used in this study shows the distribution of LBs across the functional sub-categories differed slightly across the sub-corpora (Table 5.24). The most salient findings of the normalised frequency assessment, across the sub-corpora are as follows: first, similarly to the structural analysis, the proportion of the functional sub-categories in the B1 and B2 sub-corpora were more-or-less identical, and differed from those in the C1 sub-corpus. For example, the proportion of four out of 10 functional sub-categories was nearly the same in B1 and B2 sub-corpora. For the research-oriented bundles, the 'quantification' sub-category, which indicated measures, quantities, proportions, and change bundles, such as *was a little,* and *in order to reduce*, came first in the B1 (16%) and the B2

(16%) sub-corpora, while the ‘procedure’ sub-category, which indicates events, actions, and methods, such as *to arrange their,* and *to make a good*, were of the highest proportion in the C1 sub-corpus (15%). The difference between the sub-corpora was also apparent in the text-oriented functional category, as ‘transition’ bundles, such as *on the other hand,* and *in addition to*, topped the list of the B1 and B2 writing, while ‘structuring’ bundles, such as *with respect to the,* and *in this essay*, were first in the C1 writing.

Table 5.24. Normalised frequencies and relative proportions of the functional sub-categories, across the sub-corpora. (Freq = Normalised freuency; % = Relevant ptoportion within-sub-corpus)

Sub-corpora	**B1**	**B2**	**C1**	**BAWE**
Sub-categories	**Freq (%)**	**Freq (%)**	**Freq (%)**	**Freq (%)**
Research-oriented bundles				
• Location	443 (9)	371 (7)	270 (3)	258 (6)
• Procedure	608 (12)	475 (9)	1544 (16)	785 (19)
• Quantification	799 (16)	800 (16)	1190 (12)	410 (10)
• Description	364 (7)	361 (7)	529 (5)	485 (12)
• Total	**2214 (44)**	**2007 (39)**	**3534 (36)**	**1938 (47)**
Text-oriented bundles				
• Transition signals	501 (9)	403 (8)	760 (8)	401 (10)
• Resultative signals	368 (7)	425 (8)	601 (6)	283 (7)
• Structuring	244 (4)	367 (7)	992 (10)	120 (3)
• Framing	171 (3)	291 (6)	373 (4)	541 (13)
• Total	**1284 (23)**	**1486 (29)**	**2727 (28)**	**1345 (33)**
Participant-oriented bundles				
• Stance	1532 (30)	1490 (29)	3289 (34)	756 (18)
• Engagement	123 (3)	148 (3)	196 (2)	85 (2)
• Total	**1655 (33)**	**1638 (32)**	**3485 (36)**	**841 (20)**
• Total bundles	5153 (100)	5131 (100)	9746 (100)	4124 (100)

In order to assess the significance of these findings, a chi-squared statistical analysis (and standardised residuals) was conducted for the three main categories, to determine the difference between the B1, B2, and C1 levels, compared with the RC (Table 5.25). It shows that there were significant differences among the four sub-corpora, with a chi-squared value of 363.66 and the df 6, and a P-value at 0.0001. The highlighted cells of the R-value that exceeded ±3.07 indicated the significant difference among the functional categories. These results differed from those of the functional type, reflecting a significant increase in the use of research-oriented bundles by the B1 and BAWE writers, with the C1 writers using significantly fewer of this bundle type than the other students. Another striking result was for the participant-oriented bundles, which indicated that the C1 writers used significantly more of these bundle types in their writing. Accordingly, a closer look at each functional sub-categories is needed to find the distribution of each functional sub-categories across the levels.

Table 5.25. Chi-squared and standardised residuals (R) for the functional distribution of the comparison between the B1, B2 and C1 learner levels. (italic = significant interaction)

Chi-square P < 0.05	**df = 4, P-value 1.63867E-75, $\chi 2$ = 363.66**			
	B1	**B2**	**C1**	**BAWE**
Research-oriented				
• **Count**	2214	2007	3534	1938
• **Expected**	2068	2059	3911	1655
• **R**	*4.67*	-1.66	*-10.0*	*9.87*
Text-oriented				
• **Count**	1284	1486	2727	1345
• **Expected**	1460	1453	2761	1168
• **R**	*-6.12*	1.133	-0.980	*6.71*
Participant-oriented				
• **Count**	1655	1638	3485	841
• **Expected**	1626	1619	3075	1301
• **R**	1.00	0.65	*11.5*	*-16.9*

If the residual is less than -3.078, the cell's observed frequency is less than the expected frequency. Greater than 3.078 and the observed frequency is greater than the expected frequency.

5.3.4 lexical bundles in functional categories

The second part of the functional distributions of LBs involves a more qualitative inspection, in which the concordance lines of the target bundles are further examined. In so doing, the qualitative analysis begins with a discussion of the research-oriented bundles in sub-section 5.3.4.1, followed by text-oriented bundles sub-section 5.3.4.2 and participant-oriented bundles sub-section 5.3.4.3. This section discusses the significant similarities and differences of the functional sub-categories in detail, exploring how the LBs and their discourse functions were used across the CEFR levels, and demonstrating the different choices made by the ESL learners to organise their essays.

5.3.4.1 Comparison of research-oriented Bundles in the sub-corpora

Research-oriented bundles, which correspond to the ‘referential expressions’ in the classification of Biber et al. (1999), were by far the most commonly used in the ESL learners’ sub-corpora, and the BAWE sub-corpus. These bundles help writers to structure their experience of the real world of the text, indicating time, location, and procedure, in order to quantify facts, and experiences (Hyland, 2008a). The results of the research-oriented bundles are displayed in Table 5.26.

Table 5.26. Comparison of the research-orientatated bundles across the sub-corpora. (Freq = Normalised freuency; % = Relevant ptoportion within-sub-corpus)

	B1	**B2**	**C1**	**BAWE**
Sub-functional types	**Freq (%)**	**Freq (%)**	**Freq (%)**	**Freq (%)**
Location	443 (9)	371 (7)	270 (3)	258 (6)
Procedure	608 (12)	475 (9)	1544 (16)	785 (19)
Quantification	799 (16)	800 (16)	1190 (12)	410 (10)
Description	364 (7)	361 (7)	529 (5)	485 (12)

As can be seen, the majority of the LBs identified in ESL sub-corpora fell under the quantification and procedure sub-categories. The increased use of procedure bundles across the sub-corpora, and particularly by the more advanced learners, may be due to the fact that the students sought to demonstrate their mastery of the scope and content of their essays, as shown in the example below.

- to sum up, *the purpose of* all these actions is to make animals live happier, therefore, maintain the ecological balance, hence protect our own living environment. (C1, essay156)

This bundle type was also the most frequent in the BAWE sub-corpus, accounting for 19% of the total bundles. The use of procedure bundles not only demonstrates language competence, but also organises the discourse to facilitate a better understanding of the text. In the ESL learners’ sub-corpora, such functions were realised mainly by bundles that followed the structural pattern ‘verb+ to-clause fragment’, such as in the following examples.

- That process helps transport experts *to determine the* appropriate materials for forming the surface of streets in order to increase the safety of roads. (C1,essya159)

- All that our governments can do is *try to make sure* that it is in the interests of our society and our environment. (B2, essay109)

Several other structures were also used to serve the same function, including the use of prepositional phrases and noun phrases ('with or without of-phrase fragments'), 'passive + prepositional phrase fragments', and other expressions. Meanwhile, in order to express procedure, the most commonly used bundle by far in the BAWE writing was *the use of*, whereas the bundle *in order to* was the most frequent procedure bundle in ESL learners B1, B2 and C1 writing. Examples include the following:

- *In order to* reduce Global warming, several methods are being introduced. First, we have to use public traffic, or walking, or riding a bicycle. (B1, essay 155)
- *In order to* reduce pollution, the vehicles which consume and admit petroleum products should have regular vehicle oil check. (B2, essay260)
- From my experience and observation *I think that* all people who succeeded in life had to work hard and gain more knowledge and experience in order to reach their goals. (C1, essay 328)
- *The use of* landscape was an ideal vehicle through which to explore the recent implications of Darwinism upon community. (BAWE, essay3001)

By using these bundles, the writers sought to explain the steps of the process employed to address the essay topic. However, this contrast was a key difference in how the writers discussed and constructed their ideas.

Meanwhile, quantification also appeared frequently in both the ESL learners' and the BAWE writers' sub-corpora, with an overuse by the low-level learners. This sub-category comprised about a third of the total research-oriented tokens in the ESL learners' sub-corpora. When they investigated the use of LBs across CEFR levels, Chen and Baker (2016) also found increased use of quantifying bundles across the levels involved, particularly by low-proficiency learners. The present study added to these findings by showing that this feature was also common in high proficiency ESL learners, and seemed to be a common characteristic in L2 learners' writing. Expressions in this sub-category comprised mainly of 'Noun-based' and "Preposition-based" bundles, with a very small number of other structure categories, such as ' Verb

+ to fragments' (for example, *to all the*), 'anticipatory it + verb/adjective phrase' (for example, *it is more*), 'be + noun/adjective phrase' (for example, *is one of the most*), and 'pronoun/noun phrase + be' (for example, *there are some*).

The most frequent quantification bundle in the ESL learners' sub-corpora was *a lot of*. As discussed in section 5.3.2.3, the salient overuse of the quantifier bundle *a lot of* in the ESL learners' writing may be related to the tendency for overstatement in L2 writing; thus, its use does not reflect general academic-register bundles. Meanwhile, the bundle *one of the* was the most frequent quantifying bundle in the BAWE sub-corpus.

The bundles reviewed above represented the linguistic choices made by the ESL learners and the BAWE writers to present content and "real world activities and experiences" (Hyland, 2008b, p.13). Nevertheless, as Hyland and Tse (2007, p.167) explained, in academic writing the presentation of content is also important, and is linked to the choice of the linguistic resources that serve the function of organising ideas, findings, and experiences into "convincing and coherent texts". In the ESL sub-corpora, a similar function was realised by the text-oriented bundles, as discussed in the next section.

5.3.4.2 Comparison of text-oriented Bundles in the sub-corpora

The text-oriented subcategory corresponded to the 'discourse-organiser' category in the classification of Biber et al. (1999), and was the least frequently-used bundle in ESL learners' sub-corpora, serving the function of organising the discourse. Text-oriented LBs were defined by Hyland (2008b, p.13) as those "concerned with the organisation of the text and its meaning as a message or argument". Bundles in this category are therefore used to establish additive or contrastive links between elements (e.g., *on the other hand, in addition to*), to mark inferential or causative relationships between elements (*in contrast, as a result of*), to structure signals in the text (*in this essay, in the following*), and to frame signals (*the case of, on the basis of*). Despite their low percentage across the sub-corpora, their role should not be ignored. Table 5.27 presents the usage proportion of text-oriented bundles across the sub-corpora, which consisted mainly of noun-based and preposition-based categories.

Table 5.27. Distribution of the text-oriented bundles, across the sub-corpora. (Freq = Frequency; % = percentage within-sub-corpus).

	B1	**B2**	**C1**	**BAWE**
Sub-functional types	**Freq (%)**	**Freq (%)**	**Freq (%)**	**Freq (%)**
Transition signals	501 (10)	403 (8)	760 (8)	401 (10)
Resultative signals	368 (7)	425 (8)	601 (6)	283 (7)
Structuring	244 (5)	367 (7)	992 (10)	120 (3)
Framing	171 (3)	291 (6)	373 (4)	541 (13)

Among the text-oriented sub-categories, structuring expressions were used significantly more in the ESL learners' sub-corpora than in the BAWE corpus, primarily using two structural patterns: 'prepositional phrase' and 'noun phrase + of fragments'. As Hyland (2008b) explained, the use of these expressions helps to organise the text by providing a framework within which new arguments can be discussed and refer to text stages. The expressions *first of all* and *in the essay* were the most frequent bundles used as structuring devices, and were employed to form part of expressions to direct the reader's attention to a list several things to be discussed (1), and the whole work, or part of it to summarise what is to be addressed (2), as shown in the following examples:

- *First of all*, I would like to tell advantage of playing video games. (B1, essay234)
- *In this essay*, I will discuss some of the main reasons for this and offer some suggestion about how to tackles this problem. (B2, essay8)
- *First of all*, internet and e-mail have changed the way people communicate to each other. (C1, essay203)

In contrast, the BAWE writers made considerably more use of framing bundles than the ESL learners. Such bundles are used to identify particular properties or conditions, and are more typically prepositional phrases with or without embedded of-phrase fragments. The increased use of framing expressions suggested that the BAWE writers employed such bundles that connected research ideas and provided a specific explanation that made their writing more reader-friendly, as in the following examples.

- The play explores *the ways in which* different people make sense of events in their lives, including what they merely imagine to be happening. (BAWE, essay3005b)

- *In the case of* fast mapping, learning occurs as the adult interacts with the child and there is usually an element of contrast where the new word is contrasted with a familiar word. (BAWE, essay6067e)

The difference between the ESL learners and the BAWE writers may be due to the rhetorical style of the writing genre in question, and the ESL learners appeared more often to feel the need to situate their arguments clearly to persuade a particular audience, and also felt a need to focus on other ideas relevant to their position. Anticipating ideas may also involve subsequent thoughts or possible concerns that arise in the reader's mind. Conversely, the BAWE writing more often discussed the connections between the text components, and directed the reader's attention towards specific details.

The results concurred with those of (Cooper, 2013; Pan et al., 2016; Esfandiari and Barbary, 2017), who showed that L2 learners employ fewer framing bundles in their writing than their L1 counterparts. This may be reflected in the infrequent use of preposition-based bundles in the ESL learners' sub-corpora, since framing sub-category comprises mainly of preposition-based bundles.

Despite the variations in the use of text-oriented bundles, both the ESL learners and the BAWE writers exhibited the same proportional use of resultative and transitional bundles that help to writers to forge additive links, to compare and contrast, and to establish signal conclusions of the ideas in their essays, as in the following examples:

- *To sum up,* helping reduce the effects of global warming is not only good for yourself but everyone and our home planet Earth. (B2, essay244)
- *On the other hand*, physical games help adults to stay fit, relax, eliminate stress and tension. Also, it is a perfect way to lose weight. (C1, essay113)

There were more transition bundles than other text-oriented subcategories across the sub-corpora, with the vast majority of this sub-category comprised prepositional phrases, and used to either add new information (for example, *in addition to*), to support an argument (for example, *at the same time*), or to contrast two ideas or arguments (for example, *on the other hand*).

The frequent use of connectors, within a short essay is a characteristic of non-native writers' argumentative essays, in which they are required to show their ability to build logical connections within a specific time and length of text (Paquot, 2010; Leedham and Cai, 2013). The results suggested that the ESL learners gave the impression of logicality using connectors more than the BAWE writers, as was evident in the increased frequency of transition sub-categories. Meanwhile, resultative signals also have an important rhetorical function in creating links between the components of an idea or claim, as in the following examples:

- expressing cause and effect/result;

- *As a result* of deforestation, the number of planets decrease, and the farm production reduced.(B1, essay3)
- *The effect of* global warming is also that it is causing various diseases in humans as we know that the sun's ray is causing skin cancer and other skin related diseases in humans. (B2, essay380)

- making comparisons; and

- *As a consequence,* he incorporated his personal experience of the city into his work. On the one hand, Schwarzbach stresses the fictionalization of some events of the novelist's life. He argues that Dickens's "fiction makes free and creative use of every detail of his outer and inner life". (BAWE, essays3012)
- Maybe *the reason for* this is that most of the modules are lead my native speakers. However, I honestly do not know if that is the only reason. (BAWE, essay3150)

Since most of the discourse and cohesive markers correlate with the text-oriented function of bundles, due to the importance of these expressions for learners in making their writing as coherent as possible, it can be concluded that the ESL learners in the present study primarily used text-oriented bundles to make logical connections between their ideas, and to clarify their arguments, which facilitated understand of their ideas, and forged logical correlations between paragraphs.

5.3.4.3 Comparison of participants-oriented Bundles in the sub-corpora

Although academic writing differs considerably from spoken language, both involve an interaction, in the case of the former between the writer and the reader. Participant-oriented bundles were second most frequently used category across the ESL learners' sub-corpora. The use of these expressions helps the writer to focus on the reader, to make evaluations of their arguments, and to express their view and position vis-à-vis the arguments and ideas presented, which provides the "key aspects of interaction in texts" Hyland (2008b, p.18). According to Hyland (2008b, p.147), academic writing not only focuses on the presentation of real-world activities and experiences, but also includes "using language to acknowledge, construct and negotiate social relations". Therefore, there is a need to present the arguments and interpretations concerned in convincing ways by taking advantage of language resources in their discourse communities to "express their positions, represent themselves, and engage their audiences" (ibid., p.176). The participant-oriented bundles were classified according to their functions: (1) stance expressions that refer to "ways writers explicitly intrude into the discourse to convey epistemic and affective judgments, evaluations and degrees of commitment to what they say"; and (2) engagement, which indicates "the ways writers intervene to actively address readers as participants in the unfolding discourse" (ibid., p.18).

The results presented showed that the ESL learners and the BAWE writers exhibited the same behaviour in terms of this category, choosing to employ more bundles that conveyed the writer's stance and attitudes than those that served to engage the reader, as shown in the Table 5.28.

Table 5.28. Distribution of the participant-oriented bundles, across all the corpora. (Freq = Frequency; % = percentage within-sub-corpus).

	B1	**B2**	**C1**	**BAWE**
Sub-functional types	**Freq (%)**	**Freq (%)**	**Freq (%)**	**Freq (%)**
Stance	1532 (30)	1490 (29)	3289 (33)	756 (18)
Engagement	123 (2)	148 (3)	196 (2)	85 (2)

The most frequent bundle in all the corpora was the participant-oriented, *I think that*, a finding that contradicted academic prose register norms. This bundle is typical of spoken academic discourse, but is less frequent in written genres; it is used to express an opinion, generally regarding something discussed previously (Sykes, 2017). In many cases, the word that follows this bundle refers to a noun or noun place (e.g., *I think that the government,* or *I think that children*). It is more likely that argumentative essays requiring the writer to express their opinion will include a high proportion of stance expressions, such as *I believe that*, *I think it is,* and *it is important*, an assumption confirmed by the increased use of stance bundles in the ESL learners' sub-corpora. The stance sub-category "carries meanings such as certainty, uncertainty, possibility, probability and importance that are effective means for writers to communicate their own assessments of certain propositions and their degree of confidence in these claims" (Salazar, 2014, p.104), as shown in the following examples:

- Now idea of industries come to mind that in cities there are more jobs than in villages. But *i want to* say that this not compulsory that only jobs in cities are available. (B1, essay74)
- In conclusion, *I believe that* homelessness will continue in large cities unless governments, schools and parents cooperated and take steps to address this situation in order to reduce the number of homeless people. (B2, essay123)
- Personally, *I think it is* great to help each other, share new ideas, develop new solutions, etc. It helps to create a team spirit and improve a labor productivity. (C1, essay79)

A common feature that appeared when examining the concordance lines in the ESL learners' sub-corpora was that they tended to join stance expressions with the first person pronouns *I* and *we* in bundles such as *I think it is*, *we want to*, *we need to*, and *I think that*, which were rarely found in the BAWE writing. For example:

- I agree that video games are interesting and quite fun, but *I do not* think it is a good way to keep you fit. (essay, 232)
- Finally, the main step we need to take is to live more simply. We need to reduce our consumption, recycle, and reuse. (B2, essays137)
- *I think that* famous athletes and entertainers have a great impact on our social life and make a big contribution to our society and, hence, deserve high salary. (C1 essay110)

The rare use of the personalised structure in the BAWE writing showed that the ESL learners, more than the professional writers, made significant use of the authorial presence in their essays. Moreover, most of the stance expressions found in the BAWE writing were in the form of anticipatory it, such as *it seems that, it is possible, it is beneficial, it is important to*, which indicate an impersonal tone. For instance:

- *it is a recognition* that personal good has communal determinations." Therefore, to explore the category of the outsider *it is beneficial* to analyse Eliot's literary methods, the treatment of her characters and the intellectual tropes concerning realism. (BAWE,essay6016a)

- For this reason, *it is important to* start with rules in their simplest forms and bring in more complex parts as simpler rules are mastered. (BAWE,essay6009a)

The above results represented an important distinction in the understanding of the role and use of interpersonal resources in language that might be influenced by the proficient writers' preference to avoid the authorial presence in their writing. When the use of certainty and uncertainty devices was compared, it was found that the ESL learners preferred to use certainty devices more often than uncertainty, as shown in Figure 5.8. In contrast, the BAWE writers used uncertainty devices more than certainty devices. This reflected the findings of (Ağçam, 2014; Muşlu, 2018), who reported that non-native writers use more certainty devices than native speakers and their expert L2 learner counterparts. Previous research also claimed that 'anticipatory it' occurs mostly in participant-oriented functions (Hyland, 2008b; Staples et al., 2013; Chen and Baker, 2016; Shin, 2018; Liu and Chen, 2020), a finding that was also confirmed by the present study. As discussed in section 5.3.2, ESL learners and the BAWE writers showed a similar use of 'anticipatory it + verb/adjective phrase', such as *it is important to, it is often,* and *it is difficult*, with the BAWE writers using fewer participant-oriented structures than the ESL learners. The reason for the ESL learners' overuse of participant-oriented bundles may be due to their preference for personality and impersonality in their academic prose.

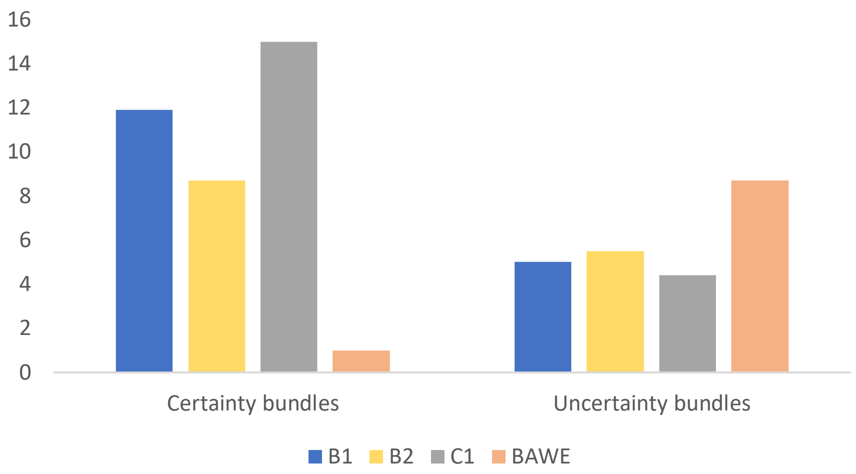


Figure 5.8. Comparison of personal certainty and uncertainty LBs across the sub-corpora.

As shown in Figure 5.8, when the use of personal certainty and uncertainty was compared across the sub-corpora, personal uncertainty was found to be less common than that of personal certainty LBs, such as *I think that, I think it is,* and *it seems that*. The findings also showed that personal certainty LBs, such as *we need to, I am going to,* and *we have seen* were more common in ESL learners' writing, whereas personal uncertainty LBs were more frequently used by the BAWE writing. Examples of this are as follows.

- Therefore, *we need to* solve this problem now. However the first step to solve this problem is to control our sleep times. (B1, essay81)
- In this essay, *I will discuss* whether we need to reconsider the types of aid we give to poorer countries. (B2, essay175)
- However, *I think that* all people should remember their history and pass it down to the next generation because this knowledge is irreplaceable and priceless for every person. (C1,essay222)
- *In my opinion* it is not easy to find appropriate listening material for students, exept the materials which are in the coursebook. (BAWE, essay3150b|)

These findings supported the view of Hyland (2005) that personal certainty LBs are used more frequently in learner corpora. He stated that self-mention represents a central pragmatic feature of academic discourse since it contributes to the writer's construction of a text and a rhetorical self. The authorial pronoun is a significant means of promoting a competent scholarly identity and gaining acceptance for one's ideas.

Therefore, self-mention is a powerful rhetorical strategy for emphasising a writer's contribution (Hyland, 2002, p.110). Meanwhile, Uysal (2012) examined the argument preferences of Turkish writers when writing in Turkish (L1) and English (L2), analysing indirectness markers, such as *I think it is* and *I believe that*, and finding a similar use among the participants. Moreover, when writing in English, the writers used these devices less than when writing their L1 language, with the Turkish writers being influenced by their L1 when writing in their L2 language.

Given the characteristics of the argumentative essay, in which writers express their personal opinion, it was expected that a number of stance expressions, such as *I think that* and *we need to*, would be present in the sub-corpora in the current study, since learners use stance expressions to present a powerful voice. In the ESL learners' writing, there was a clear overuse of the stance function, compared with the BAWE writing, with the ESL learners in general using speech-like bundles, such as *I think that, I think it is,* and *I want to* more than the BAWE writers. The underuse of stance expressions in the BAWE corpus may be due to the belief that such expressions are not favoured in academic research, which is governed by views that value empirical and quantitative objectivity (Mirhosseini, 2017, p.269), and which prefer the use of a "neutral and detachedly descriptive" language.

5.3.5 The relationship between structural and functional categories

Previous research reported a strong relationship between the structural and functional distribution of LBs (Biber et al., 2004; Pan et al., 2016). For instance, most participant-oriented bundles consist of clausal bundles, whereas phrasal expressions tend to be composed of noun-phrase and prepositional-phrase bundles. Table 5.29 below represents the distribution of structural types of LBs across functional categories. It can be seen that the majority of the research-oriented and text-oriented bundles were composed of noun-phrase and preposition-phrase bundles, while the participant-oriented bundles in all the sub-corpora were dominated by verb-based bundles. The associations identified in each sub-corpus are discussed in this section.

Table 5.29. The relationship between the structural and functional categories across the sub-corpora. (N = Number of bundles; % = Percentage in each sub-corpus)

Type	**Research-oriented**		**Text-oriented**		**Participant-oriented**		**Chi-square result**
	N	%	N	%	N	%	
B1							
Noun-based	52	73%	8	11%	11	16%	
Preposition-based	41	59%	5	7%	23	33%	102.77**
Verb-based	96	39%	117	47%	34	14%	
Other	19	28%	14	21%	34	51%	
B2							
Noun-based	42	64%	22	33%	2	3%	
Preposition-based	34	54%	22	35%	7	11%	77.37**
Verb-based	78	37%	31	15%	103	49%	
Other	11	21%	21	40%	21	40%	
C1							
Noun-based	71	72%	27	27%	1	1%	
Preposition-based	41	59%	24	35%	4	6%	123.94**
Verb-based	111	45%	14	6%	120	49%	
Other	29	38%	22	29%	25	33%	
BAWE							
Noun-based	140	71%	52	27%	4	2%	
Preposition-based	68	52%	56	43%	7	5%	164.86**
Verb-based	169	43%	80	20%	143	37%	
Other	17	20%	46	53%	24	28%	

In the B1 writing, the chi-squared test showed a significant association between structural and functional categories, *x*2 (102.7) and df of 6, a significant *P*-value at p < .05. Further analysis using the standardised residuals (R) revealed that the research-oriented bundles were associated significantly with the noun-based category, the text-oriented bundles were associated significantly with the verb-based category, and the participant-oriented bundles were associated significantly with the ‘other expressions’ category.

In the B2 sub-corpora, the chi-squared test demonstrated a significant association between the structures and functions of the LBs, with a chi-squared value of 77.37 and df 6, significant at *P*<.05. The R-value revealed that the research-oriented bundles were also associated highly with the noun-based category, and the participant-oriented bundles were associated significantly with the verb-based category. Unexpectedly, the text-oriented bundles did not show any significant distribution in the B2 writing.

In contrast, the statistical analysis of the C1 and BAWE sub-corpora detected a greater significant relationship between the structures and functions categories, with a chi-squared value of 77.37 and df 6, significant at *P*<.05 in the C1 writing, and a chi-squared value of 164.86 and df 6, significant at *P*<.05 in the BAWE writing. Interestingly, the R values of both sub-corpora showed a similar significant increase in the noun-based category in the use of research-oriented bundles, preposition-based bundles in the text-oriented category, and a greater significant increase of verb-based bundles in the participant-oriented category.

In short, the relationship between the structural and functional categories was evident in the four sub-corpora in this study, in which the LBs incorporating noun-based bundles were usually highly associated with the research-oriented bundles across the four sub-corpora. The strong bond these categories can be attributed to the extensive use of noun phrase bundles, such as *a lot of*, *the end of the*, and *the majority of*, to describe the location, time, quantity, and procedure of academic essays.

It should be noted that the association of noun-based bundles and the research-oriented function has been proven to be used significantly in written form, whereas verb-based bundles tend to be prominent in the participant-oriented function in a wide range of spoken language (Biber et al., 2004; Hyland, 2008a; Chen, 2008; Beng and Keong, 2017).

The results, therefore, provided tentative initial evidence that the structural categories and discourse functions of LBs were closely associated. The results also showed that the B1 and B2 writers tended to use a similar number of LBs, the structure and function correlations of which differed from those in the C1 and the BAWE writing.

The use of LBs by ESL learners in their academic writing is an area in which less proficient writers sometimes deviate from more proficient writers. Hence, more proficient ESL learners lead less proficient writers in the transition to expert writing, as measured by the grammatical features of LBs. This reflected the findings of previous studies, such as that conducted by Biber Biber et al. (2004); Cortes (2004); Hyland (2008a), showing that writers in different disciplines, genres, or registers exhibit different choices in their preference of linguistic devices.

5.3.6 RQ2 discussion

RQ2 What differences exist in the structures and functions of LBs in ESL learners B1, B2 and C1 argumentative essays and proficient student writers?

The structural investigation of the LBs was conducted broadly using a modified scheme established by Biber et al. (1999) that divided the bundles into three structural types: noun-based, verb-based, and preposition-based. The bundles that ESL learners B1, B2 and C1 mainly used belong to Verb-based structural group such as '1st/2nd person pronoun + VP fragment' and '(Verb/adjective) + to-clause fragment'. Three major findings can be summarised according to the results obtained by the chi-squared analysis discussed in this section. First, there were significant differences between the ESL learners in terms of the usage frequency of the different structural categories into which the target bundles were classified. Second, although all the sub-corpora contained more verb-based bundles, only the difference in the usage frequency of the C1 writers was significant. Third, a detailed analysis of the structural sub-categories also showed a significant increased use of noun-based bundles by the C1 and BAWE writers, and of preposition-based bundles in the B1 and B2 writing, indicating that the distribution of the different categories of LBs in the ESL learners' essays grew closer to those in the academic prose of English, since Biber et al. (2004) found that academic writing in English relies on noun and preposition phrases.

A detailed analysis of the difference between the B1, B2, and C1 learners and the BAWE writers, in terms of the usage frequency of the structural sub-categories. We found that 'noun phrase with of-phrase fragment', 'other prepositional phrase expressions', '(Verb/adjective) + to-clause fragment', '1st/2nd person pronoun' and noun phrase 'with other post-modifier fragment' were the five most varied, and also most frequent structural patterns in the ESL learners' sub-corpora, and that the proportional distribution of the target bundles used across the learners' sub-corpora was broadly similar, with the exception of the different use of '1st/2nd person pronoun'.

An important structural difference between the most frequent LBs present in the ESL learners' and the BAWE sub-corpora was the fact that the former favoured the use of verb-based bundles beginning with personal pronouns, particularly 1st and 2nd person (*I, you, we*) in the initial position of the verb-phrases in the academic essays. The representative target bundles classified under this subcategory in ESL learners' sub-corpora, including *I have to, I think that,* and *I think it is*, were rarely found in the BAWE sub-corpus. These bundles are most commonly used to express doubt or uncertainty in the English language, and constitute a common pattern in L2 students' EFL argumentative writing, as confirmed by (e.g., Hong, 2013; Kim, 2013; Yoon and Choi, 2015b). However, these bundles were found to be highly personal and spoken-like, which points to learners' writing as being more assertive and less tentative than that found in the RC. In the English language, these bundles are those most commonly used to indicate doubt or uncertainty, but have a great subjectivity, and are not commonly used in academic writing. A possible explanation for the high proportion of personal pronouns might have resulted from the practical notion of the argumentative essays, which required demonstration and illustration to provide information to support the view presented. As discussed in section 5.2.6, ESL learners employed more LBs that satisfied the specific purposes of argumentative writing. It might have been expected that students would have given priority to stance bundles for such a purpose; however, stance bundles take a large proportion in their bundle use as the functional analysis shows. Though personal pronoun comprises one important element of stance in academic discourse (Hyland, 2005), bundles including personal pronouns words are necessarily stance bundles as shown in the functional analysis. While some of them (e.g., *I think that, I believe that*) do express personal opinions, feelings or attitudes, others may just refer to the writers themselves but generally convey other discourse functions such as quantification (*one of my*), or description (*I was very*), to name but a few. Therefore, this fact highlights why the functional classification of LBs needs to take into account the discursive context in which they are specifically used.

Another distinguishing feature of the ESL learners and the BAWE writers was the use of 'passive verb + prepositional phrase fragment', which indicates the impersonal voice for a locative or logical relation (Hyland, 2008b). The findings were surprising, because the passive voice is usually employed in formal writing, such as academic papers, in which actions themselves are often considered to be more important than the person or object that performs the action. The use of passive voice in the BAWE writing exhibited the general tendency of academic writing to contain more of this voice, which is also associated with stance expression (Biber et al., 1999). Since the BAWE writers were regarded as experts' writers, the BAWE sub-corpus was likely to contain more academic expressions than those of the ESL learners. Therefore, the infrequent use of this structure by the ESL learners was an indication of misuse. While it was beyond the aims of this study to examine this point further, it was a feature that reflected the distinguishing characteristics of the ESL learners' and the BAWE writers' essays. Therefore, although LBs are frequently used in ESL learners' writing, some written functions are rarely used at all levels. This finding contradicted that of Hyland (2008b); Wei and Lei (2011), who found that at least advanced learners use the passive voice far more frequently than the professional writers.

A comparison of the structural distribution of LBs across the ESL learner sub-corpora produced another interesting finding. Although there is a wider variety of noun-based bundles in B1 and B2 writing, their overall frequency across the levels was not sufficiently high to be more frequent than in the C1 and RC sub-corpora. Therefore, the distribution of different categories of LBs in high-level (C1) essays by ESL learners' essays has shared some characteristics of LBs in academic writing. In an attempt to explain the results in the context of expert writers versus L2 learners, previous research (Biber et al., 1999; Hyland, 2008b; Chen and Baker, 2010; Liu, 2012) found that the prevalence use of noun-based bundles is a distinctive feature of academic writing, and an indication of an author's higher proficiency level.

A detailed investigation of the analysis of the most frequent LBs in ESL learners' sub-corpora revealed that one main three-word LB contributed to this increase in B1 and B2 sub-corpora, namely "*a lot of*". B1 and B2 learners' writing relies heavily on the concept of quantity, to the extent that a number of quantifiers become statistically significant. The increased use of a number of quantifier bundles thereby creates informal in the rhetoric of B1 and B2 students' English argumentative writing. The case of "*a lot of*" and their shared local textual function confirms this point. The salient overuse of the quantifier bundle "*a lot of*" in ESL learners' writing may also be related to the tendency for overstatement in L2 writing. Chen and Baker (2010) discovered that even at advanced level, L2 learners tended to create an impression of generalisation-tone in their writing. Hinkel (2005) also claimed that writings by many L2 learners are full of overstatements that are rare in formal writing and create a colloquial style of writing when presenting their arguments. A comparison of learner corpus data with written and spoken native, Gilquin and Paquot (2007) point out that learners also keen to use more overstatement features in their writing which make their writing sound as spoken-like, as this is a common feature in spoken discourse.

Therefore, whilst this study did not confirm the overstatement at C1 level, it did partially substantiate that at B1 and B2 levels.

It would be better to increase learners' awareness that the meaning of such phrases is nonspecific and subjective, and that there often are generally accepted ideas for what constitutes 'a lot'. This knowledge could better help learners to use these bundles effectively. The findings suggested that ESL students' writing shared certain common characteristics with the spoken register. This may be associated with the influence of teaching, which emphasises expressions, and might also reflect ESL students' learning strategy of memorising taught expressions.

Turning to the functional analysis of the target bundles, the first analysis explored the concordance lines, in order to allocate the target bundles in this study to a corresponding functional category adopted from Hyland (2008b) categorisation (research-based, text-based, and participant-based), according to their discourse function in the text. The type and the normalised token distributions of the LBs were similarly distinguished across the sub-corpora. Similar to the structural analysis, the functional analysis found little evidence of clear distinctions between the ESL learners and the BAWE writers' essays. The three major findings can be summarised according to the results of the LB functional analysis.

First, research-oriented bundles were the most prevalent bundle type in both the ESL learners and the BAWE sub-corpora. As Hyland (2008b, p.49) explained, this bundle type "helps writers to structure their activities and experiences of the real world", and includes for example, *the importance of*, and *the end of*. Second, the difference between this category and the other two categories was relatively extreme, a finding that reflected the primary feature of academic writing, namely the focus on the subject of the research. The increased use of research-oriented bundles was consistent with the findings of (Ädel and Erman, 2012; Ruan, 2017). Third, in terms of their use of the other two functional categories, the ESL learners employed more type and token participant-oriented bundles than text-oriented, a pattern similar to that identified by (Biber et al., 2004). In contrast, the BAWE writers used more type and token text-oriented bundles, reflecting the findings of (Huang, 2014), who reported that both junior and senior Chinese learners in both oral and written modes favoured the use of research-oriented and participant-oriented bundles over text-oriented bundles. Regardless of their language proficiency, academic writers focus on facts and evidence related to an essay's topic. Since the writing task in this study required the ESL learners to compose an argumentative essay, providing a convincing argument in support of the conclusion they had reached, the greater use of research-oriented bundles may be a sign that their writing at all levels relied heavily on facts and evidence to support their argument, as with the proficient writers. This was supported by the finding that the ESL learners used slightly more participant-oriented bundles in their writing than the proficient student writers in the BAWE sub-corpus, suggesting that the non-native writers preferred to argue their position using opinion statements.

However, the greater use of research-oriented bundles by the B1 and B2 learners can be interpreted as a sign of informal writing. The potential reason for this lies in examining the research-oriented sub-categories, namely location, procedure, quantity, and structure, used by the learners. Quantification expressions that specify the measurable extent or amount were the most frequent LBs employed, with an overuse at the lower level—this sub-category comprised about a third of the total research-oriented tokens in the B1and B2 sub-corpora. The main three-word LBs contributing to this increase was the informal quantifier bundles *a lot of*, which is associated with the spoken register as discussed above. This finding suggested that the B1 and B2 writers were more reliant on quantifier bundles, a fact that may be related to the issue of overstatement in L2 writing; thus, their usage did not reflect general academic-register bundles as shown in structural analysis. This may also be a technique employed by low-level learners to address their limited vocabulary and store of LBs, as with "phraseological teddy bears" (Hasselgård, 2019, p.15). The learners relied on certain phrases that they felt comfortable using them. They may have been exposed to these structures through reading and therefore believed they could be used safely when required. By combining sentence fragments from their reading, the less proficient ESL writers in this study were able to enhance their vocabulary limitations using structures they could be certain would be considered 'proper' English.

Another important issue arose when the participant-oriented function was compared across the corpora, as it was evident that stance expressions were used more frequently by the ESL learners than the BAWE writers. This finding reflected that of Cooper (2013), who reported the presence of a high frequency of LBs that performed a stance function in an IELTS sub-corpus. These expressions refer to "ways writers explicitly intrude into the discourse to convey epistemic and affective judgments,

evaluations and degrees of commitment to what they say" Hyland (2008b, p.167), such as *I believe that*, *I think it is*, and *it is important*. When each functional category was compared in detail in the present study, it was found that personal stance bundles, such as *I think that*, were used more often than impersonal stance bundles in all the learners' writing. When the use of certainty devices was compared, it emerged that the ESL learners used uncertainty bundles more often than certainty bundles, in contrast with the BAWE writers. Given the characteristic of the argumentative essay that requires writers to express their personal opinions, it was expected that a number of stance expressions, such as *I think that* and *we need to*, would be present in the sub-corpora, as learners use stance expressions to present a powerful voice. In contrast, the underuse of stance expressions in the BAWE writing may be due to the belief that these expressions are not favoured in university writing, which is governed by views that value empirical and quantitative objectivity (Mirhosseini, 2017), and a preference for "neutral and detachedly descriptive" language (Mauranen and Bondi, 2003, p.269).
A detailed examination of the stance LBs revealed that verb-phrase bundles were used frequently by both the B2 and C1 writers confirming the connection between the structural and functional distribution of LBs and reflecting the findings of (Hernández, 2013; Muşlu, 2014).

An examination of the target bundles also revealed a connection between the research-oriented bundles and noun-based bundles in this study, concurring with the findings of (Biber et al., 2004; Hyland, 2008a; Chen, 2008; Beng and Keong, 2017). The evidence for an increase in noun-based and research-oriented LBs in the C1 writers' essays also supported the usage-based framework of language learning (e.g., Ellis et al., 2016; Crossley et al., 2019), as the higher-level ESL writers in this study reflected the distributional characteristics of LBs in academic writing over time. The strong relationship between the participant-oriented bundles and the verb-based bundles may be due to the extensive reliance on first person pronouns in the participant-oriented bundles. Moreover, the verb-based LBs that included 'to clause fragments', such as *to sum up* were also found extensively in the participant-oriented bundles among the ESL learners' sub-corpora.

To conclude, the findings of the structures and functions distribution revealed variation in the use of LBs across the levels. In brief, the structural and functional analysis revealed the following.

- The structural distribution of the target bundles used across the learners' sub-corpora generally presented the same picture, except for the use of '1st/2nd person pronoun' and 'noun phrase with of-phrase' bundles;
- The majority of the structural types in the ESL learners' sub-corpora were based on verb-based bundles, the majority of which performed stance expressions, thus "conveying the writer's attitudes and evaluations" Hyland (2008b, p.14);
- C1 writing contains more written-like elements in comparison with B1 and B2 levels;
- ESL learners used of *I* plus an active verb, as opposed to the passive voice. Indeed, the ESL learners in the present study exhibited an overall higher proportion in the sub-categories that are more representative of spoken language.
- Both ESL learners and BAWE writers show the same patterns of the usage identified by previous studies that among the three functional categories, when used in academic writing, research-oriented bundles represent the most commonly used functional type, the difference between this category and the other two categories being quite extreme.
- When it comes to usage of the other two functional categories, ESL learners B1, B2 and C1 were different from the BAWE writers and they used more of the participant-oriented bundles than text-oriented, as opposed to the pattern found in (Cooper, 2013; Chen and Baker, 2016). The difference between these two categories is less pronounced than that between them and the research-oriented, text-oriented concern the organisation and meaning of the text, while participant-oriented focuses on the writer or reader of the text. Hence, it seems that regardless of the ESL learners' levels, they focus more on providing more subjective, personal judgment, convey the writer's attitudes and establishing a closer relationship with the reader than on presenting to the reader the interconnectedness of their ideas, the structure, coherence and cohesion of their texts.

5.4 Keyness analysis

As discussed in section 2.5, the keyness principle concerns keyword analysis. A keyness analysis of language varieties is often conducted to encompass the differences between ESL learners' academic writing. The term 'keyness' can be used to indicate whether a specific bundle is significantly overused or underused in a target corpus, when compared with a reference sub-corpus (RC). In this chapter, the following RQ will be addressed:

> **RQ3** What are the characteristics of keybundles deployed in ESL learners' essays in comparison with the BAWE writers?

As discussed in section 4.7, keybundles were extracted using the *Keyword* function of *WordSmith* tool (*WST*) by comparing each ESL learners' sub-corpora with the BAWE sub-corpus. In addition, two statistical measures were considered when identifying keybundles, Log likelihood statistic and the BIC score. Both metrics were used to complement the previous quantitative comparison made between the ESL learners' sub-corpora with the BAWE.

The *P-value* representing the degree of danger of error was set at the default *WST* value of 0.000001 (one in a million) to decrease the number of keybundles used for the investigation. After setting the *P-value*, the ESL learners' bundle lists were compared individually with the BAWE bundles list. Lists of the keybundles were then created by the *WST*. Despite the difference between the metrics, the result showed that both BIC score and the LL statistic in *WST* produced the same ranking order of keybundles identified in this study. Therefore, the identified keybundles from the B1, B2, and C1 sub-corpora were finalised with a keyness value of the BIC score. All the keybundles retrieved were then crosschecked to ensure that the keybundles met the cut-off point and the dispersion criteria employed in this study, namely they occurred at least four times in 100,000 words, across three texts. It is important to note that this study focussed only on overused keybundles; underused keybundles form an essential area for future research.

A keyness analysis of the overall ESL learners' sub-corpora and the BAWE sub-corpus, with the latter as the RC, found that the bundles that referred to the essay topic concerned, such as *to global warming*, *temperature of the*, *in city life*, *science and* *technology*, and *in the countryside*, had an extraordinarily high keyness value in the ESL learners' sub-corpora. As Scott and Tribble (2006, p.63) explained, "keywords are mostly connected to what the text is about and are important to it, with some intruders which suggest something about the style, and which often repay further analysis". These 'intruders' refer to bundles that include grammatical words, such as *we*, *I*, *can*, and *these*, that are suggestive of the writing style. This was an expected result, since the essay topics required the ESL learners to discuss ideas related to these topics, and to form an argument to convince a reader. As these bundles were content-based, this study focused on the general bundles used differently by the participants at the three different proficiency levels, comparing them with the BAWE writers. As discussed in section 4.6.4, all the content-based bundles were excluded from the analysis, as they did not reflect the use of general academic language.

After excluding the content-based bundles, a total of 43, 45, and 122 three-word keybundles were significantly ($P < 0.000001$) used more by the B1, B2, and C1 learners, respectively, according to their keyness value (See Appendix G). A significant difference was found between the keybundles identified in the three sub-corpora; around 11% of the bundles were significant in their usage frequency in the B1 and B2 sub-corpora, while 23% of the bundles were significantly frequent in the C1 sub-corpus, with a higher keyness value. The highest keyness value was found in C1 list for the bundle *I think that*, followed by *a lot of* in B2 and B1 sub-corpora lists. Thus, it was evident that the C1 writers had more keybundles than the B1 and B2 writers, compared with the RC. A sample of 20 keybundles from each ESL learners list was subjected to more detailed analyses, according to their keyness value, as shown in Table 5.30.

Table 5.30. Top 20 three-word keybundles in the B1, B2, and C1 sub-corpora with a significantly different usage frequency to those in the BAWE sub-corpus. (K = keyness value; Italic = top 20 most frequent bundles)

B1 level		**B2 level**		**C1 level**	
Keybundles	**K**	Keybundles	**K**	**Keybundles**	**K**
a lot of	131	*a lot of*	144	*I think that*	890
first of all	102	*first of all*	132	*first of all*	513
I want to	102	*point is that*	117	*second of all*	431
day by day	73	*I want to*	117	*I believe that*	372
I do not	68	*we need to*	78	*I think it*	300
in my opinion	68	*in this world*	73	on the other	232
want to become	58	*they do not*	67	*to sum up*	190
in the field	58	*there are many*	65	*I do not*	131
we need to	58	in my opinion	64	*I will give*	131
point is that	54	the whole world	54	to support my*	131
we have to	54	is to become	54	*I want to*	126
to get a	49	*on the other*	50	*his or her*	124
I am going	49	we have to	49	for several reasons	111
around the world	49	are very different	49	*the opportunity to*	110
things that are	44	I do not	49	when I was	106
all over the	39	I will discuss	49	the one hand	102
you have to	39	they have to	44	he or she	102
there are many	39	are aware of	39	aspect of this	97
to become a	34	to go to	39	which I will	97
I am very	34	day by day	39	*it is a*	95

Significant at ($p < 0.000001$). A bundle which is positively key occurs more often than would be expected by chance in comparison with the reference corpus.

In terms of the four-word bundles' list, a total of eight (B1), nine (B2), and 58 (C1) keybundles were significantly ($P < 0.000001$) overused in comparison with BAWE. A significant difference was found between the three sub-corpora, with approximately 30%, 32%, 39% of the B1, B2, and C1 sub-corpora being keybundles. As can be seen in

Table 5.31, the highest keyness value in B1 list was for the bundle *in the field of*, the bundle *in this essay I* in the B2 list, and the bundle *I think it is* in C1 list. ESL learners from different proficiency levels exhibited different use of LBs from each other.

Table 5.31. Top eight 4-word bundles in B1, B2 and C1 sub-corpora with significantly different frequency from those in BAWE corpus. (K = keyness value; Italic = top 20 most frequent bundles)

B1 level		**B2 level**		**C1 level**	
Keybundles	**K**	**Keybundles**	**K**	**Keybundles**	**K**
in the field of	58	in this essay I	57	I think it is	256
first of all it	44	on the other hand	45	to sum up I	223
I am going to	44	they are aware of	34	on the other hand	198
in this essay I	38	this essay will examine	34	in the following paragraphs	126
anywhere in the world	29	a second point is	34	I think that the	126
a second point is	29	different from each other	34	reasons to support my	111
I do not think	29	another point is that	29	in conclusion I think	102
increasing day by day	29	there are a lot	39	reasons which I will	97

Judging from the components making up a bundle in Table 5.30 and Table 5.31, the bundles with the first, fourth, and fifth highest keyness value in the three-word list were *I think that, I believe that*, and *I think it*. It suggests that the advanced ESL learners were more likely to use first person pronouns than the BAWE writers in their academic essays. Interestingly, the bundle *I think that* can also be seen as an overstating bundle, as Aijmer (2001) connected the increased use of *I think* by ESL learners in their argumentative essays with the attempt to build their argument and make it more persuasive, which might be communicatively unnecessary, or overly wordy. An example of the first person pronoun with the highest keyness value from the C1 sub-corpus is as follows:

- *I think that* computers play an essential role in our lives and they bring numerous benefits to our community. (C1, essay 92)

Another form of keybundle that was used significantly by the ESL learners, compared with the BAWE writers, was the connector bundles (e.g., *on the other hand)*. A total of 32 connector keybundles, occurring more than 800 times in ESL learners' sub-corpora, were found in the keybundle lists. Both the frequency and the keyness analysis confirmed that these bundles were used significantly by the ESL learners, suggesting that it might be a characteristic of ESL learners' language use. Therefore, it appears that all three levels use the first-person pronoun and connectors categories more significantly in their writing. In light of that, the next sections will discuss these two key categories in detail.

5.4.1 Connectors

The term 'connectors' is used in this study to refer to lexical items that demonstrate a cohesive tie between sentences and paragraphs of an essay, in order to show a coherent relationship between them. Examples of these bundles are *on the other hand,* and *in addition to*. They were referred to a lot of alternative terms such as "linking adverbials" by (Biber et al., 1999), and cohesive conjunctions (Halliday and Hasan, 1976). According to Biber et al. (1999, p.875,558), the primary function of connectors is "to state the speaker/writer's perception of the relationship between two units of discourse" and "to make semantic connections between spans of discourse of varying length". Therefore, using connectors in argumentative essays is one way to achieve paragraph coherence, a key factor in assessing second language writing (Rachmawati and Susanti, 2016). At the same time, the use of connectors is often found problematic for L2 learners, and previous research realised that non-native learners tend to both overuse and underuse connectors in their writing (Granger and Tyson, 1996; Hinkel, 2004; Leńko-Szymańska, 2008; Shea, 2009). In this sense, the usage of connectors in ESL learners' sub-corpora worth to be closely examined in their argumentative essays.

A number of categorisations have been made for connectors so far. For instance, Biber et al. (1999) classified them according to their semantic categories, which are enumeration, summation, apposition, result/inference, contrast/concession, and transition. Liu (2008) also divided them into four main categories as an additive, adversative, causal/resultative and sequential. Apart from these, the classification proposed by Quirk et al. (1985) has been widely used in language research. Quirk and his colleagues classified connectors into seven categories based on their semantic use (Table 5.32). This classification has been adopted due to its clarity, and because it was commonly used in academic texts. The main seven categories concerned can be divided further into sub-categories, which were not considered in the present study, due to its scope and purpose.

Table 5.32. Semantic categories of connectors. Source: Quirk et al. (1985, p.624-636)

Classification of connectors
Listing connectors assign numerical labels to the items listed (e.g., *first, second, third*). In addition, they indicate relative priority and create integral structure to a text. They can also signal that an item has a similar force to a preceding one (e.g., *equally, similarly*) or, on the other hand, assess an item as adding greater weight to a preceding one (e.g., *above all*) (Quirk et al. 1985: 634–637).
Summative connectors precede an item which is to be looked at in relation to specific items that have gone before. The same applies also to **appositive conjuncts**, but while summative conjuncts introduce an item that embraces the preceding one (e.g., *all in all*), the appositive conjuncts rather express the content of the preceding item/s (e.g., *for instance*) (Quirk et al. 1985: 637).
Resultive connectors indicate a conclusion, summary, a result, etc. (e.g., *as a result, in conclusion*). In a similar way, inferential conjuncts indicate a conclusion that is based on logic and supposition (e.g., *in other words*) (Quirk et al. 1985: 638).
Contrastive connectors "present either contrastive words or contrastive matter in relation to what has preceded" (Quirk et al. 1985: 638) (e.g., *on the other hand, in contrast, however*).
Transitional connectors are used to "shift attention to another topic to a temporally related event" (Quirk et al. 1985: 639). *By the way* and *in the meantime* are examples of this type of connector.

Table 5.33 presents a list of the connectors that were keys in the ESL learners' sub-corpora, compared with the BAWE sub-corpus, along with their normalised frequency and keyness value. Both the frequency and the keyness analysis confirmed that these bundles were overused in ESL learners' writing, suggesting that they might be characteristics of ESL learners' language.

Table 5.33. Keybundles connectors in the ESL learners' sub-corpora. (Bold = shared by the three levels; italics – shared by two levels)

Connectors	Frequency	Keyness	Connectors	Frequency	Keyness
			B1		
first of all	21	102	First of all it	9	43
to sum up	12	29	*a second point is*	6	29
			B2		
first of all	27	132	*on the other hand*	39	31
on the other	42	50	*a second point is*	7	20
to sum up	12	30	another point is that	6	16
the main reason	6	29	a third point is	6	16
			C1		
first of all	106	521	*on the one hand*	20	96
second of all	89	430	first of all I	18	87
on the other	48	232	second of all I	17	82
to sum up	50	189	in addition to those	13	62
the one hand	21	101	in addition to these	11	63
in addition to	29	84	first of all a	9	43
my point is	10	48	however I believe that	9	43
because it will	7	38	in order to succeed	8	38
one more reason	6	29	however I think that	7	33
the main reason	6	29	to summarize I think	7	33
to sum up I	46	222	second of all a	6	29
on the other hand	41	198	in order to get	6	29
in conclusion I think	21	101			

The table above shows that a total of four connectors at B1 level, eight at B2 level, and 25 at C1 level were keys, compared with the RC. The results indicate that in argumentative essays, ESL learners use connectors in a pattern related to their writing development, measured by the increased use of LBs, as various connectors were used more frequently in C1 level writing than in the B1 and B2 levels. A closer inspection at Table 5.33 shows that ESL learners' lists are full of listing connectors *first of all, a second point is, another point is,* and *second of all a*. These connectors are used to express simpler relationships between propositions, as shown in the examples below.

- *First of all*, I think active video games are a good way to keep healthy body. There are lots of types of exercise you can do such as basketball and dance you can play with them in your home. (B1, essay235)
- *First of all*, cities are seen to offer different job opportunities since there are very different branches of businesses in cities. (B2, esay217)
- *First of all*. It is better and more efficient to build universities in cities because lots of students can join them. Secondly, small universities will lack support and will be boarding. (C1 sub-corpus, essay 26)

Another connector type that showed significant use across the levels was the summative category. This was expected, as Biber et al. (1999) observed that summative connectors are used as discourse markers when composing that conclusion of a text. The example below shows how the ESL learners in the present study used these bundles to summarise and link the information in their essay.

- *To sum up*, we have look at different ways to use technology to tackle environmental problems instead of using it for play. (B1, essay 35)
- *To sum up*, helping decrease the effects of global warming is not only good for yourself but for everyone and for our Earth. (B2, essay244)
- *To sum up* all mentioned above, I think that we need to be more careful with the natural resources we use and, moreover, we can do something to preserve them. (C1, essay188)

The following significant connector category was contrastive connectors use to introduce information that contrasts with or differs from information given in previous sentences, which often lead to main points that writer want to make (e.g., such as *on the other hand* and *on the one hand*) (Biber et al., 1999), and therefore, “contribute to the interactive nature of academic discourse” and “enable voices other than the author’s to enter the text” (Povolná, 2016, p.57). Although contrastive connectors can also be placed within the sentences, they mainly occur at the beginning of sentences, especially in the academic writing. An example of these connectors is *however*, which was demonstrated by Biber et al. (1999) to be one of the four most frequently used contrastive connectors in academic discourse, and he described it as “uniformly preferred” (ibid, p.889) to indicate contrast by writers of this type of discourse. An example of the use of this connector in the ESL learners’ writing is as follows:

- *However, I believe* that internet gave us more advantages and opportunities than disadvantages and problems. (C1, 180)

Another example of these connectors type is *on the other hand*, which found to be highly used in academic writing (Hyland, 2008b; Leedham, 2011), was one of the most frequent connectors in the B2 and C1 levels writing. For example:

- Village is a small area where there is lack of facilities. *On the other hand,* City is the area where every kind of facility is available. People who live in villages are quite simple but people of cities are modern. (B2, essay343)
- *On the one hand,* saving land in its natural condition brings many benefits. (C1, essay2)

To conclude, the investigation of the connector bundles indicated that the ESL learners, particularly the C1 writers, followed by the B2 writers, tended to rely on these expressions to a greater extent than the B1 writers, therefore, as their proficiency level increased, the ESL learners employed more connectors in their essays. Listing connectors were the most frequent expressions in all the three-level, indicating that ESL learners might know both how and when to use them. Further research should be undertaken to investigate these expressions in ESL learners’ writing.

5.4.2 Personal pronouns

Another category that showed an overuse in the keyness analysis was personal pronouns. The presence and use of these bundles were described by Luzón (2009, p. 193) as “refuting the traditional view that this type of discourse is impersonal and objective”, since they are a powerful strategic resource for the construction of an authoritative self through the realisation of various functions (Ivanič and Camps, 2001;

Martínez, 2005; Leedham, 2011). Previous research included anecdotal disagreement regarding the overuse and underuse of personal pronouns in non-native writing, with some previous studies finding that L2 learners tend to use few personal pronouns (Pan et al., 2016; Hyland, 2002; Wei and Lei, 2011), while some others found a significant use in non-native writing (Mccrostie, 2008b; Lee and Chen, 2009; Leedham, 2011). These studies demonstrated that the excessive use of personal pronouns contributed to the speech-like quality of a learner's academic written production.

Regardless of these contradictory results, the efficient and thoughtful use of personal pronouns generates a positive impression of a writer who has "a confident and expert mind in full control of the material, making judgements and passing comment on issues of concern to the discipline" (Hyland, 2004a, p.123). Especially in argumentative essays, writers have to state their opinions on the topic, as the reader wants to know the writer position on this argumentative issue. Hyland (2004b, p.143) stated that 'self-mention I plays a crucial role in mediating the relationship between writers' arguments and the expectations of their readers', but the incorrect use and overuse of this feature can be a manifestation of inexperienced writers unfamiliar with the genre and register conventions, and an indication of a novice writer (Mccrostie, 2008b).

The keyness analysis found significant use of self-mention expressions by ESL learners at all levels, with increased use over the competency levels. Table 5.34 shows the keybundle lists of the personal pronoun that were used significantly across the levels, compared with the BAWE sub-corpus.

Table 5.34. Lists of keybundles for the personal pronouns in the ESL learners' sub-corpora. (Bold = shared by the three levels; italics – shared by two levels)

Connectors	Frequency	Keyness	Connectors	Frequency	Keyness
B1					
I want to	21	102	I have been	7	34
I do not	14	68	we had a	7	34
we need to	12	58	*I will discuss*	6	29
I am going	10	48	I am going to	9	43
you have to	8	38	I do not think	6	29
I am very	7	34			
B2					
I want to	24	117	*I believe that*	8	39
we need to	16	78	we can say	7	34
we have to	10	48	he has to	6	29
I do not	10	48	*in this essay I*	21	56
I will discuss	10	48			
C1					
I think that	205	890	I think it is	53	256
I believe that	77	512	to sum up I	46	222
I think it	62	430	I think that the	26	125
I do not	27	130	first of all I	18	87
I will give	27	130	I am sure that	17	82
I want to	26	125	second of all I	17	82
I am sure	18	87	I will give my	14	67
I did not	20	58	I think that every	14	67
I prefer to	11	53	I think that it	13	62
I like to	10	48	I did not like	8	38
we need to	9	43	I think that a	7	33
I will list	8	38	I think that this	7	33
I base my	7	33	*in this essay I*	15	33
I think the	7	29			

The table above shows a disparity between the ESL learners' use of several pronoun groupings, confirming that the first person pronouns were the most significant keybundles across the sub-corpora. This result was consistent with that of previous studies, which found that non-native writers make significant use of the first person pronoun *I* in their writing (e.g., Mccrostie, 2008b; Lee and Chen, 2009; Leedham, 2011). Furthermore, the bundle with the strongest keyness at the B1 and B2 levels was

I want to, whereas the bundle *I think that* had the strongest keyness value in the C1 writing. Examples of these bundles are as follows:

- At the last, *I want to* say that cities and villages are good. It is up to you if you want to change or not. you can enjoy living in city and village. (B1, essay141)
- At the end *I want to* say that if you enjoy the beauty of nature than you should go to the village and enjoy it. (B2, essay330)
- *I think that* computers play an important role in our lives and they bring many benefits to society. Moreover, children can learn by use of computers. (C1, essay18)

The increased use of the first-person pronouns (*I want to, I do not, I am going)* highlights the writer's voice when conveying the connotations of an argument. The significant occurrence of the first-person pronoun indicated clearly that it was a predominant feature of their essays that the participants used to show their presence. However, when comparing the three levels, it was apparent that the C1 writers used significantly more personal pronouns than the writers of the other levels, as the occurrence frequency of the first person pronoun use rose sharply from the B levels to the C1 level.

A closer inspection to the first person pronouns across the sub-corpora showed increased use of the opening provider that presents the author’s opinions, ideas, arguments, or judgments (see examples below), and usually collocates with part of the mental process of cognition, such as *think* and *feel* (Halliday and Matthiessen, 2013). As shown in Table 5.34 above, a range of verbs collocate with first person pronouns in ESL learners’ Keybundle lists (*e.g., believe*, *think*, and *know*), for which the C1 writers more frequently used them. Examples of the use of these verbs by the ESL learners are as follows:

- To sum up, *I think that* computer technology gives people several benefits including the chance to improve one's knowledge and be more self-confident, persistent and experienced in this world. (C1, essay116)
- I think most things people buy are not really important, but they are misled by advertisements provided by consumer society. At last, *I believe that* people do not know anymore what they really need. (B2, essay273)

- First of all, *I want to* say that problems make people stronger. By passing difficulties people get important knowledge and experience. (B1, essay381)

The significant use of these bundles can be attributed to the differences between the sub-corpora in terms of the first-person pronouns' discourse functions. The results demonstrated that a great variety of verbs were used to indicate an opinion in the C1 level writing, including *think*, *want*, *believe*, *need*, and *prefer*. Thus, the difference between the sub-corpora lay in the opinion provider's usage, and in the fact that more tokens of those verbs were found in C1 writing. This confirmed that the less proficient ESL writers lacked the expressions necessary to express their opinions or ideas, and arguments in their writing. This lack of variety may indicate a developmental writing problem for less proficient ESL learners.

In summary, the statistical analysis consolidated the previous findings of the most frequent LBs identified in the ESL learners' sub-corpora. It also provided essential evidence highlighting a particular genre for further investigation. The statistical analysis of the keybundles shown in Table 5.30 and Table 5.31demonstrates strong evidence of the difference in use of LBs between the ESL learners' levels.

5.4.3 RQ3 discussion

RQ3 What are the characteristics of keybundles deployed in ESL learners' essays in comparison with the BAWE writers?

The measurements of keyness serve to distinguish LBs, which differ significantly between corpora due to being either overused or underused within one corpus in relation to a reference corpus. In this study, the analysis of differences compares the use of LBs in the BAWE sub-corpus with those in the B1, B2 and C1 sub-corpora. These differences highlight idiosyncratic uses, as well as significant variations between the types of LBs used by ESL learners and proficient student writers. Keyness analysis of the ESL learners and BAWE sub-corpora showed that first person pronouns and connectors were the most prevalent bundles distinguishing ESL learners' writing.

In terms of the first keyword category identified in the ESL learners' sub-corpora, the results revealed that the self-mentioned form is used frequently in ESL learners' writing. That increase can be partly due to how argumentative essays are structured, which requires students to present their own opinions without reference to outside sources. Hyland (1994, P.240) states that "Rather than being factual and impersonal, effective academic writing actually depends on interactional elements which supplement propositional information in the text and alert readers to the writer's opinion'. I believe that the choice of using a certain personal pronoun in academic writing, and particularly in argumentative essays, can often show the writers' position, and their relationship with readers.

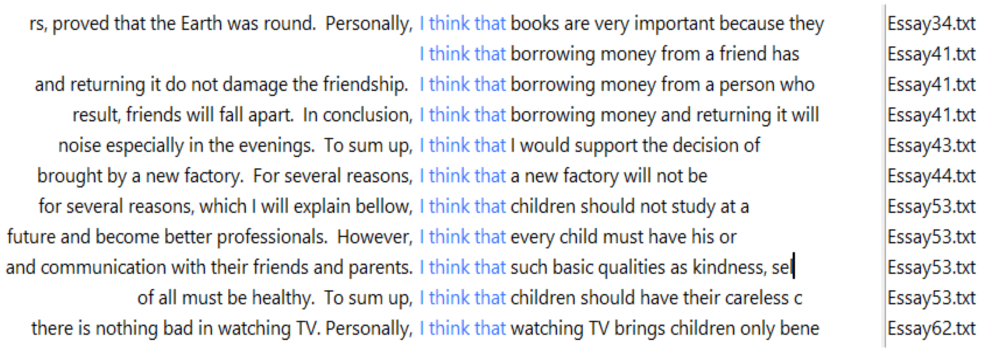
When the first person pronouns bundles were examined, increased use of the opening provider bundles became apparent, which presenting the author's opinions, ideas, arguments or judgments, usually collocating with parts of the mental process of cognition such as thought and feelings (Halliday and Matthiessen, 2013). Thus, the results for using first person pronouns confirmed previous researchers' findings (e.g., Cobb, 2003; Mccrostie, 2008a; Joharry, 2016), where L2 learners favoured the use of first person pronouns more than experts and native English speakers. More specifically, ESL learners use the first person *I* more as sentence-initial than in the middle of their writing. This suggests that the use of the personal pronoun *I* in ESL learners' argumentative essays has a role to play in particular discourse strategies, as supported by concordance analysis and explained in section 5.2.6.

The concordance analysis of the keybundles in ESL learners' sub-corpora revealed the personal pronouns *I* is used mostly to state a purpose (e.g., *in this essay I*) or an argument (e.g., *I think that*), personal matter and to present conclusions (e.g., *to sum up I*). As repeatedly mentioned, this may be a consequence of the argumentative essay type, in which writers are encouraged to share their opinions on a given topic. Thus, writers can create a positive self-representation through appropriate use of these expressions and establish strong writer-reader relationships (Dueñas, 2007). The previous section has shown that the increased use of the personal pronouns in C1 writing may be attributed to the predominant use of the *I think* which is supported the claims by Natsukari (2012); Alward (2019), who found the phrase *I think* is excessively used in EFL learners' essays. In contrast, the increased use of the first person pronouns in B1 and B2 is due to frequent use of the bundle *I want to*. This bundle can describe the writer's personal matters or opinions, such as personal identity and experience. It can be seen that ESL learners are mainly used personal pronouns *I* to stating their position in relation to the essay topic. Therefore, we could claim the ESL learners shared the same features of the written discourse at all levels. Below are some extracts in which the first person pronouns *I* is used.

- Finally, *I want to* express my opinion about how to make this world a better place. (B1, essay 34)
- In conclusion, *I want to* say that if people get high salaries their contribution to society is huge. (B2, essay 132)
- *I think that* computers play an essential role in our lives and they bring many benefits to our society (C1, essay3)

Although the use of first person pronouns can be acceptable in argumentative writing, the overuse of these expressions can make learners' writing lack objectivity and generality which could be problematic in academic writing. Akahori (2007) investigated the use of first person pronouns *I* in American and British university students in the Louvain Corpus of Native English Essays (LOCNESS) with Japanese EFL learners' argumentative essays taken from International Corpus of Learner English. The study found that EFL learners' writings lacked argumentativeness due to subjective perspectives, which was seen in the overuse of first person pronouns. Thus, While the use of personal pronouns can shape the writers' identities as academic writers, the overuse of these expressions might make their writing tedious and wordy, particularly in short essays.

A possible reason for the overuse of certain expressions is a lack vocabulary knowledge. As Hinkel (2005) found that ESL learners repeat the same idea, which is probably due to the lack command of L2 vocabulary. A closer inspection of the concordance lines confirmed the overused of these expressions by some ESL learners across the sub-corpora, as shown in Figure 5.9.



```
   rs, proved that the Earth was round.  Personally, I think that books are very important because they     Essay34.txt
                                                     I think that borrowing money from a friend has           Essay41.txt
     and returning it do not damage the friendship.  I think that borrowing money from a person who           Essay41.txt
          result, friends will fall apart.  In conclusion, I think that borrowing money and returning it will Essay41.txt
        noise especially in the evenings.  To sum up, I think that I would support the decision of            Essay43.txt
    brought by a new factory.  For several reasons, I think that a new factory will not be                    Essay44.txt
    for several reasons, which I will explain bellow, I think that children should not study at a             Essay53.txt
 future and become better professionals.  However, I think that every child must have his or                 Essay53.txt
and communication with their friends and parents. I think that such basic qualities as kindness, sel          Essay53.txt
                   of all must be healthy.  To sum up, I think that children should have their careless c     Essay53.txt
   there is nothing bad in watching TV. Personally, I think that watching TV brings children only bene        Essay62.txt
```

Figure 5.9. A screenshot of the concordance line of the bundle I think that in C1 sub-corpus.

To conclude, the overuse of the first person pronouns suggests that these expressions among ESL writers are a predominant feature of their argumentative essay to show their presence, with an increase of the first person *I* than other personal pronouns. In addition, the overall increase in the frequency of personal pronouns *I* across most learners when writing argumentative essays suggests that some learners are not to some extent aware of the genre-specific characteristics. Yet, the overuse of certain bundles, such as *I think, and I want to* shows that learners tended to rely on a limited number of bundles they feel comfortable using in order to achieve the purpose of delivering their arguments. The comparison between ESL learners' levels disclosed different patterns of such bundles to the argumentative writing. C1 learners used *I* more frequently and with more variety than B1 and B2 levels, overall suggesting increases in the knowledge of using first person pronouns bundles at C1 level with respect to the characteristics of argumentative genre. Thus, in terms of pedagogical implications, raising awareness of the first-person pronoun's functions can serve in argumentative writing as well as can help learners acquire the knowledge of using these expressions most effectively in their academic writing.

In addition to the first-person pronoun *I*, this thesis identified another salient Keybundles type in ESL learners' sub-corpora; i.e., connector expressions. It was found that connector expressions (e.g., *first of all, on the other hand*) are more frequently used by ESL learners. The analysis of all the connectors used in the ESL learners' essays shows learners were conscious of using various connectors to organise their arguments and to provide counterarguments. Such connectors enhance cohesion, and improves the links between sentences (Field and Oi, 1992). Crewe (1990, p320) argued that inclusion of connectors in L2 written discourse is one way to achieve a more academic style and even to "impose surface logicality on a piece of writing where no deep logicality exists" (p. 320), an issue which other researchers have discussed (e.g., Paquot, 2010). Genre is clearly an important factor, for example the use of "chains of connective devices" (Paquot, 2010, p.174) in the short essays which characterise language tests due to its need not only for covering the topics but also making a logical relationship between sentences within a controlled number word count.

However, the result also showed discrepancies in the use of connectors between the three ESL learners' levels. Particularly, C1 writers followed by B2 writers, tended to rely on connector expressions more frequently than B1 writers do. Consequently, through the abundant use of connector expressions, ESL learners tried to achieve overall coherent texts. It can therefore be assumed that as the level increase, writers become aware of the importance of building and structuring their arguments more clearly; therefore, the reader can follow the writer's reasoning relatively easily. Thus, proficiency levels play an important role in the use of connectors by ESL learners in their academic writing.

The concordance analysis examining the functions of connectors' expressions in ESL learners' sub-corpora provided additional detail. Among the connector expressions identified in ESL learners, the listing expressions (*first of all, in addition to*) had the highest keyness values across all levels. These connectors use to contrast or insert arguments into their discussion. The predominance use of listing connectors seems logical if we consider chronological progression, which is typically adopted in argumentative essays. The increased use of listing connectors was observed in many previous studies, implying that learners engaged in explaining and enumerating previously discussed content (Lee, 2004; e.g., Lei, 2012; Park, 2013; Kim, 2019). For example, Lei (2012) study revealed that Chinese learners, unlike native English speakers, favoured using listing and contrasting connectors to build their argument. Crewe (1990), also argues that this could be attributed to the fact that learners were utilising connectors in an attempt to generate a coherent feel to their writing.

A possible reason for the increased use of certain connector types is, as Hasselgård (2019) described, in an L2 we "regularly clutch for the words we feel safe with: our 'lexical teddy bears'" (p. 237). An inspection of the keybundles lists showed a shared bundle in two different connector types across the levels, with an increase at C1 level. This means that the "lexical teddy bears" of B1 and B2 levels become more favoured by C1. The evidence collected demonstrating the overuse of certain connectors by ESL learners reflects the work of Youngdong (2020), which showed that ESL learners use connectors as grammatical markers, and tend to use items they are familiar with in their writing. As also noted by Milton (1998), ESL learners prefer to use connectors they feel familiar with and know they can use accurately.

Another interpretation of the increased use of connectors in ESL learners' essays corresponds with evidence presented by Halliday et al. (2014), suggesting that in English, the overuse or underuse of specific connectors is one of the principal variables of English discourse. Therefore, in a piece of writing, the presence or absence of various connector devices do not augment the sense of the text, rather it is the appropriate use of these connector markers. Thus, in this regard, a degree of argumentative essays is also one aspect to take into account in this study, requiring ESL learners to concisely argue and defend a point of view about a topic by affirming their argumentation and providing supporting evidence. This writing genre requires effective research, good organisation and the inclusion of logical connections between sentences and paragraphs. Therefore, extensive use of connectives that tightly link one sentence to another and carry the argument forward ensures writers produce a more structured and harmonious text. Exploring the sub-corpora confirmed that most of the key connectors from Table 5.34 are more prevalent and frequent in higher learners' levels. Examination of the individual sub-corpora suggests that this overall increase in employing these connectors is due to an increase in the number of learners who used these expressions in their writing in the way of achieving cohesion in text, meaning that progression of the use of connectors correlates to language proficiency.

At the same time, this assumption could lie in the effect of process-based or argumentative essays taught in language centres, with teaching of certain connectors in an attempt to make students aware of the effectiveness of these expressions in presenting and organizing their arguments. Using such a method, it is assumed that students acquire the necessary knowledge of textual features of writing, to produce a successful text (Archibald and Jeffery, 2000). Firoozjahantigh et al. (2021) examined the effect of process-based instruction of writing on the IELTS writing Task Two performance of Iranian EFL learners. They argue that this method has been proven to be significantly helpful in developing L2 learners' academic writing skills. Therefore, the increased use of connectors expressions across the levels could be attributed to this method., as it is likely to influence students toward the use of certain connectors in their writing to achieve cohesion in text.

To sum up, the result of the connectors analysis revealed that ESL learners tended to overuse certain connectors significantly including *to sum up* and *first of all* in their argumentative essays. It suggested that ESL learners used these connectors quite often to make reader easily follow their arguments. Furthermore, the finding of this analysis revealed an increase in the prevalent and frequent use of connectors across the levels. Therefore, the results indicate a potential relationship between students' levels and the frequency with which connectors are used. This conclusion can also be proven by findings reported by Milton (1998); Leńko-Szymańska (2008), who observed the increased use of connector expressions in L2 learners' writing.

In order to achieve academic writing competence, learners must not only heed the organisation of their ideas in the text, but also the coherent construction of sentences and paragraphs. According to Oshima and Hogue (2007), a coherent paragraph flows smoothly from beginning to end. Three ways to achieve paragraph coherence are using nouns and pronouns consistently throughout a paragraph, using transition signals to show the relationships among ideas, and setting ideas into a logical order, such as logical division. In the present study, the keyness analysis findings showed that the ESL learners used connectors and first-person pronouns significantly helping to achieve text coherence, which is an important characteristic of the argumentative essay. However, the study noted an improvement in the use of connectors and first-person pronouns expressions across the levels, suggesting a positive relationship between the use of these expressions and learners' levels. The findings presented in this chapter revealed that, as the proficiency level increase, ESL learners become confident to use personal pronouns and connectors expressions as they believed that these expressions are showed authority and achieve text coherence. At the same time, language teachers must also raise their students' awareness of the dangers of the overuse of these expressions, once they understand their functions, since such overuse can reduce their strategic force in academic writing. The key here is to use these expressions sparingly.

5.5 Longitudinal study

5.5.1 Frequency distribution of B1, B2 and C1 sup-corpora

This section addresses the final research question, 'To what extent does an increased use of LBs correlate with learners' level of proficiency?'. This chapter investigated the development in the use of LBs across proficiency levels (B1, B2, and C1). For the purpose of this analysis, nine ESL learners were tracked over six months, producing three ESL learners' sub-corpora: B1, B2, and C1. These bundles were constituted approximately 20,000 words in each (see section 4.5.4). The study utilised a corpus-based approach to investigate the most frequently occurring three- and four-word LBs used by these ESL writers in their argumentative essays.

As discussed in section 4.8, *WST* was used to provide lists of the most frequent three- and four-word LBs in ESL learners sub-corpora. In order to extract the target bundles, the cut-off point was set at 40 times per million words (four times per 100,000 words) with at least three different texts (see section 4.6). The LBs that met the study's cut-off criteria were analysed in terms of their frequency, structures (Biber et al., 1999), and functions (Hyland, 2008b). Each sub-corpus was analysed separately, and the lists of the most frequent bundles were then compared to determine whether the use of LBs correlated with the level of proficiency. The results are shown in the form of descriptive statistics, as presented in Table 5.35 and Figure 5.10. The columns in Table 5.35 denote 1) the ESL learners' sub-corpora, according to their CEFR level; 2) the number of words in each sub-corpus, as computed by the *WST*; 3) the number of bundle types extracted from each sub-corpus; 4) the percentage of the total bundles in each sub-corpus, based on the frequency count; and 5) the normalised frequency.

Table 5.35. The target bundles extracted from the ESL learners' sub-corpora (longitudinal data). (% = Percentage of total bundles in the sub-corpus; Freq = Normalised frequency)

Length/Corpus	**No. of word**	**bundle type**	**%**	**Freq***
3-word B1	21,873	132	8	2501
4-word B1		21	2	434
3-word B2	18899	135	9	3159
4-word B2		29	3	651
3-word C1	19955	149	10	3473
4-word C1		39	3	741

* T he normalised frequency per 100,000 words, calculated as (absolute frequency/total number of words*100,000).

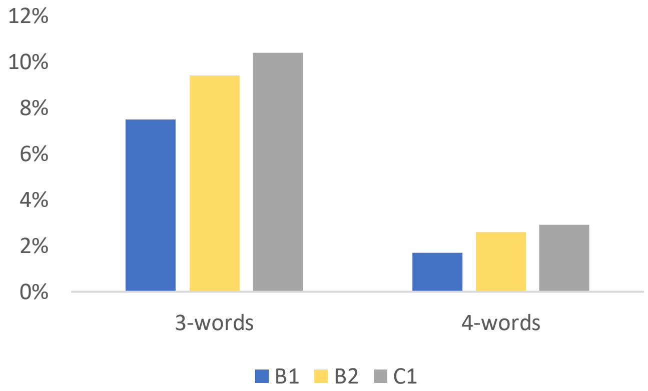


Figure 5.10. Percentage of the usage frequency of the target bundles in the three sub-corpora.

The results revealed an increased frequency in the type and token of the target LBs across the levels. In terms of the total number of different bundles used, there was a gradual increase in the use of the target LBs from 153 at B1, 164 at B2, to 188 at C1 level, while the tokens also showed an increase in the total bundles from 2,935 at B1, 3,810 at B2, and 4,215 at C1.

The degree to which the difference between the sub-corpora was significant was calculated using the chi-squared test to determine whether or not the distribution observed was due to chance, and to give a better estimation of the keyness of the keybundles, as shown below.

Table 5.36. Chi-squared analysis of the difference between the sub-corpora.

	B1	**B2**	**C1**
Observed value			
• 3-words	2501	3159	3473
• 4-words	434	651	741
Expected value			
• 3-words	2446	3175	3512
• 4-words	489	635	702
Chi-square P < 0.05	Chi-Square = 10.5063, df = 2, P-value = 0. 005231		

The results of the chi-squared test indicated a significant difference in the use of the target bundles across the CEFR levels, with a chi-squared value of 10.5063 and df 2, with a significant *P*-value at 0.005231.

The findings supported the argument that there was a direct proportionality between the use of LBs and language competence, and was consistent with the assumption that the use of LBs increases over proficiency levels, since the difference in their use between the ESL learner levels was significant. The result also addressed the issue discussed by Chen and Baker (2010, p.44), concerning "whether there is a relationship between proficiency and the number of formulaic expressions used".

While their research was not conclusive regarding a connection between LB usage and student performance, the findings of the present study suggested that the increased use of LBs may play a significant role in relation to student's level, as there was a clear dramatic increase in the use of LBs from B1 and B2 to C1 levels.

In order to present an overview of the commonalities and the development in the use of LBs across the three sub-corpora, the top 20 three- and four-word LBs in the B1, B2, and C1 sub-corpora are shown in Table 5.37 and Table 5.38, along with their frequency per 100,000 words.

Table 5.37. The 20 most frequent three-word bundles in the B1, B2, and C1 sub-corpora. (Freq = Frequency), Italic = shared bundles

B1	**Freq**	**B2**	**Freq**	**C1**	**Freq**
in conclusion I	82	as a result	85	*on the other*	110
across the world	64	*one of the*	79	*most of the*	75
it is common	59	in order to	74	in my opinion	70
a lot of	55	a lot of	63	in order to	65
I believe that	41	*on the other*	63	it is not	55
one of the	41	in my opinion	58	*one of the*	55
they want to	41	the fact that	53	*it is a*	50
that in the	37	I believe that	53	to sum up	50
in the past	37	*it is a*	48	in the world	45
field of opportunities	32	*most of the*	48	the development of	45
it comes to	32	*the number of*	48	to increase their	45
it is a	32	as well as	42	in the future	40
it is not	32	due to the	42	to say that	40
most of the	32	the learning process	42	and they are	35
on the other	32	to sum up	37	around the world	35
the number of	32	in other words	37	compared to the	35
the quality of	32	more and more	37	in conclusion I	35
to say that	32	there is a	37	in terms of	35
and so on	27	according to the	32	learn how to	35
because they are	27	because of the	32	*the number of*	35

Table 5.38. The 20 most frequent four-word bundles in the B1, B2, and C1 sub-corpora. (Freq = Frequency), Italic = shared bundles

B1	Freq	B2	Freq	C1	Freq
In conclusion I think	55	*on the other hand*	63	*on the other hand*	100
is one of the	32	in the learning process	32	is one of the	30
on the other hand	32	on the one hand	32	I would like to	25
when it comes to	32	as a result of	26	it is important to	25
huge field of opportunities	23	it is necessary to	26	as long as we	20
I would like to	23	it is true that	26	I do not agree	20
in the near future	23	one of the most	26	in this essay I	20
my point of view	23	it is argued that	21	at the same time	15
the main reason for	23	a lot to do	16	believe that it is	15
it is true that	18	as a result the	16	does not mean that	15
a vital role in	14	become more and more	16	have the opportunity to	15
all the possibilities for	14	can be seen as	16	in conclusion I prefer	15
first of all a	14	I believe that the	16	it is clear that	15
I agree with the	14	in addition to this	16	it is not the	15
I believe that in	14	it is clear that	16	my point of view	15
I strongly believe that	14	it is easier to	16	on the one hand	15
in this type of	14	play an important role	16	some other sets of	15
it would be better	14	this essay will discuss	16	that this is a	15
most of the time	14	this is because many	16	to the development of	15
the huge part of	14	to be more specific	16	some people think that	15

As Table 5.37 and Table 5.38 show, the usage of 20 most frequent three-word LBs ranged between 27-82/100,000 times at B1 level, 32-85/100,000 times at B2 and 35-110/100,000 times at C1 level. Meanwhile the overall frequency four-word bundles ranged between 14-55/100,000 times at B1, 16-63/100,000 times at B2 and 15-100,000 times at C1. It can be seen that the overall frequency of three- and four-word LBs increased across the levels, there was a wide difference between the learners at 55 in the B1 sub-corpus, 63 in the B2 sub-corpus, and 100 at C1 level. Hence, a narrower type and token of LBs was used by the learners at B1 level, with an increase in the pattern of average token frequency of LB types as the students progressed to higher levels of study. Again, strong evidence of the increased use of LBs was identified, as the C1 writers employed more bundles than the other writers. This may infer that higher ESL writers show a higher preference for the use of LBs in their academic writing.

When comparing the LB types across the three lists, it was apparent that five three-word and one four-word LBs were shared across the three sub-corpora, to an increased degree at B2 and C1 levels, with the exception of the bundles *one of the* and *the number of* that showed a lower frequency in the C1 sub-corpora than in the B2 sub-corpora. The significant increased use of LBs at B2 level indicated a key effect of the level on LB usage frequency, suggesting that proficiency levels had a significant effect on the usage frequency of LBs in ESL learners' academic writing.

The frequency of the bundles shared by all the three sub-corpora was cross-checked against the results from the cross-sectional study to determine the extent to which the ESL learners used these bundles (Table 5.39). Interestingly, the three-word bundles *on the other* and *it is a*, and the four-word bundle *on the other hand* were shared across the two studies, indicating that the ESL learners favoured these three bundles, with a high normalised frequency across the levels.

Table 5.39. Comparison of the shared bundles in the cross-sectional and longitudinal studies.

Bundle	**B1**	**B2**	**C1**	**B1**	**B2**	**C1**
	Cross-sectional			**Longitudinal**		
On the other	62	84	93	32	63	110
It is a	32	28	113	34	48	50
On the other hand	60	78	80	32	63	100

To conclude, the comparison of the frequency of the three- and four-word LB types and tokens used by the ESL learners showed a notable increase in their overall frequency and prevalence across the levels, with the writers at higher levels exhibiting significantly higher usage frequencies of both type and token over time. This finding is consistent with that of Chen and Baker (2016), Qin (2014), and Vo (2016), who found that the use of LBs increases across CEFR levels. The structural and functional analysis of the LBs is examined in the next section to identify evidence for this claim.

5.5.2 Structural analysis of the target bundles

The structural distribution of the LBs in this study was conducted in only one way, namely via the token distribution of the target bundles. The bundle types were discarded, as some of the sub-categories in each sub-corpus failed to represent a sufficient amount of data required for a chi-squared test. Following the cross-sectional study design, the LBs were classified into four main sub-categories (noun-based, preposition-based, verb-based and other), and 12-sub-categories. Some bundles did not fall into these sub-categories, and were allocated to the 'other' category, which was excluded from the analysis, as some LBs were not present in sufficient numbers for a chi-squared test. The results obtained from the preliminary analysis of the frequency and relative proportion of the structural distribution in the four sub-corpora is shown in Table 5.40 and Figure 5.11.

Table 5.40. Distribution of target bundles according to structural taxonomy. (Freq = Frequency; %= percentage within-sub-corpus)

	B1	B2	C1
Structural types	**Freq (%)**	**Freq (%)**	**Freq (%)**
Noun-based	668 (23)	836 (22)	817 (26)
Preposition-based	736 (25)	1085 (28)	912 (22)
Verb-based	1290 (44)	1482 (39)	1684 (40)

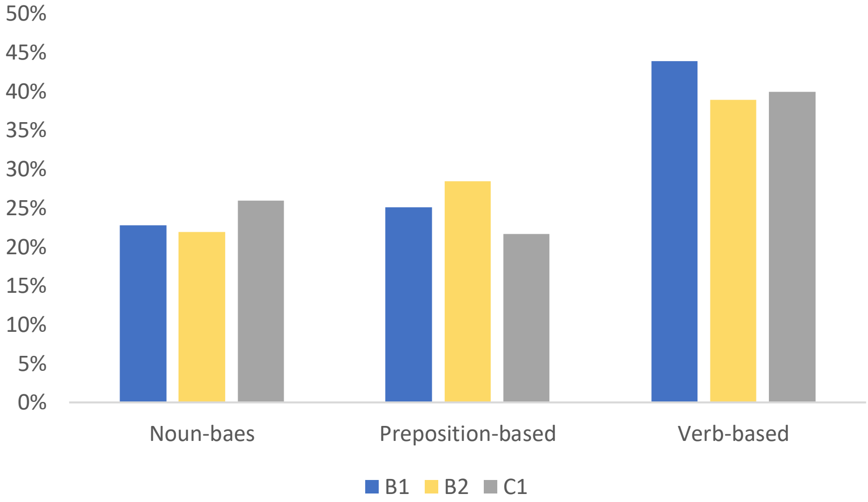


Figure 5.11. Overall distribution of the structural types across the sub-corpora.

It can be seen that the LBs in the ESL learners' sub-corpora (B1, B2, and C1) employed all the three main structural categories in the taxonomy. Verb-based bundles, such as *not be able to* and *are based on*, was the foremost category in the ESL learners' sub-corpora, which more commonly used in conversation register (Biber et al., 2004; Altenberg, 1998).

When comparing the development in the use of LBs across the levels, it was apparent that there was a significant effect at C1 level on the frequency of noun-based bundles (e.g., *one of the most*), indicating that the usage frequency in the ESL learners' argumentative essays increased by the C1 level from the B1 and B2 levels. Meanwhile, at the B1 and B2 levels, there was no significant change in the frequency of noun-based bundles, suggesting that proficiency level had a significant effect on the use of LBs by the ESL learner advanced level (C1). Another interaction that was observed was the frequency of verb-based bundles, such as *I would like*, which showed a significant decrease at B2 level, and a slight increase at C1 level, indicating that proficiency level had an effect on the use of these bundles by ESL learners in their argumentative essays. Meanwhile, the learners used more preposition-based bundles at B2 level, which increased significantly in B2 level, but decreased rapidly at C1 level.

In order to provide statistical evidence for the significant change in the structural categories between the sub-corpora, it was necessary to identify the differences by calculating the means of the chi-squared test. As in the case of the cross-sectional study, this test was used to evaluate whether or not the differences between the sub-corpora, in terms of the distribution of the structural tokens, were random. Table 5.41 shows the results of the chi-squared test (and standardised residuals), conducted on the use of the three main categories: noun-based, preposition-based, and verb-based LBs, across the ESL learners' sub-corpora, and compared with the RC.

Table 5.41. Standardised residuals (R) in a chi-squared test for structural distribution. (italic = significant interaction)

2.46146E-12	**χ2 = 60.33; df=6, p < 0.05.**		
	B1	B2	C1
Noun-based			
• **Count**	668	836	1092
• **Expected**	792	948	939
• **R**	-2.1	-2.8	*4.6*
Preposition-Based			
• **Count**	736	1085	912
• **Expected**	745	891	883
• **R**	0.7	*5.6*	*-4.7*
Verb-based			
• **Count**	1290	1482	1684
• **Expected**	1297	1551	1536
• **R**	2.5	-2.4	0.1

If the residual is less than -2.99, the cell's observed frequency is less than the expected frequency. Greater than 2.99 and the observed frequency is greater than the expected frequency

The table above revealed an extremely significant difference among the three sub-corpora, with a chi-squared value of 60 and a df of 6, which far exceeded the value required for the highest significant *P*-value at 0.0001. There was, therefore, significant differences between the sub-corpora, in terms of the three structural categories. Further analysis using the standardised residuals (R) revealed that the ESL learners' levels were positively correlated, as the learners' essays included consistent use of noun-based bundles at B1 and B2 levels, with a significant increase in their usage frequency at C1 level over time. Similarly, the learners' use of preposition-based bundles increased significantly at B2 level, indicating that they increased their use of these bundles at an earlier level than they did noun-based bundles. In contrast, although the verb-based bundles showed an increased use at B1 level, they were not found to be significantly overused over time.

Interestingly, all three structural categories contributed to the difference between the ESL learners' sub-corpora, although some of the results did not reach the required threshold of ±2.9. It can be argued that the results of the normalised frequencies demonstrated that there were significant differences in the way in which all grammatical forms within the structural types were used in the ESL learners' essays.

An overview of the differences between the sub-corpora considered the sub-categories of each structural type. Table 5.42 displays the normalised frequency and proportion of the noun-based bundles across the levels.

Table 5.42. Distribution of the Noun-based LBs across the levels (Freq = Frequency. %= percentage of the total bundles in the sub-corpus.

Sub-corpus	**B1**	**B2**	**C1**
Sub-categories	**Freq (%)**	**Freq (%)**	**Freq (%)**
Noun phrase with other post-modifier fragment	174 (6)	228 (6)	276 (6)
Noun phrase with of-phrase fragment	494 (16)	608 (16)	814 (19)
Total Noun-based	668 (22)	836 (22)	1092 (25)

As the table above shows, there was an increase in the use of noun-based bundles, specifically in the use of the structural form 'noun phrase with other post-modifier fragments', with 16% at B1 and B2 levels, and 19% at C1 level. A detailed investigation of the analysis of the most frequent LBs in the ESL learners' sub-corpora revealed that there were two main three-word LBs that contributed to this increase, one referring to quantity (*most of the*), and the other to quality (*the development of*).

The bundles *most of the*, which increased in frequency across the levels, may have been used more because of a high tendency to employ formal quantifier bundles in argumentative writing at C1 level, as the following example illustrates:

- Research have found that *most of the* accidents are caused by inexperienced drivers. (C1, essay 21)

Although quantifiers are basic words, they are a statistically frequent category in argumentation that implicitly conveys rhetorical meaning. Since argumentative writing is required by most English proficiency tests, such as the International English Language Testing System (IELTS), test-takers should become familiar with persuasive kinds of rhetoric, and increase their awareness of the important features in written discourse. Meanwhile, the bundle *the development of* was used far more frequently at C1 level than at any other levels, with B1 level seeing no use of this bundle at all. Example of the use of the bundle the development of in the C1writing is displayed below.

- Therefore, some people think that the education system is the only important factor to *the development of* a country, and they may be right. (C1, essay 3)

This result suggested that NP with of-phrase fragments were more frequently used at advanced C1 learners' level. This may be due to the inherently complex structural form of these bundles that require writers to deliver their message or argument formally, a key feature of academic writing (Biber et al., 1999).

In terms of the preposition-based bundles, there was an increase in the frequency of these bundles in the ESL writers' essays from B1 to B2 level, although there was a decreasing trend in their use at C1 level, as shown in Table 5.43.

Table 5.43. Distribution of preposition-based bundles. (Freq =Frequency; % = Percentage of the total bundles in the sub-corpus)

Sub-corpus	**B1**	**B2**	**C1**
Sub-categories	**Freq (%)**	**Freq (%)**	**Freq (%)**
Prepositional phrase with embedded of-phrase	105 (4)	148 (4)	70 (2)
Other prepositional phrase expressions	631 (21)	868 (24)	742(20)
Total preposition-based	736 (25)	1000 (28)	812 (22)

The increase identified in the use of preposition-based bundles at B2 level was specifically in the use of the structural form 'Other prepositional phrase expressions' at 21% at B1 level, and 24% at B2 level. A detailed investigation of the analysis of the most frequently used LBs in the ESL learners' sub-corpora revealed that two main three-word LBs contributed to this increase, namely *as a result* and *in other words*. The latter was used more by the ESL learners at B2 level than B1 level. A possible explanation for the increased use of this bundle is that the ESL learners at B2 level preferred to take the reader's needs into consideration, and to employ a clear explanation for the preposition mentioned earlier, as shown in the example below:

- *In other words*, male violence against women is a characteristic of someone, who was treated badly neglected in their childhood (B2, essay5)

Another bundle that showed the same tendency was the connector *as a result* that is usually employed at the beginning of a sentence, as in the example below; the learners at B2 level used this more than those at B1 level. Arguably, the increased use of connectors by the ESL in their argumentative essays was due to the fact that they were expected to use linking expression and discourse marker to organise information, in order to structure their discourse logically. While some previous studies connected their findings to L1 influence (Granger and Tyson, 1996; Hinkel, 2002), attributed this to writers attempting to organise their ideas according to the argumentative essay's structure, as in the example below.

- *As a result*, the government would be able to provide more funds for more productive purposes, such as the development of infrastructure, industries, hospitals, and so on, which may help improve a country. (B2, essay11)

This section discussed the proportion of bundles across the structural sub-categories of the target bundles at B1, B2, and C1 levels, showing the variation in the use of LBs among the grammatical structures within the ESL learners' essays. The next section discusses the functional analysis of the bundles identified in ESL learners' sub-corpora.

5.5.3 Functional analysis of lexical bundles across the sub-corpora

A major finding of the longitudinal study was that as the level of study increased, the writers used both a greater number and a greater variety of LBs in their argumentative essays. Therefore, it was necessary to investigate the functional categories associated with this increase in the ESL learners' argumentative essays. Following the same procedure applied in the cross-sectional study (Section 5.3.3), Table 5.44 and Figure 5.12 present the normalised frequency of the functional types, across the sub-corpora.

Table 5.44. Overall type distribution of the functional types across the sub-corpora. (Freq =Frequency; % = Percentage of the total bundles in the sub-corpus

Functional categories	**B1**	**B2**	**C1**
	Freq (%)	**Freq (%)**	**Freq (%)**
Research-oriented bundles	1463 (50)	1757 (46)	1814 (43)
Text-oriented bundles	864 (29)	1259 (33)	1348 (33)
Participant-oriented bundles	608 (21)	794 (21)	1042 (25)
Total	2935	3810	4204

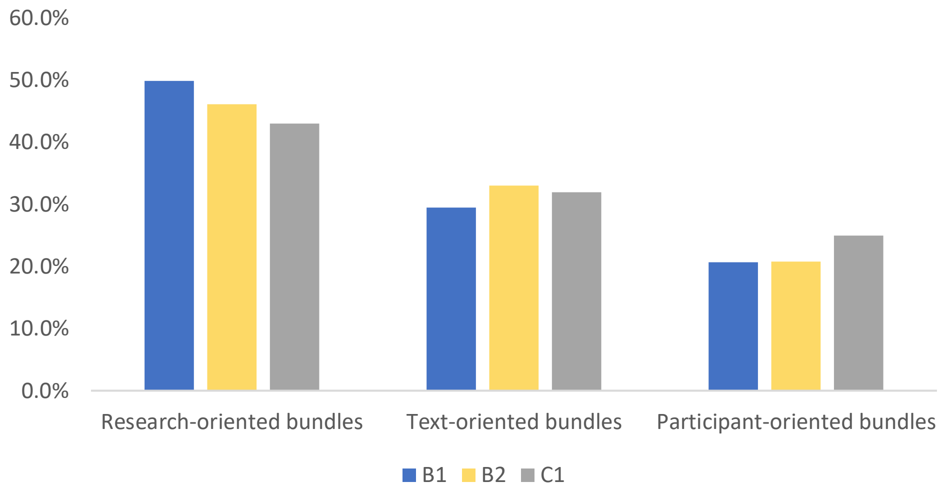


Figure 5.12. Overall type distribution of the target bundles, according to functional taxonomy, across the sub-corpora.

As shown in the table and figure above, there was not much difference across the levels among the functional categories. A comparison of each category's normalised frequency in the ESL learners' sub-corpora found that research-oriented bundles were the foremost category, accounting for 50% at B1 level, 46% at B2 level, and 43% at C1 level. This bundle type helps writers to organise their ideas and findings, and is comprised four sub-categories: location, procedure, quantification, and topic. This finding reflected the primary feature of academic writing, namely a focus on the subject of the research. Text-oriented bundles came second in terms of their usage frequency in the ESL learners' sub-corpora, with almost a third of the total bundles in each sub-corpus being of this type. This functional category helps writers to organise their text and its message, and includes transition (e.g., *on the other hand*); resultative (e.g., *as a result*); and framing (e.g., *the extent to which*) signals in the text. Meanwhile, participants-oriented bundles came last at around 20% of the total bundles across the sub-corpora.

A comparison of the increased use of the LBs across the levels revealed three interactions, the first of which was the frequency of text-oriented bundles, such as *on the other hand*, which increased at B2 level, as shown in Figure 5.12. There was also a significant interaction of C1 level and participant-oriented bundles, *such as I would like to*. In contrast, the use of research-oriented bundles decreased at B2 and C1 levels, compared with B1 level, indicating a negative effect, with a more frequent use of research-oriented bundles at B2 level that declined rapidly to C1 level. Meanwhile, the use of text-oriented bundles increased slightly at B2 level, and the use of participant-oriented bundles at the lower levels increased rapidly at C1 level. In order to provide statistical evidence of the significant differences in the functional categories between the sub-corpora, it was necessary to determine them by calculating the means of the chi-squared test, which was used to evaluate whether or not the differences between the sub-corpora, in terms of the distribution of the functional tokens, were random. Table 5.45 shows the results of the chi-squared test (and standardised residuals) conducted for their use in the three main categories, noun-based, preposition-based, and verb-based, across the ESL learners' sub-corpora.

The results of the chi-squared test indicated that there was a significant difference in the functional distribution of the target bundles between the ESL learners' sub-corpora, with a chi-squared value of 43.7 and df of 6, that far exceeded the value required for the highest significant *P*-value at 0.0001. The standardised residuals were then calculated to identify the cells that contributed to the differences. The highlighted cells of the R-value that exceeded ±2.9 indicated the significant difference among the functional categories.

Table 5.45. Standardised residuals (R) in a chi-squared test for functional distribution. (italic = significant interaction)

1.5E-08	**$\chi 2$ = 43.7; df=6, p < 0.05.**		
	B1	**B2**	**C1**
Research-oriented			
• Count	1463	1757	1814
• Expected	1339	1738	1918
• R	*4.1*	-0.4	*-3.2*
Text-oriented			
• Count	864	1259	1348
• Expected	925	1201	1325
• R	-2.1	*3*	2.8
Participant -oriented			
• Count	608	794	1042
• Expected	671	871	962
• R	-2.6	-2.8	*5.1*

If the residual is less than -2.9, the cell's observed frequency is less than the expected frequency. Greater than 2.9 and the observed frequency is greater than the expected frequency.

As shown in the table above, the ESL learner levels differed in terms of the proportion of the functional distribution of LBs employed. In terms of the research-oriented bundles, ESL learners at B1 level showed a significant increase in their use, while the C1 learners used significantly fewer of this bundle type. Meanwhile, the use of text-oriented bundles increased at B2 level, as did the use of participant-oriented bundles in ESL learners' writing at C1 level. The result revealed significant change across the levels over time, therefore further statistical analysis was undertaken regarding the functional sub-categories, as the variation of the main functional categories impacted the significant change over time.

Starting with research-oriented, Table 5.46 illustrates the proportions of sub-functions of research-oriented bundles in each sub-corpus, which can be used to show the changes of research-oriented bundles in ESL learners' writing over time.

Table 5.46. Distribution of the research-oriented bundles across the levels (Freq= Frequency; % = percentage of the total bundles in the sub-corpus)

Sub-categories	**B1**	**B2**	**C1**
	Freq (%)	**Freq (%)**	**Freq (%)**
Location	343 (12)	138 (4)	386 (9)
Procedure	288 (10)	640 (17)	531 (13)
Quantification	379 (13)	603 (16)	521 (12)
Description	453 (15)	376 (10)	176 (9)

We can see that the use of location/time (*in the past*) and description (*the quality of the*) bundles contributed to the significant increase in the ESL learners' writing at B1 level.

Qualitative analysis of the concordance lines of 'location bundles' shows that this group consisted exclusively of expressions indicating time (60%) and starting with prepositional phrase (*in the future, at the same time*). At the same time, the majority of description bundles were performed by noun phrase + of structures expression (*the importance of, the problems of the*). As stated by Biber and Gray (2010), Noun-based and preposition-based bundles are more common in academic writing. Although B1 writing is embedded with verb-based bundles, they are sharing characteristics of written production.

Turning to text-oriented bundles, there appears to be a boundary that differentiates between ESL learners' levels, as shown in Table 5.47.

Table 5.47. Distribution of the text-oriented bundles across the levels (Freq= Frequency; % = percentage of the total bundles in the sub-corpus)

Sub-categories/ Sub-corpus	**B1**	**B2**	**C1**
	Freq (%)	Freq (%)	Freq (%)
Transition signals	151 (5)	466 (12)	551 (13)
Resultative signals	357 (12)	439 (12)	286 (7)
Structuring	169 (6)	127 (3)	246 (6)
Framing	187(6)	228(6)	266(6)

It was somewhat surprising that even though the use of participant-oriented increases between B1 and B2 over time, only one sub-category does change that much.

Only the functional sub-category 'transition signals' was used more frequently at B2 than B1, and a detailed investigation revealed that the bundle *on the other hand* contributed to the significant increase in the ESL learners' writing at B2 level. This bundle is a type of connector expressions whose purpose is to contrast two different views of such a problem or an issue, and include various meanings, such as in contrast, alternatively, and conversely, as shown in the following example.

- *on the other hand,* people studying at the universities for getting knowledge and improving their job position and invent new things in that specific field. (B2, essay167)

The increased use of these expressions might be due to the organisational structure of the argumentative essay. In this writing genre, learners should heed not only in the organisation of ideas, but also their coherence. According to Oshima and Hogue (2007, p.79), a coherent paragraph flows smoothly from beginning to end, and can be achieved using three steps: "using nouns and pronouns consistently throughout a paragraph, using transition signals to show relationships between ideas, and arranging ideas into some kind of logical order". Therefore, connectors are an important tool for achieving paragraph coherence.

The third general category that showed an increase across the ESL levels was participant-oriented bundles, which focused on the writer or reader of the text, and were used mainly to show hedging, to express an attitude, to stress emphasis, and to indicate epistemic meaning. Table 5.48 presents the occurrence of the participant-oriented sub-categories across the levels.

Table 5.48. Distribution of the Participant-oriented bundles across the levels. (Freq= Frequency; % = percentage of the total bundles in the sub-corpus)

Sub-categories	**B1**	**B2**	**C1**
	Freq (%)	**Freq (%)**	**Freq (%)**
Stance expressions	581(20)	757(20)	997 (24)
Engagement	27 (1)	37 (1)	55 (1)

It is clear that there was increased use of the sub-function 'stance expressions' at C1 level, with the writers employing the bundles *it is not* and *it is a* frequently in their writing (see the example below), which is commonly used as an impersonal pronoun to refer to a thing that was previously discussed or identified in the text. The reason for the increased use of these expressions in the learners' essay maybe because of the key feature of argumentative writing that requires writers to express opinions, and to comment on and evaluate propositions, whilst allowing the writer to remain in the background.

- *It is a* key responsibility of the government to protect the environment by spending more money on research and modern equipment. (C1, essays14)
- Language is the best way of communication. Nowadays, *it is not* enough to be able to speak one language to communicate with the outside world. (C1, essay19).

The analysis employed by this study compared the use of LBs, in terms of their frequency, structure, and functions in the argumentative essays produced at three ESL learner levels (B1, B2, and C1). The detailed investigation revealed a number of distinctive features that differed according to the learners' proficiency level.

i. RQ4 discussion

RQ4 To what extent does an increased use of LBs correlate with learners' level of proficiency?

The fourth research question concerns of the changes in the usage frequency of LBs across the proficiency levels. The analysis in the longitudinal study found that different categories of LBs underwent distinct patterns of change across the proficiency levels. By conducting frequency, structural and functional analysis of the target LBs in the ESL learners' writing, it was evident that there were marked changes in how the LBs were used in the argumentative essays of the ESL learners at different proficiency levels. The differences between the three sub-corpora contributed to the understanding of how ESL learners employed LBs. According to the analysis presented in section 5.5, the use of LBs increased across the proficiency levels; as might be predicted, the C1 level writers made greater use of extended three- and four-word LBs in their argumentative essays than the other levels.

In terms of the frequency effect of the use of LBs across the proficiency levels, the study revealed a notable increase in the frequency of LBs at B2 level. The significant increase of LBs at B2 level indicates that their frequency increased at B2 compared to B1 over time. From B2 to C1, there was also a significant change in the frequency occurrences of LBs over time. These findings support the argument that there is a positive relationship between the use of LBs and proficiency level over time. Overall, the results of the present study suggest that, despite LBs being seemingly difficult to acquire (Liu, 2012; Gil and Caro, 2019), six months might be sufficient time for novice writers to gradually begin to produce more LBs and more professional output. These results appear to be consistent with (Crossley and Salsbury, 2011), who analysed the development of LBs in the spoken language of ESL learners to examine how the frequency of LBs changed over the course of one year. The result suggested that LB accuracy increased as a result of time spent learning English in an L2 native country, and thus as a function of increased English language proficiency. The findings of this thesis complement those of (Crossley and Salsbury, 2011) and increase our understanding of the use of LBs in a number of important ways. First, while Crossley and Salsbury (2011) found that the use of LBs developed over time in L2 spoken production, the current study found an increase in the use of LBs across the proficiency levels of written production over time. A strong argument can therefore be put forward for the developmental mechanisms involved in L2 spoken discourse (informal register) and L2 written discourse (formal register) of LBs. Moreover, while (Crossley and Salsbury, 2011) confirmed that L2 learners developed in the accuracy of LBs over a period of one year, this thesis observed a significant increase in LB production over six months. This suggests that ESL learners' knowledge of LBs can improve over a relatively short period of time. However, even though the use of LBs increased across the proficiency, there were some variations on their grammatical distribution.

Assigning functional distribution to each of the three- and four-word LBs identified in the ESL learners' sub-corpora using Hyland (2008b) functional taxonomy showed a significant change of the functional distribution of LBs among the levels over time. In terms of the research-oriented bundles, the results showed that ESL learners are mastering these expressions at an early stage (B1 level). This indicates the ESL learners provide the motivations of their research by highlighting how important the topic is. However, there was a significant decrease of these bundles at advanced level C1 writing, instead, there was a significant increase of participant-oriented bundles at C1 level. It is clear that ESL learners are shifting from using bundles that are focusing on a research topic to use more bundles that involve the writer's personal voice. Similarly, a significant interaction between the frequency of LBs in the ESL learners' sub-corpora and text-oriented showed that text-oriented that occurred more frequently in the ESL learners' sub-corpora were used at B2 level.

Meanwhile, regarding structural distribution, each of the three- and four-word LBs identified in the ESL learners' sub-corpora using Biber et al. (1999) structural taxonomy. The results showed that ESL learners used significantly more noun-based LBs in their essays at C1 level, which is inconsistent with the findings of (Chen and Baker, 2010). Conversely, an in-depth analysis revealed that the increase in the range of noun-based bundles at C1 level tended to be due to the greater use of quantifier expressions, focusing on several referential bundles. Previous studies have claimed that these bundles are more often used in conversation, especially the nominal phrase with the informal marker *a lot of* (increasingly used at B1and B2 level), and suggested that such use of conversation-type bundles is a feature unique to L2 learner writing (Staples et al., 2013; Chen and Baker, 2016; Bychkovska and Lee, 2017). At the same time, a manual check of the concordance line of noun-based bundles across the sub-corpora showed a rapid decrease in the informal quantifier *a lot of* at 63 in B2 and 25 at C1. Instead, the ESL learners at C1 used a more formal quantifier in delivering their argument (e.g., *most of the, one of the, the number of*). Therefore, the change from using more formal quantifiers than informal quantifiers at C1 level suggests that ESL learners' writing improves over time. This reflects a strong tendency among ESL learners to support their arguments by drawing on the concept of quantity. In other words, whilst a quantifier is vague and used more often in spoken language, using a noticeable quantity of something is a dominant stylistic strategy in EFL learners' argumentation.

Turning to the preposition-based bundles, the results revealed a significant increase in these bundles at B2 level over time, with learners achieving this increased use at an earlier level than they did with noun-based bundles. However, the use of these expressions decreased significantly at C1 level, indicating that ESL learners are shifting from more preposition-based bundles to noun-based bundles in their writing. At the same time, the increase in preposition-based bundles at B2 level was specifically traced back to the increase of connector expressions in learners' essays at B2 level. This result confirms previous findings and provides additional evidence that connectors are increasingly used in ESL learners' academic writing. As discussed in section 5.3.1, using connectors in argumentative essays is a key to achieve paragraph coherence, and is one of the most important factors in the assessment of second language writing (Rachmawati and Susanti, 2016). This is expected as these expressions are useful for showing the relationship between a piece of information and the writer's point of view, or for developing the writer's arguments.

Meanwhile, although the underuse of verb-based bundles in the B2 level essays was not significant, the findings of this study also show that verb-based structure had a negative effect on LBs use over time, which suggests that the use of verb-based structures decreased overtime. It is apparent from the results that the significant increase in the use of noun-based bundles at C1 level indicates an improvement in ESL learners' academic writing skills, since Biber et al. (2004) found that academic writing in English relies on noun phrases. This finding also supported the usage-based framework of language learning (Ellis et al., 2016; Crossley et al., 2019), as high-level ESL writers in this study were close to achieving the distributional characteristics of LBs in academic writing.

The structural and functional analysis revealed variations across the levels in the use of LBs of initial status, and also a change across the proficiency levels. Additionally, a positive correlation was observed, as the essays that contained one category of LB less frequently included more of these LBs, and vice versa, over time. For example, the ESL learners used noun-based LBs less frequently at B1 level, but used them increasingly over time. This highlighted the homogeneity of the writers in this study, in terms of the rate of change in the frequency of LB usage, since the ESL writers constituted a single cohort undertaking the same language course.

The significant increase in the frequency of noun-based, preposition-based, text-oriented, and participant-oriented bundles was evidence that "beginner learners' collocational knowledge can improve over a relatively short period of time" (Siyanova-Chanturia, 2015, p.158). This contradicted the findings revealed that L2 learners use fewer frequent collocations over time (Bestgen and Granger, 2014; Siyanova-Chanturia and Spina, 2020). These contradictory findings in research may be due to the different methodology employed, and to the different nature of the participants and their characteristics. It should be noted that the findings of the frequency occurrence of LBs in this study were not directly comparable with the other studies discussed, due to the differing variables concerned, such as proficiency level, the participants' L1 background, the research methods, and the writing style, therefore comparisons should only be made with caution.

Reflecting the observations of Chen and Baker (2016), this study confirmed that the ESL learners at B2 level exhibited significant development over time, whereby their writing neared the distributional characteristics of LBs in English academic prose, as the B2 bundles contained as many bundle characteristics of written discourse. According to the CEFR guidelines, B2 writers "Can use a variety of linking words efficiently to mark clearly the relationships between ideas" (Europe, 2020). In the present study, the increased use of linking words was evident in the increased use of participant-oriented and preposition-based bundles, a feature that was not present in the B1 level writing.

In summary, in terms of the fourth research question regarding the change in the usage frequency of LBs across proficiency levels, this study found the following:

- The different ESL learner levels exhibited a distinct pattern of change across the levels;
- There was a significant increase in the usage frequency of noun-based bundles in the ESL learners' writing at C1 level;
- B2 level was arguably the level that shows transition in the use of LBs, as sharing as many characteristics as possible of written production.

229 Conclusion

6.1 Main findings

This thesis focused on lexical bundles (LBs) as a target linguistic feature, motivated by the significant role they play in fluent linguistic production and "a key distinguishing feature of particular modes, registers and genres" (Hyland and Jiang, 2018, p.1). This study employed a corpus-based approach to determine the extent to which LBs provide cohesion, by revealing how the building blocks used in discourse are used differently in the academic writing produced by L2 learners of varying proficiency levels. This study fulfils two major research objectives. The primary objective is to investigate the use of LBs within academic writing at three CEFR levels: B1, B2 and C1, to produce empirical data concerning possible variations in bundles identified at three ESL learners' levels. That will be useful to understand language variation specifically in English language learners (rather than university students or expert writers) academic writing. The second objective is to track the developmental use of LBs in argumentative essays by ESL learners over time at three CEFR levels, so as to provide empirical data to measure the relationship between LBs and language proficiency levels.

To achieve these objectives, the study identified and compared the frequency, structure and discourse function of the most frequently LBs in argumentative essays of three ESL learners' levels, namely B1, B2 and C1. The need to conduct research into academic writing was initially motivated by the requirement to understand the use of LBs in argumentative essays, especially the bundles that differ between ESL learners' levels, in order to facilitate their use by teachers and learners. Based on the findings of this thesis, therefore, some language features can be identified as providing insights into future teaching practices regarding LBs.

This research has provided empirical corpus-based evidence on how ESL learners' writers used LBs in terms of frequency, structures, and functions at different proficiency levels. Based on the findings of this thesis, specific features in the aspect of academic writing in ESL learners B1, B2 and C1, as well as learners' common characteristics regardless of proficiency have been identified. Below I summarise the findings for the research questions in order to have an overview of the study.

In relation to the frequency of LBs used at each of the three ESL learners' levels, the following conclusions were drawn. First, the most frequently used LBs in C1 writing was the bundle *I think that*, while the bundle *a lot of* was preferred in B1 and B2 writing. Second, C1 writers employed a larger number of LB types and tokens compared to B1 and B2 writers. This could result from C1 writers having a higher preference for LBs to build their arguments. This finding also begins to address questions posed by Li and Schmitt (2009, p98-99) concerning whether "the appropriate and diverse use of lexical phrases has an effect on the evaluation of academic writing". Therefore, apart from showing a higher frequency of LBs, the number of different LBs also increased at C1 level. We were able to see that C1 writers tend to use more LBs with greater varieties than B1 and B2 levels. This finding brings us to a tentative conclusion that the increased use of LBs in ESL learners in cross sectional-study only found at C1 writers. However, although ESL learners showed great varieties use of LBs, they used distinctive bundles rarely or never found in the expert writing, which indicates differences in their bundle use.

Regarding the variations in the structural use of LBs across sub-corpora, the analyses detected some similarities and differences across the levels (Section 5.3.1). ESL learners' levels showed some similar tendencies with regard to the construction of LBs. In view of this, the largest number of LBs across the levels were those constructed with verb-based bundles. These findings correspond to previous observational studies, which suggest that verb-based bundles are in the prominent structural category used by ESL learners across the levels (Chen and Baker, 2010; Chen and Baker, 2016; Ruan, 2017). Although verb-based bundles are used more widely in spoken language than in academic prose (Biber et al., 2004), it was also a distinctive feature in the proficient student writers writing in this study (BAWE). This suggests that the claim made by Biber et al. (1999)et al. (i.e., that verb bundles are more likely to be found in conversation than written discourse) is not applicable to every genre of writing. It is likely, then, associated with the writing genre, since they are increasingly used in all sub-corpora. In addition, sometimes, as the argumentative essay calls for this use, conversational bundles generally include a personal pronoun and active verb, and this structure is needed if, as in this case, students are required to show their viewpoint. Therefore, the students' reliance on this structural type might be a characteristic of LB use in ESL learners' argumentative essays.

At the same time, examining the structural subcategories also showed similarities in the use of LBs across the levels. This indicates that ESL learners prefer using relatively the same structures of LBs to deliver their message. The similarity between the ESL learners and the BAWE writers also found their preference to use noun-based bundles and preposition-based bundles sub-categories. However, the verb-based sub-categories showed apparent differences for containing a more formal writing style in BAWE writing than the ESL learners' writing, as ESL learners used the personal pronouns in their essays, as opposed to the passive voice. While 1st/2nd person pronoun + VP fragment bundles is likely to be the most frequent sub-category at C1 level and the Noun phrase with of-phrase fragment at the BAWE. A possible explanation of the similarity between the levels could be that LBs' structural distribution did not reflect the differences in language proficiency between ESL learners' levels. Therefore, what contributes to learners' level of proficiency is the frequency use of LBs rather than structures.

When the functions of the LBs are compared, ESL learners tend to use certain LBs serving functions that similar across the levels. It was found that LBs indicating procedure, quantity, and resultative signals (research-oriented bundles) seem to be predominant in ESL learners' writing, as compared to the other two major functions of these word combinations. These bundles were focused more on the external relations in the world describing time and place relations (location), size and quantity (quantification), the study itself (description) and research procedures. This finding accurately reflects the most outstanding features of academic writing; focusing on the subject of the research (e.g., Hyland, 2008b; Allen, 2009; Du, 2013). It seems that regardless of language proficiency, academic writers focus principally on facts and evidence relating to the essay topic. Given that the writing task used in this study requested that ESL learners write an argumentative style essay providing a convincing argument relating to the conclusions they had come to, greater use of research-oriented phrases might have signified that ESL learners' writing at all levels relied more heavily on facts and evidence to support their arguments, similar to the proficient students (BAWE sub-corpus).

A closer inspection of the functional sub-categories of the identified bundles revealed considerable discrepancies across the sub-corpora. For example, the analyses of the research-oriented sub-categories demonstrated a preponderance of informal quantifying bundles (e.g., *a lot of*) deviating from typical academic prose found in both B1 and B2 sub-corpora. These bundle patterns are common features in spoken discourse and show informal writing style. In contrast, C1 writers are “more mature academic writers” and utilised more procedural expressions, which not only shows greater language competence but also allows them to organise the discourse so that audiences have a clearer understanding of the text. These findings indicate that C1 writers are more likely to differentiate between formal and informal language, as they exhibited similar features of spoken and written production. Meanwhile, B1 and B2 writing exhibited more speech-like features than written ones. Therefore, these functional differences simply reflect the discrepancies in general English proficiency between ESL learners' levels.

In addition to the differences between the levels, this study found that ESL learners exhibited the same use of stance bundles, which convey a sense of certainty, uncertainty and desire. Although LBs articulating with stance expressions are less common in written language (Biber et al., 2004; Biber and Barbieri, 2007; Csomay, 2013), ESL learners increasingly used these expressions in their argumentative essays. As discussed earlier, stance bundles may be characteristics of argumentative essay writing in an L2 context. It is more likely that argumentative essays requiring the writers to express their opinion, causing a high proportion of stance expressions, such as *I believe that, I* think it is, and *it is important*, an assumption confirmed by the increased use of stance bundles in the ESL learners’ sub-corpora.

In relation to the grammatical variation in the use of LBs in ESL learners’ writing, it is also interesting to note that keyness analysis of the overall ESL learners and BAWE sub-corpora, with the latter as the reference corpus, revealed two distinguishing characteristics associated with LBs produced by ESL learners. Similar to the previous studies (Biber et al., 2004; Biber and Conard, 2005; Chen and Baker, 2010), LBs embedded with first person pronouns (e.g., *I think that, I think it is, I want to*) were overused across the ESL learners’ levels. The increased use of the first-person pronoun helps readers follow up the argument, and helps writers to build their identity by highlighting their voice when conveying the connotations of an argument. The results obtained from the keyness analysis showed that ESL learners at C1 level had a stronger and firmer authorial position than learners at other levels, as they made more frequent use of self-mentions (*I*) to present their viewpoint in their writing. Therefore, an understanding of the accurate use of first-person pronouns is of great value to ESL learners. They must know, in the process of writing an argumentative essay, how to articulate their personal view to their reader.

Another characteristic found in ESL learners' writing was connector bundles. The results are in line with previous research, which suggests that L2 learners are engaged in explaining and listing what was already mentioned in their writing (Granger, 2004; Lee, 2004; Chen and Baker, 2010; Park, 2013; Kim, 2019). These expressions are crucial in academic writing and one way to achieve paragraph coherence, which is one of the most important aspects in assessing second language writing. Referring to the CEFR scale, in order to achieve a B2 or B1 level, it is necessary to use a wide range of linking devices in writing correctly. Therefore, they are an important tool for achieving written proficiency in English. According to the literature, the use of connectors should somehow function as an indicator of the text's coherence. If such a relationship can be demonstrated, this will help improve the quality of the text. Overall, the analysis of the most common connector expressions in ESL learners' sub-corpora showed a positive increase across the proficiency levels.

Finally, and perhaps more importantly, time interacts with proficiency (CEFR levels) in its effect on the use of LBs. However, time affected ESL learners' proficiency levels differently. In general, whilst there was an increase in the frequent and mutually attracted LBs across all proficiency levels, the growth was by far strongest for B1 level compared to the other groups. The results of the longitudinal study revealed that the proportion of LBs used by ESL students were significant predictors of learners' writing proficiency, and increased significantly across the levels over time. It was found that ESL learners produced more frequent LBs as their levels increased. It appears that, following exposure to the L2 language, learners acquire more extensive LBs. This finding may be attributed to the inherent nature of LBs, as it is acknowledged that most LBs are semantically transparent (e.g., *it is important, as a result, a lot of, on the other hand*). They often consist of high-frequency words, which are likely to be known to ESL learners.

At the same time, it was surprising that even though the use of LBs increased across the levels, different categories of LBs underwent distinct patterns of changes over proficiency levels. By analysing the frequency, structure and function of the target LBs in ESL learners' writing, it is clear that there are indeed marked changes in how

LBs are used in the argumentative essays of ESL learners of different proficiency levels over time. LBs change from being characterised by informal verb-based structures to a more academic writing style with noun-based and prepositional-based bundles at higher proficiency levels over time. At the same time, there is a tendency toward LBs that convey the writer's attitudes and evaluations with the progress of writing proficiency over time, which is attributed to the argumentative writing genre.

According to the cross-sectional study findings, CEFR C1 level is the main stage at which ESL learners show signs of development and begin to be aware of the characteristics of academic writing, as the C1 sub-corpus contains a number of elements of formal writing style. Meanwhile, learners at the B1 and B2 levels have made neither a quantity gain nor a quality gain in terms of bundle use in their essays, which are clearly characterized by an informal style that represents the typically spoken register. However, according to the evidence from the longitudinal study, the CEFR-B2 level is the stage at which learners show an increase and development in the use of LBs over time, and begin to grasp the distinction between spoken and written production. It is apparent that time reflected positively on the development use of LBs in ESL writing, as bundles used at C1 level are highly characterised by the academic writing genre, whereas B1 bundles are more likely to be characterised by an informal style that represents spoken production.

6.2 Relating research findings to the CEFR

Examining the use of LBs in ESL learners' argumentative essays at the B1, B2 and C1 levels related/linked to the description of some CEFR can-do statements across the levels. For example, with reference to CEFR, C1 writers are described as ''Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say. The flexibility in style and tone is somewhat limited' Europe (2001, p. 187). As C1 writers found to use LBs extensively in terms of types and frequency in their argumentative essays, the qualitative components of this study are broadly consistent with the characteristics of C1 writing described in the CEFR above. For example, writers at higher levels (C1) used more LBs and had a more comprehensive range of grammar with improved accuracy.

The next statements are associated with the scale of the coherence, which reflects a clear distinction in the description of language development across the proficiency levels, as seen in the table below.

Table 229.1: CEFR Coherence descriptor

C1
Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.
B2
Can use a variety of linking words efficiently to mark clearly the relationships between ideas.
B1
Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.

As shown in the above table, whilst only B2 level learners were expected to use a variety of linking words (used as a connector expression in this thesis), the CEFR does not predict a greater density of connectors as used by more advanced learners. Based on the results of RQ3, a limited number of connectors are found at B1 and B2 levels, while ESL learners at C1 levels show greater use of a larger number of different connectors and to the extent that clearly distinguishes them from the lower levels. The results do, however show that learners use a wide range of different connectors later than predicted in the CEFR at B2 level. This may indicate that a revision of the CEFR-scale is needed at this point.

Although B1 and B2 learners tended to overuse certain connectors in their writing (e.g. *on the other hand, as a result*, *first of all*), this phenomenon, as discussed earlier, may be associated with the "lexical teddy bears" (Hasselgård, 2019). Common connectors give novice writers a degree of comfort, and therefore they view these as being suitable to use safely when required. It is also likely that learners' use of certain bundles is due in part to their using the same connectors in different rhetorical functions, as seen in the use of the bundle *on the other hand* by B1 writers. As discussed earlier, the use of explicit coherence phrases like connector expressions is only one way of building coherent texts. The increased use of connectors does not necessarily make a text coherent, and it is possible to make a coherent text with the controlled use of explicit markers of coherence relations, as stated in the CEFR description.

In addition to the above results, a B2 writer described as one who ‘can make distinctions between formal and informal language with occasional less appropriate expressions’, and their ‘language lacks, however, expressiveness and idiomaticity and use of more complex forms is still stereotypic’ (Europe, 2003, p.187). Based on the findings, the use of informal language manifested across the levels, as ESL learners shared certain common characteristics with the spoken register (e.g., the increased use of verb-based bundles).

It is apparent that comparing the use of LBs across proficiency levels in several areas of interest, including frequency, structure and functional distribution, attaches considerable importance to communicative needs and shows that the learners need to have a wide range of LBs to deal with simple needs in academic writing. However, the findings suggest that in the B1–C1 range at least, the CEFR scale is most generalizable across languages at C1. Thus, LBs seem to render differences that are sufficiently fine-grained to distinguish between the levels. This is probably because the development of LBs continues even at the highest levels, and possibly never ends, since it involves the growth of lexis, which is constant even in the L1.

6.3 Contributions of the study

An extensive literature review revealed that several prior studies focused on the use of LBs in argumentative essays written by ESL learners of different proficiency levels. Most of these previous studies (Cooper, 2013; Staples et al., 2013; Qin, 2014; Chen and Baker, 2016) focused their attention on assignments written by L2 learners, compiling data from ready-made corpora or academic English language tests provided by students from different study years at the university. Therefore, the results of the study with regard to comparisons of LBs, and performance development of ESL learners of different proficiency levels can be considered as a contribution to existing knowledge. In particular, the findings increase our awareness in relation to the similarities, differences and development of the LBs used by ESL learners at different levels. The LBs frequently identified can be used as model patterns for teaching and learning purposes.

The second contribution at a descriptive level relates to comparing ESL learners at B1, B2 and C1. The development of sub-corpora of ESL learners from different proficiency levels has meant that this study's findings can help address questions regarding variations between learners' levels in academic writing. For example, the use of connectors, personal pronouns and stance bundles support and builds on the results of previous studies (Staples et al., 2013; Chen and Baker, 2016). Furthermore, the investigation into differences in LBs helps to clarify the gap between ESL learners, thereby shedding light on what students require to become more proficient writers (Cortes, 2004; Jablonkai, 2009).

Another contribution concerns the acquisition of bundles and arises directly from the longitudinal nature of the second aim of this study. It seems evident from the findings of this study that there is a positive correlation between the use of LBs and proficiency levels, although only to a limited extent. This is important as it begins to address the questions posed by Biber (1990) as to whether bundles correlate to proficiency level.

6.4 Pedagogical contributions

The study has revealed three pedagogical contributions:

1. The identification of LBs in this study can contribute to existing knowledge of which type of LBs are commonly found at each level, so that the pattern can be applied when teaching and learning academic writing more generally. In addition, ESL learners' keybundles in relation to those of BAWE writers can be used to raise awareness of what words are not commonly found or which are deemed irrelevant to proficient writers.
2. Despite the prevalence of the argumentative essay type at university levels in ESL/EFL contexts, this study has contributed to the research on LBs and added to the growing body of knowledge concerning how bundles are used in argumentative essays by ESL learners with different proficiency levels.
3. These results also demonstrate how useful keyword analysis is for determining the characteristics of a particular language register. LB studies also help to identify genre features, thereby providing valuable information for L2 learners.

6.5 Implications

As the study focused on argumentative essays written by ESL learners, the current research results have implications for English for Academic Purposes (EAP). First and foremost, the study contributes to the understanding of the differences between writing produced by ESL learners at different proficiency levels, especially EAP learners. The analysis has confirmed some characteristics of ESL learner language that were already suspected in an L2 context, for instance the underuse of the passive voice across the levels, and disproved others, such as the overuse of first-person pronouns. Other aspects can only be uncovered through quantitative and qualitative analysis, for example, the fact that although ESL learners overuse participant-oriented bundles, they are only significantly used by C1 writers.

Moreover, this study has also suggested that these cases of overuse and underuse, regardless of their syntactic form, can be used as a reference in terms of the development of written competence. Improving students' awareness of the function of these bundles would equip them with better productive knowledge of LBs and enable them to use LBs carefully by paying more attention to common misuse and overuse; this would further improve their written proficiency. Future EAP pedagogical material should focus on these areas to help bridge the gap between what EAP learners are capable of and their intended proficiency.

In addition, because of the transparency of LBs in academic register, students should be taught these expressions explicitly in order to recognize these bundles and their functions. The current study found that the LBs used by ESL learners were phrasal rather than clausal bundles, which appeared to be similar to the bundles found in academic discourse. However, a qualitative analysis showed that C1 learners used bundles to structure discourse to a greater extent than B1 and B2 writers, who seemed to write more bundles in conversational rather than academic register and to use them in a more informal style, namely, to simply mark quantification rather than structure discourse. It appears that ESL learners need to be taught the functional use of these expressions if they are to be used accurately. As C1 writers who have been shown to be competent were more likely to use more and varied LBs, all students should be taught these more advanced uses of bundles as they appear to make a difference to ESL learners' writing skills.

Another point of concern is that my teaching experience suggests that students are rarely or never exposed to words or phrases in a corpus due to limited classroom time. Teachers need to find ways of exposing students to the most frequent bundles, rather than assuming that students will acquire these expressions independently, as students need multiple and various exposures to a word or phrase before they fully understand these items. Learners need to learn these expressions in context and not using stand-alone lists that come and go infrequently. Research has shown that highlighting important words for discussion in reading activities and practice is one way of providing the necessary input (Zimmerman, 1997; Horst and Meara, 1999; Appel, 2011b). For example, Appel (2011b) examined the use of LBs in university EAP samples. The study found that although less proficient writers produced more LBs than advanced level writers, a closer investigation showed that many of these expressions were sourced from the reading articles they were provided with before beginning their writing. Therefore, students should be frequently exposed to LBs and should be assigned writing tasks in which they are asked to somehow manipulate the LBs. Boers et al. (2006) suggested that highlighting formulaic sequences in teaching materials can improve overall language fluency for non-native English speakers.

Finally, because C1 writers use various and frequent bundles from the academic register, textbooks should have units that demonstrate how to structure an essay with bundles from academic writing, as both proficient ESL learners and proficient student writers use bundles in this way. In addition, it is easier to generalise what language is favoured by learners through keyword results with high log-likelihood scores and LBs that meet specific cut-off thresholds for the inclusion of LBs for widespread use. This study identified the most frequent three and four-word LBs in ESL learners' writing, which could indicate the most useful LBs for ESL students to learn. The discourse function of the identified bundles can also act as a guide for teachers on the context and purpose in which they can be used in written discourse. English language institutions should use this type of research as a resource to create a detailed style guide that could be referenced by teachers when assessing written discourse.

6.6 Limitation

During the course of this study, which began with a compilation, processing, and analysis of B1, B2 and C1 ESL learners' sub-corpora and ended with relating the results to the descriptors of written language competence in the Common European Framework of Reference, a wide range of decisions had to be made. These decisions have led to a focus on certain aspects while excluding others. This section will briefly describe the limitations of the central corpus study, starting with the methodological limitations and ending with the limitations in relation to the interpretation of the results which concern both the representativeness of the corpus and the statistical analysis, as discussed below.

1. Ideally, both cross-sectional and longitudinal data would be larger than they are at present. Although they were representative samples of the language varieties under investigation, some bundles could not be analysed statistically in more detail due to the small dataset. For instance, when analysing LBs structurally, the number of occurrences in each sub-category fell very quickly, presenting a problem with regard to an accurate analysis. With a limited sub-corpora size, caution must be applied, as the findings might not be transferable to other learner groups. To make the results of this study more accurate and generalizable, it might be preferable to examine a larger corpus.
2. In addition, I would have preferred to study balanced gender, age, or L1 background of ESL learners in order to investigate the possible effect of these factors on the results concerning the differences in the acquisition and application of LBs in academic writing. This was not possible due to time constraints and the challenges of offering an equal number of essays in order to build the corpus. This could be an interesting area for further investigation, which would allow for a better representation and could inform researchers as to whether gender plays a role in the use of LBs.
3. Scrutinizing the learners' written competence, as opposed to simply analysing learners' use of the LBs, for instance, meant that the net had to be widely cast, as it were. In other words, the study could not limit itself to the analysis of the frequency of LBs. For each bundle type, structure/function, and other discourse markers, choices had to be made regarding which LBs to include and exclude in the analysis. Inevitably, due to time constraints, not all LBs were included. A notable example is the decision to exclude spelling errors. As discussed in section 4.6.7, during the retyping process, some spelling mistakes were identified in the ESL learners' essays. These errors covered both minor and major mistakes. Since the essays were mostly not available as computer-read texts, which resulted in the

very difficult and time-consuming process of preparing texts for the concordancing program, it was possible that the essays might have contained some typographical errors. In addition, since the *WordSmith* tool would only extract recurrent multiword sequences which were identical in form (i.e., spelling), it was necessary to have standardized spelling as much as possible throughout the corpora. As spelling was not the focus of the current study, and to avoid the researcher having to use intuition, the decision was made to exclude the small number of spelling errors identified across the essays. Therefore, there are some inescapable limits to the authenticity of the data. However, the LBs that were included arguably cover such a wide range that they make it possible to perform a relevant analysis of the ESL learners in written discourse.

4. There is a possibility that some useful LBs have been excluded from the analysis, or that unnecessary items have been included, simply because the decision was based mainly on the researcher's judgment. Meanwhile, the functional and structural classification of LBs is based on human interpretation and therefore may not be entirely free from bias and/or subjectivity. This will therefore continue to be a limitation in this type of research until human judgment is removed entirely. One way of tackling this issue, thereby increasing reliability, would be to employ multiple reviewers in order to evaluate the appropriacy of selected bundles, and to categorize them structurally and functionally more reliably. However, due to various circumstances, this was not possible in the current study.
5. The large number of LBs identified, and their structural and functional sub-categorisations, included in the analysis meant that this thesis had to limit itself – for the most part – to quantitative analyses. Only significant results were analysed qualitatively in more detail. Of course, the analysis of the sub-corpora under investigation would have been even more in-depth had time and resources permitted to supplement the extensive quantitative analysis with an even more detailed qualitative one of all the identified bundles.
6. The learner corpus research described in this thesis has its object of study essays produced by ESL learners of English from six language centres in the UK. It is important to bear in mind which language variety a specialised corpus represents, and restrict claims based on research to that particular variety, and not make claims beyond it. This is associated with the limited research examining the language presented in ESL programs and language used in various academic disciplines, and

means that the results of this study are restricted to what was examined. For instance, findings from a C1 sub-corpus might tentatively be extended to all advanced C1 learners of English in the UK, but not to all advanced C1 learners of English globally. We have seen throughout this thesis that there are as many similarities as there are differences between various ESL learners and university students. The findings, therefore, provide a valuable basis for the comparison of ESL learners at B1, B2, and C1 levels.

7. Another limitation related to the reference corpus used in this study, and it could also be argued that as the BAWE corpus was compiled more than twenty years ago, there might be a temporal gap, and the language used may no longer accurately represent current usage. This viewpoint is supported by Hyland and Jiang (2018), who reported a considerable change in the functional distribution of LBs in response over time. Therefore, choosing the newest reference corpus could shed light on the accuracy of ESL learners' levels, though to the best of my knowledge, this is the only publicly available corpus that can be comparable with the target sub-corpora.

8. The final limitation is related to linking the results to the CEFR descriptors. The main focus of this study is examining the use of LBs for each of the CEFR levels, which will differentiate one level from adjacent ones. For example, the occurrence or not of LBs in a learner's output can diagnose their CEFR level and/or distinguish them from other learners whose use of the same bundles differs (significantly) from that of the first learners. However, the CEFR descriptions were not very explicit about the use of formulaic language, specifically LBs, for language competence, and the written descriptor scales take insufficient account of how variation in terms of this phenomenon may affect performances by raising or lowering the actual level; across all three levels there was only limited evidence of statements drawing on Can-Do objectives. In addition, there is no evidence for the CEFR scales that suggest learners at a certain level are able to perform tasks associated with the lower levels. The CEFR is criticized by second-language acquisition researchers in that a pathway of progression is assumed, even though specific stages have not been fully confirmed by research into the development of proficiency over time (Alderson, 2007; Hulstijn, 2007; Figueras, 2012).

To the best of my knowledge these points have not been investigated, and it is precisely this lack of empirical research that gives rise to increasing misgivings about the capability of the framework in its current format to measure the possibility of finding correlations between the development of formulaic language, particularly LBs, and the proposed language proficiency levels of the CEFR scale (Council of Europe, 2003). These points have an inherent need for Can-Do statements to be broad enough in terms of the acquisition of formulaic language to be adaptable to diverse learners' requirements.

Despite the aforementioned limitations, I have sufficient confidence to reassure that I took multiple decisions to ensure that this research is both authentic and trustworthy.

6.7 Future work

The outcomes of this thesis, as well as its limitations, open much scope for further research in a variety of related aspects.

1. The first extension of this research that comes to mind would be to replicate it using oral data from learners of different proficiency levels enrolled in EAP courses. By comparing the results with the current study, more insight might be gained into the characteristics of ESL learners' language. This in turn might allow more definite conclusions to be drawn regarding the use of LBs, and more accurate advice being given to ESL learners at each level.

2. There is an inherent difficulty attached to research investigating very specific corpora, in that availability of texts is harder attain. For instance, to build a corpus of IELTS task2 texts, a researcher needs the exact texts used and cannot replace with texts of perceived similar functions, since doing so would limit the applicability of the results. Despite the discussed hardship, employing larger corpora and use the full capacity of the BAWE as a RC would lead to more accurate analysis and more generalizable findings. For instance, some LBs could not be examined thoroughly due to their rarity combined with the size of the corpus. Using larger corpora might make it possible to investigate these relatively rare bundles, as well as the infrequently occurring categories (e.g., underuse of passive voice), with much more accuracy.

3. Further research should also be carried out into the differences between the LBs used by ESL learners and native speakers. This would help learners gain a command

of the register by identifying how to reduce the gap between student writing and that of native speakers.

4. Further research could analyse more positive and negative keybundles and LBs to better understand other characteristics of ESL learners' writing. Finally, research into the types of bundles used by ESL learners in their argumentative essays and textbooks in EAP courses should also be undertaken in order to examine the relationship between ESL learners' levels and their study materials.

To conclude, the current study showed that LBs are indeed an effective way for differentiating between learners' proficiency levels. Other than contradicting or reaffirming past findings, this study also extends beyond the descriptive analysis of learner writing to argue that learners' language is still influenced by specific features, such as genre writing. Thus, this thesis has made a significant contribution to the existing body of research on corpus linguistics and provided a range of new insights into ESL learners' language.

6.8 Concluding remarks

An important contribution this study has made to corpus linguistics is the development of cross-sectional and longitudinal corpora of ESL learners writing based argumentative essays. Analysis of these ESL learners' sub-corpora and, in conjunction with a corpus of university students writing from the same genre, has added to research findings on ESL students' understanding of the use of LBs. The finding of this study showed that, contrary to the similarities of the use of LBs between B1 and B2 levels, the variation uses of these expressions correlated to English language performance in the case of advanced C1 level. Another finding showed that the grammatical distribution of the target bundles used across the learners' sub-corpora generally presented the same picture; LBs at all levels were full of verb-based and research-oriented bundles. In addition, there are particular features that are mostly characteristic of ESL learners' writing, the highly overused of verb-based bundles and research-oriented bundles. Furthermore, salient bundles were mostly used in reference to the written genre (argumentative essays); and confirmed the overuse of connector and personal pronoun discussed in chapter 3. It was argued that the writing genre is highly influenced the prevalence use of certain LBs in ESL learners' writing. Finally, there is a positive relationship between the use of LBs and academic performance, as learners developed their use of LBs over time.

To conclude, this study provided evidence for the significant variations and developmental use of LBs in B1, B2 and C1 argumentative essays. The nature of the variations and development suggests that the writing style expected in argumentative essays corresponds to that expected in ESL learners' writing. This leads to the conclusion that LBs could serve as a predictor of students' ability to produce academic writing. At the same time, caution should be exercised in relation to interpreting the findings, as it is uncertain whether these bundles appear across different types of writing tasks or if they are specific to certain task types. Therefore, further research into the types of LBs that occur in writing other than argumentative essays is needed. This would shed light on whether certain LBs are prevalent in all academic writing while others are content-specific.

Bibliography

Ädel, A. & Erman, B. 2012. Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for specific purposes 31*(2)**,** 81-92.

Ädel, A. & Römer, U. 2012. Research on advanced student writing across disciplines and levels: Introducing the Michigan Corpus of Upper-level Student Papers. *International Journal of Corpus Linguistics, 17*(1)**,** 3-34.

Adolphs, S. 2006. *Introducing electronic text analysis: A practical guide for language and literary studies*, Routledge.

Adolphs, S., Brown, B., Carter, R., Crawford, P. & Sahota, O. 2004. Applying corpus linguistics in a health care context. *Journal of Applied Linguistics and Language Research, 1*(1)**,** 9-28.

Ağçam, R. 2014. A corpus-based study on author stance in academic English. *American Journal of Educational Research, 2*(12)**,** 1230-1236.

Agresti, A. 2018. *An introduction to categorical data analysis*, Wiley.

Aijmer, K. 2001. I think as a marker of discourse style in argumentative student writing. *Gothenburg studies in English,* (81)**,** 247-257.

Akahori, N. 2007. The excessive use of the first-person pronoun "I" in English compositions by Japanese students. [Building learner corpora of Japanese learners of English and the contrastive analyses]. (Report for Grant-in Aid for Scientific Research (B) No. 15320059). Tokyo: Brainsnetwork.

Alali, F. & Schmitt, N. 2012. Teaching formulaic sequences: The same as or different from teaching single words? *JALT Journal 34, 3*(2)**,** 153-180.

Alderson, J. C. 2007. The CEFR and the need for more research. *The Modern Language Journal, 91*(4)**,** 659-663.

Alhassan, L. & Wood, D. 2015. The effectiveness of focused instruction of formulaic sequences in augmenting L2 learners' academic writing skills: A quantitative research study. *Journal of English for Academic Purposes, 17***,** 51-62.

Alipour, M. & Zarea, M. 2013. A Disciplinary Study of Lexical Bundles: The Case of Native versus Non-Native Corpora. *Taiwan International ESP Journal, 5*(2)**,** 1-20.

Allan, R. 2016. Lexical bundles in graded readers: To what extent does language restriction affect lexical patterning? *System, 59***,** 61-72.

Allen, D. 2009. Lexical bundles in learner writing: An analysis of formulaic language in the ALESS learner corpus. *Komaba Journal of English Education, 1***,** 105-127.

Altenberg, B. 1998. *On the phraseology of spoken English: The evidence of recurrent word-combinations*, na.

Altenberg, B. & Tapper, M. 1998. The use of adverbial connectors in advanced Swedish learners' written English. *Learner English on computer***,** 80-93.

Alward, A. 2019. Exploring Self-Mention in The Yemeni EFL Argumentative Writing Across Three Proficiency Levels. *Issues in Language Studies, 8*(2)**,** 48-60.

Anada, R. P., Arsyad, S. & Dharmayana, I. W. 2018. Argumentative Features of International English Language Testing System (IELTS) Essays: A Rhetorical Analysis on Successful Exam Essays. *International Journal of Language Education, 2*(1)**,** 1-13.

Appel, R. 2011a. Lexical bundles in l2 english academic writing.

Appel, R. & Wood, D. 2016. Recurrent word combinations in EAP test-taker writing: Differences between high-and low-proficiency levels. *Language Assessment Quarterly, 13*(1)**,** 55-71.

Appel, R. F. 2011b. *Lexical bundles in university EAP exam writing samples: CAEL test essays.* Carleton University.

Applebee, A., Langer, J., Mullis, I., Latham, A. & Gentile, C. 1994. NAEP 1992 writing report card (Report No. 23-W01). Washington, DC: US Government Printing Office.

Archibald, A. & Jeffery, G. C. 2000. Second language acquisition and writing:: a multi-disciplinary approach. *Learning and instruction, 10*(1)**,** 1-11.

Ari, O. 2006. Review of Three Sofware Programs designed to identify lexical bundle. *Language Learning & Technology, 10*(1)**,** 30-37.

Aston, G. 1997. Small and large corpora in language learning. *Practical applications in language corpora***,** 51-62.

Aull, L. 2015. *First-year university writing: A corpus-based study with implications for pedagogy*, Springer.

Aull, L. L. & 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. 2004. Querying keywords: Questions of difference, frequency, and sense in keywords analysis. *Journal of English linguistics, 32*(4)**,** 346-359.

Baker, P. 2006. *Using corpora in discourse analysis*, A&C Black.

Baker, P. 2010. Corpus methods in linguistics. *Research methods in linguistics***,** 93-113.

Bal, B. 2010. *Analysis of Four-word Lexical Bundles in Published Resesarch Articles Written by Turkish Scholars.* Georgia State University.

Barnbrook, G. 1996. Language and computers: a practical introduction to the computer analysis of language. *Edinburgh Textbooks in Empirical Linguistics*.

Beng, C. & Keong, Y. 2017. Comparing structural and functional lexical bundles in MUET reading test. *Pertanika Journal of Social Sciences and Humanities, 25*(1)**,** 133-148.

Berber-Sardinha, T. Comparing corpora with WordSmith Tools: How large must the reference corpus be? Proceedings of the workshop on Comparing corpora, 2000. Association for Computational Linguistics, 7-13.

Bestgen, Y. 2018. Evaluating the frequency threshold for selecting lexical bundles by means of an extension of the Fisher's exact test. *Corpora, 13*(2)**,** 205-228.

Bestgen, Y. & Granger, S. 2014. Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing, 26***,** 28-41.

Biber, D. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and linguistic computing, 5*(4)**,** 257-269.

Biber, D. 1993. Representativeness in corpus design. *Literary and linguistic computing, 8*(4)**,** 243-257.

Biber, D. 2006a. Stance in spoken and written university registers. *Journal of English for Academic Purposes, 5*(2)**,** 97-116.

Biber, D. 2006b. *University language: A corpus-based study of spoken and written registers*, John Benjamins Publishing.

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

Biber, D. & Conard, S. 2005. The Frequency and Use of Lexical Bundles in Conversation and Academic Prose. *Lexicographica, 20***,** 56-71.

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

Biber, D., Conard, S. & Leech, G. 2002. *Longman Student Grammar of Spoken and Written English,* Harlow,Essex, Longman.

Biber, D., Conard, S. & Rebben, R. 1998a. *Corpus Linguistics,* Cambrodge, Cambridge [etc].

Biber, D. & Conrad, S. 1999. Lexical bundles in conversation and academic prose. *Language and Computers, 26***,** 181-190.

Biber, D. & Conrad, S. 2009. Register, genre, and style, Cambridge, CUP.

Biber, D., Douglas, B., Conrad, S. & Reppen, R. 1998b. *Corpus linguistics: Investigating language structure and use*, Cambridge University Press.

Biber, D. & Egbert, J. 2018. *Register variation online*, Cambridge University Press.

Biber, D. & Gray, B. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes, 9*(1)**,** 2-20.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. & Quirk, R. 1999. *Longman grammar of spoken and written English*, MIT Press Cambridge, MA.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H. & Demecheleer, M. 2006. Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language teaching research, 10*(3), 245-261.

Breyer, Y. 2011. Corpora in language teaching and learning. *Potential, Evaluation, Challenges. Collection: English Corpus Linguistics, 13*.

Burgess, S. & Cargill, M. 2013. Using genre analysis and corpus linguistics to teach research article writing. *Supporting research writing.* Elsevier.

Burstein, J., Elliot, N. & Molloy, H. 2016. Informing Automated Writing Evaluation Using the Lens of Genre: Two Studies. *Calico journal, 33*(1), 117-141.

Butler, J. A. & Britt, M. A. 2011. Investigating instruction for improving revision of argumentative essays. *Written Communication,, 28*(1), 70-96.

Bychkovska, T. & 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. & 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*(5), 31-64.

Byrne, S. 2016. *An examination of successful language use at B1, B2 and C1 level in UCLanESB speaking tests in accordance with the Common European Framework of References for Languages.* University of Central Lancashire.

Byrnes, H. 2007. Introduction to Perspectives. *The Modern Language Journal*, 641-685.

Cacchiani, S. Keywords and key lexical bundles as cues to knowledge construction in RAs in economics. PALC 2009, 2011. Peter Lang, 335-350.

Campbell, Y. C. & Filimon, C. 2018. Supporting the argumentative writing of students in linguistically diverse classrooms: An action research study. *RMLE Online, 41*(1), 1-10.

Candarli, D. 2020. A longitudinal study of multi-word constructions in L2 academic writing: the effects of frequency and dispersion. *Reading and Writing*, 1-33.

Chen & Baker 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30-49.

Chen, L. 2008. *An investigation of lexical bundles in electrical engineering introductory textbooks and ESP textbooks.* Carleton University Ottawa.

Chen, X. & Xiao, G. A Corpus-based Study of the Structures and Functions of Academic Chunks in the Discipline of Engineering. 5th International Conference on Information Engineering for Mechanics and Materials, 2015. Atlantis Press.

Chen, Y. 2009. *Investigating Lexical Bundles Across Learner Writing Development.* PhD, Lancaster University, United Kingdom.

Chen, Y. & Baker, P. 2016. Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics, 37*(6), 849-880.

Chou, L. 2011. An Investigation of Taiwanese Doctoral Students' Academic Writing at a US University. *Higher Education Studies, 1*(2), 47-60.

Chujo, K. & Utiyama, M. 2006. Selecting level-specific specialized vocabulary using statistical measures. *System, 34*(2), 255-269.

Cobb, T. 2003. Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *Canadian Modern Language Review, 59*(3), 393-424.

Cobuild, C. 1992. *English usage*, Harper Collins Publishers.

Cock, D. 2000. *Repetitive phrasal chunkiness and advanced EFL speech and writing,* Amsterdam, Rodopi.

Coe, C. O. E. 2014. *Education and languages, language policy: The Common European Framework of Reference for Languages: Learning, teaching and assessment* [Online]. Available: https://www.coe.int/en/web/common-european-framework-reference-languages/ [Accessed 23/03/2018].

Cohen, J. 1988. Statistical power analysis. *Current directions in psychological science, 1*(3), 98-101.

Conklin, K. & Schmitt, N. 2008. Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *29*(1)**,** 72-89.

Conklin, K. & Schmitt, N. 2012. The processing of formulaic language. *32***,** 45.

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

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

Cooper, P. A. 2016. *Academic vocabulary and lexical bundles in the writing of undergraduate psychology students.* University of South Africa.

Cooper, T. 2013. Can IELTS writing scores predict university performance? Comparing the use of lexical bundles in IELTS writing tests and first-year academic writing. *Stellenbosch Papers in Linguistics Plus, 42***,** 63-79.

Cortes, V. 2002. *Lexical bundles in freshman composition,* Amsterdam, John Benjamins Publishing Company.

Cortes, V. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23***,** 397–423.

Cortes, V. 2006. Teaching lexical bundles in the disciplines: An example form a writing intensive history class. *Linguistics and Education, 17(4)***,** 391-406.

Cortes, V. 2013. The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for academic purposes, 12*(1)**,** 33-43.

Cortes, V. 2015. Situating lexical bundles in the formulaic language spectrum: Origins and functional analysis developments. *Corpus-based research in applied linguistics.* John Benjamins.

Cowie, A. 1998. *Phraseology: Theory, analysis, and applications*, OUP Oxford.

Coxhead, A. & Byrd, P. 2007. Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of second language writing, 16*(3)**,** 129-147.

Crewe, W. 1990. *The illogic of logical connectives*.

Crossley, S. & Mcnamara, D. 2014. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26***,** 66-79.

Crossley, S. & Salsbury, T. 2011. The development of lexical bundle accuracy and production in English second language speakers. *International Review of Applied Linguistics in Language Teaching, 49*(1)**,** 1-26.

Crossley, S., Skalicky, S., Kyle, K. & Monteiro, K. 2019. Absolute frequency effects in second language lexical acquisition. *Studies in Second Language Acquisition, 41*(4)**,** 721-744.

Crowhurst, M. 1990. Teaching and learning the writing of persuasive/argumentative discourse. *Canadian Journal of Education/Revue canadienne de l'éducation***,** 348-359.

Csomay, E. 2013. Lexical Bundles in Discourse Structure: A Corpus-Based Study of Classroom. *Applied Linguistics, 34(30***,** 369–388.

Culpeper, J. 2009. Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics, 14*(1)**,** 29-59.

Dastjerdi, H. V. & Samian, S. H. 2011. Quality of Iranian EFL learners' argumentative essays: Cohesive devices in focus. *Mediterranean journal of social sciences, 2*(2)**,** 65-65.

David, M. & Sutton, C. D. 2004. *Social research: The basics*, Sage.

De Cock, S. 2003. *Recurrent sequences of words in native speaker and advanced learner spoken and written English: A corpus-driven approach.* UCL-Université Catholique de Louvain.

De Schryver, G.-M. 2012. Trends in twenty-five years of academic lexicography. *International Journal of Lexicography, 25*(4)**,** 464-506.

Demetriou, T. 2019. *Predicting IELTS ratings using vocabulary measures.*

Destigter, T. 2015. On the ascendance of argument: A critique of the assumptions of academe's dominant form. *Research in the Teaching of English*, 11-34.

Dontcheva-Navratilova, O. 2012a. LEXICAL BUNDLES IN ACADEMIC TEXTS BY NON-NATIVE SPEAKERS. *Brno Studies in English, 38*(2).

Dontcheva-Navratilova, O. 2012b. Lexical bundles in academic texts by non-native speakers.

Dörnyei, Z. 2007. *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*, Oxford University Press Oxford.

Dryer, D. B. 2013. Scaling writing ability: A corpus-driven inquiry. *Written Communication, 30*(1), 3-35.

Du, J. 2013. *The use of lexical bundles by Chinese EFL English-major undergraduates at different university levels: A corpus-based study of L2 learners' examination essays.* Chinese University of Hong Kong.

Dueñas, P. M. 2007. 'I/we focus on…': A cross-cultural analysis of self-mentions in business management research articles. *Journal of English for Academic Purposes, 6*(2), 143-162.

Dunning, T. E. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics, 19*(1), 61-74.

Durrant, P. 2015. Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics, 38*(2), 165-193.

Ellis, N., Römer, U. & O'donnell, M. B. 2016. Constructions and usage-based approaches to language acquisition. *Language Learning, 66*(1), 23-44.

Ellis, N., Simpson-Vlach, R. & Maynard, C. 2008. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *Tesol Quarterly, 42*(3), 375-396.

Englishuk. 2019. *Student Statistics Report 2019* [Online]. Vienna, Austria: BONARD Available: https://www.englishuk.com/uploads/assets/members/publications/statistics/2019/English_UK_Student_Statistics_Report_2019_Executive_Summary.pdf [Accessed 1/07 2019].

Erman, B. & Warren, B. 2000. The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse, 20*(1), 29-62.

Esfandiari, R. & Barbary, F. 2017. A contrastive corpus-driven study of lexical bundles between English writers and Persian writers in psychology research articles. *Journal of English for Academic Purposes, 29*, 21-42.

Europe, C. O. 2001. *Common European Framework of Reference for Languages: learning, teaching, assessment*, Cambridge University Press.

Europe, C. O. 2003. *Relating Language Examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEF). Manual: Preliminary Pilot Version. DGIV/EDU/LANG 2003,5. Strasbourg: Language Policy Division.*

[Online]. [Accessed 12/4 2018].

Europe, C. O. 2020. *Common European Framework of Reference for Languages: learning, teaching, assessment*, Cambridge University Press.

Evans, S. & Green, C. 2007. Why EAP is necessary: A survey of Hong Kong tertiary students. *Journal of English for Academic Purposes, 6*(1), 3-17.

Feak, C. & Dobson, B. 1996. Building on the Impromptu: A Source-Based Academic Writing Assessment. *College esl, 6*(1), 73-84.

Ferretti, R. P., Andrews-Weckerly, S. & Lewis, W. E. 2007. Improving the argumentative writing of students with learning disabilities: Descriptive and normative considerations. *Reading & Writing Quarterly, 23*(3), 267-285.

Ferris, D. & Tagg, T. 1996. Academic listening/speaking tasks for ESL students: Problems, suggestions, and implications. *TESOL quarterly, 30*(2), 297-320.

Field, Y. & Oi, Y. 1992. A comparison of internal conjunctive cohesion in the English essay writing of Cantonese speakers and native speakers of English. *RELC journal, 23*(1), 15-28.

Figueras, N. 2012. The impact of the CEFR. *ELT journal, 66*(4), 477-485.

Firoozjahantigh, M., Fakhri Alamdari, E. & Marzban, A. 2021. Investigating the Effect of Process-based Instruction of Writing on the IELTS Writing Task Two Performance of Iranian EFL Learners: Focusing on Hedging & Boosting. *Cogent Education, 8*(1), 1881202.

Firth, J. 1957. *Papers in Linguistics,* London, Oxford University Press.

Fleiss, J. L., Levin, B. & Paik, M. C. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions, 2*(212-236), 22-23.

Flowerdew, J. 1996. Concordancing in language learning. *The power of CALL*, 97-113.

Flowerdew, L. 2004. The argument for using English specialized corpora to understand academic and professional language. *Discourse in the professions: Perspectives from corpus linguistics*, 11-33.

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, Skehan P, & Swain M (Eds.), Researching pedagogic tasks: second language learning, teaching, and testing (pp. 75–93).

Foster, P., Tonkyn, A. & Wigglesworth, G. 2000. Measuring spoken language: A unit for all reasons. *Applied linguistics, 21*(3), 354-375.

Francis, W. Language corpora BC. Directions in Linguistics: Proceedings of Nobel Symposium, 1992. 17-32.

Freddi, M. 2005. Arguing linguistics: Corpus investigation of one functional variety of academic discourse. *Journal of English for Academic Purposes, 4*(1), 5-26.

Fromm, G., Grama, D. F., Beilke, N. S. V. & Santos, C. G. 2020. Wordsmith Tools e Sketch Engine: corpora/Wordsmith Tools and Sketch Engine: an analytical-comparative study for scientific research with corpora manipulation. *Revista de Estudos da Linguagem, 28*(3), 1191-1248.

Gabrielatos, C. 2007. Selecting query terms to build a specialised corpus from a restricted-access database. *ICAME journal, 31*, 5-44.

Garner, J., Crossley, S. & Kyle, K. 2019. N-gram measures and L2 writing proficiency. *System, 80*, 176-187.

Geiser, S. & Studley, W. R. 2002. UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. *Educational Assessment, 8*(1), 1-26.

Gere, A. R., Aull, L., Escudero, M. D. P., Lancaster, Z. & Lei, E. V. 2013. Local assessment: Using genre analysis to validate directed self-placement. *College Composition and Communication*, 605-633.

Geva, E. 1992. The role of conjunctions in L2 text comprehension. *TESOL quarterly, 26*(4), 731-747.

Gil, N. N. & Caro, E. M. 2019. Lexical bundles in learner and expert academic writing. *Bellaterra Journal of Teaching Learning Language Literature, 12*(1), 65-90.

Gilquin, G., Granger, S. & Paquot, M. 2007. Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes, 6*(4), 319-335.

Gilquin, G. & Gries, S. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus linguistics and linguistic theory, 5*(1), 1-26.

Gilquin, G. & Paquot, M. Spoken features in learner academic writing: Identification, explanation and solution. Proceedings of the 4th Corpus Linguistics Conference, University of Birmingham, 2007. 27-30.

Glaboniat, M., Müller, M., Schmitz, H., Rusch, P. & Wertenschlag, L. 2005. *Profile deutsch*, Langenscheidt Berlin.

Goh, G. 2011. Choosing a reference corpus for keyword calculation. *Linguistic Research, 28*(1), 239-256.

Goźdź-Roszkowski, S. 2011. *Patterns of linguistic variation in American legal English: A corpus-based study*, Peter Lang Frankfurt am Main.

Grabowski, Ł. 2013. Register variation across English pharmaceutical texts: A corpus-driven study of keywords, lexical bundles and phrase frames in patient information leaflets and summaries of product characteristics. *Procedia-Social and Behavioral Sciences, 95*, 391-401.

Grabowski, Ł. 2015. Keywords and lexical bundles within English pharmaceutical discourse: A corpus-driven description. *English for Specific Purposes, 38*, 23-33.

Granger, S. 1998a. The computer learner corpus: a versatile new source of data for SLA research. *Learner English on computer*, 3-18.

Granger, S. 1998b. Prefabricated patterns in advanced EFL writing: Collocations and lexical phrases. *Phraseology: Theory, analysis and applications*, 145-160.

Granger, S. 2002. A bird's-eye view of learner corpus research. *Computer learner corpora, second language acquisition and foreign language teaching, 6*, 3-33.

Granger, S. 2004. Computer learner corpus research: Current status and future prospects. *Applied Corpus Linguistics.* Brill Rodopi.

Granger, S. 2012. Learner corpora. *The encyclopedia of applied linguistics*.

Granger, S. & Meunier, F. 2008. *Phraseology: An interdisciplinary perspective,* Amsterdam, John Benjamins.

Granger, S. & Tyson, S. 1996. Connector usage in the English essay writing of native and non-native EFL speakers of English. *15*(1), 17-27.

Gray, B. & Biber, D. 2013. Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics, 18(1)*, 109-136.

Greenbaum, S. 1991. ICE: The international corpus of English. *7*(4), 3-7.

Gries, S. 2008. Corpus-based methods in analyses of second language acquisition data. *Handbook of cognitive linguistics and second language acquisition.* Routledge.

Groom, N. 2010. Closed-class keywords and corpus-driven discourse analysis. *Keyness in texts*, 59-78.

Güngör, F. & Uysal, H. 2016. A Comparative Analysis of Lexical Bundles Used by Native and Non-native Scholars. *English Language Teaching, 9*(6), 176-188.

Haefner, J. 1992. Democracy, pedagogy, and the personal essay. *College English, 54*(2), 127-137.

Halliday, M. & Hasan, R. 1976. *Cohesion in English,* London, Longman.

Halliday, M. & Matthiessen, C. 2013. *Halliday's introduction to functional grammar*, Routledge.

Halliday, M. a. K., Matthiessen, C. M., Halliday, M. & Matthiessen, C. 2014. *An introduction to functional grammar*, Routledge.

Harmer, J. 2001. The practice of English language teaching. *London/New York*, 401-405.

Hartse, J. H. & Kubota, R. 2014. Pluralizing English? Variation in high-stakes academic texts and challenges of copyediting. *Journal of Second Language Writing, 24*, 71-82.

Hasselgård, H. 2019. Forthcoming. Phraseological teddy bears: frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. *Corpus Linguistics, Context and Culture. Berlin: De Gruyter Mouton.*

Hasselgren, A. 1994. Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics, 4*(2), 237-258.

Heath, S. B. 1993. Rethinking the sense of the past: The essay as legacy of the epigram. *Theory and practice in the teaching of writing: Rethinking the discipline, 105131.*

Heino, P. 2010. Adverbial connectors in advanced EFL learners' and native speakers' student writing.

Heng, C. S., Kashiha, H. & Tan, H. 2014. Lexical Bundles: Facilitating University" Talk" in Group Discussions. *English Language Teaching, 7*(4), 1-10.

Henriksen, B. 2013. Research on L2 learners' collocational competence and development–a progress report. *L2 VOCABULARY ACQUISITION, KNOWLEDGE AND USE new perspectives on assessment and corpus analysis*, 29-56.

Hernández, P. 2013. Lexical bundles in three oral corpora of university students. *Nordic Journal of English Studies, 12*(1)**,** 187-209.

Hewings, M. 2010. Materials for university essay writing. *English language teaching materials***,** 251-278.

Hewings, M. & Hewings, A. 2002. "It is interesting to note that…": a comparative study of anticipatory 'it'in student and published writing. *English for Specific Purposes, 21*(4)**,** 367-383.

Hinkel, E. 2002. *Second language writers' text: Linguistic and rhetorical features*, Routledge.

Hinkel, E. 2003. Adverbial markers and tone in L1 and L2 students' writing. *Journal of Pragmatics, 35*(7)**,** 1049-1068.

Hinkel, E. 2004. Teaching Academic ESL, Writing: Practical Technigues in Vocabulary and Grammar. New Jersey: Lawrence Earlbaum Associates. Inc.

Hinkel, E. 2005. Hedging, inflating, and persuading in L2 academic writing. *Applied Language Learning, 15*(1/2)**,** 29.

Hoey, J. 2009. The two-way likelihood ratio (G) test. Citeseer.

Hoffmann, S., Evert, S., Smith, N., Lee, D. & Berglund-Prytz, Y. 2008. *Corpus linguistics with BNCweb-a practical guide*, Peter Lang.

Hong, S. 2013. An n-gram analysis of Korean English learners' writing. *13*(2)**,** 313-336.

Horkoff, T. & Mclean, S. 2015. *Writing for Success: 1st Canadian Edition*.

Horst, M. & Meara, P. 1999. Test of a model for predicting second language lexical growth through reading. *Canadian Modern Language Review, 56*(2)**,** 308-328.

Howarth, P. 1998. Phraseology and second language proficiency. *Applied linguistics, 19*(1)**,** 24-44.

Howarth, P. A. 1996. *Phraseology in English academic writing: Some implications for language learning and dictionary making*, Walter de Gruyter.

Hua, Z. & David, A. 2008. Study design: Cross-sectional, longitudinal, case, and group. *The Blackwell guide to research methods in bilingualism and multilingualism***,** 88-107.

Huang, J. 2005. Challenges of academic listening in English: Reports by Chinese students. *College student journal, 39*(3)**,** 553-570.

Huang, K. 2014. A corpus study of Chinese EFL majors' phraseological performance. *HKU Theses Online (HKUTO)*.

Huang, K. 2015. More does not mean better: Frequency and accuracy analysis of lexical bundles in Chinese EFL learners' essay writing. *System, 53***,** 13-23.

Huhta, A., Alanen, R., Tarnanen, M., Martin, M. & Hirvelä, T. 2014. Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing, 31*(3)**,** 307-328.

Hulstijn, J. 2007. The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*(4)**,** 663-667.

Hunston, S. 2002. *Corpora in applied linguistics*, Ernst Klett Sprachen.

Hyland, K. 1990. A genre description of the argumentative essay. *RELC journal, 21*(1)**,** 66-78.

Hyland, K. 1994. Hedging in academic writing and EAF textbooks. *English for specific purpose, 13*(3)**,** 239-256.

Hyland, K. 1998. *Hedging in scientific research articles*, John Benjamins Publishing Company Amsterdam.

Hyland, K. 2002. Authority and invisibility: Authorial identity in academic writing. *Journal of pragmatics, 34*(8)**,** 1091-1112.

Hyland, K. 2004a. *Disciplinary discourses, Michigan classics ed.: Social interactions in academic writing*, University of Michigan Press.

Hyland, K. 2004b. *Genre and second language writing*, University of Michigan Press.

Hyland, K. 2005. Stance and engagement: A model of interaction in academic discourse. *7*(2)**,** 173-192.

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*, 4–21.

Hyland, K. 2009a. *Academic discourse: English in a global context*, A&C Black.

Hyland, K. 2009b. Writing in the disciplines: Research evidence for specificity. *Taiwan International ESP Journal, 1*(1), 5-22.

Hyland, K. 2012. Bundles in academic discourse. *Annual review of applied linguistics, 32*, 150-169.

Hyland, K. & Jiang, F. 2018. Academic lexical bundles. *International Journal of Corpus Linguistics, 23*(4), 383-407.

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

Ivanič, R. & Camps, D. 2001. I am how I sound: Voice as self-representation in L2 writing. *Journal of second language writing, 10*(1-2), 3-33.

Jablonkai, R. 2009. In the light of": A corpus-based analysis of lexical bundles in two EU-related registers. *Corvinus University of Budapest: WopaLP, 3*, 1-27.

Jalali, H. 2015. Examining Novices' Selection of Lexical Bundles: The Case of EFL Postgraduate Students in Applied Linguistics. *Journal of Applied Linguistics and Language Research, 1*(2), 1-11.

Jalali, H., Rasekh, A. & Rizi, M. 2009. Anticipatory'it'lexical bundles: A comparative study of student and published writing in applied linguistics. *Iranian Journal of Language Studies, 3*(2).

Jalali, H. & Zarei, G. 2016. Academic writing revisited: A phraseological analysis of applied linguistics high-stake genres from the perspective of lexical bundles. *Journal of Teaching Language Skills, 34*(4), 87-114.

Jalali, Z., Moini, M. & Arani, M. 2014. Structural and functional analysis of lexical bundles in medical research articles: A corpus-based study. *International Journal of Information Science and Management (IJISM), 13*(1).

Jaworska, S., Krummes, C. & Ensslin, A. 2015. Formulaic sequences in native and non-native argumentative writing in German. *International Journal of Corpus Linguistics, 20*(4), 500-525.

Jo, C. W. 2017. *Exploring Adolescents' Persuasive Essays: Toward Promoting Linguistic and Intercultural Competence.* Doctoral dissertation, Harvard University.

Joharry, S. 2016. *Malaysian learners' argumentative writing in English: A contrastive, corpus-driven study.* Doctoral dissertation.

Johnston, K. 2017. *Lexical bundles in applied linguistics and literature writing: A comparison of intermediate English learners and professionals.* Doctoral dissertation, Portland State University.

Jones, C. & Waller, D. 2015. *Corpus linguistics for grammar: A guide for research*, Routledge.

Jones, C., Waller, D. & Golebiewska, P. 2013. Defining successful spoken language at B2 level: findings from a corpus of learner test data. *The European Journal of Applied Linguistics and TEFL, 2*(2), 29-46.

Jones, M. & Haywood, S. 2004. Facilitating the acquisition of formulaic sequences. *Formulaic sequences*, 269-300.

Juknevičienė, R. 2009. Lexical bundles in learner language: Lithuanian learners vs. native speakers. *Kalbotyra, 61*(61), 61-72.

Karabacaka, E. & Qinb, J. 2013. Comparison of lexical bundles used by Turkish, Chinese, and American university students. *Procedia - Social and Behavioral Sciences, 70*, 622 – 628.

Kashiha, H. & Heng, C. 2013. An exploration of lexical bundles in academic lectures: examples from hard and soft sciences. *The Journal of AsiaTEFL, 10*(4), 133-161.

Kazemi, M., Katiraei, S. & Rasekh, A. E. 2014. The impact of teaching lexical bundles on improving Iranian EFL students' writing skill. *Procedia-Social and Behavioral Sciences, 98*, 864-869.

Kecskes, I. 2016. Deliberate creativity and formulaic language use. *Pragmemes and theories of language use.* Springer, Charm.

Kennedy, G. An Introduction to Corpus Linguistics. Studies in Language and Linguistics, 1998. Citeseer.

Kim, E. 2020. A Contrastive Corpus-Based Analysis of Lexical Bundles between English L1 and English L2 Writers in Medical Journal Abstracts. Ewha Womans University, Seoul.

Kim, J. 2013. Lexical Bundles in Korean College Students' English Essays: A Corpus-based Comparative Study. *English Langauge Eduaction 19*, 157-179.

Kim, J. 2019. The Use of Adverbial Connectors in Korean College Students' English Essays: A Corpus-based Comparative Study. *English Language Education, 35*(2)**,** 83-104.

Knudson, R. E. 1992. The development of written argumentation: An analysis and comparison of argumentative writing at four grade levels. *Child study journal*.

Knudson, R. E. 1994. An analysis of persuasive discourse: Learning how to take a stand. *Discourse Processes, 18*(2)**,** 211-230.

Krummes, C. & Ensslin, A. 2015. Formulaic language and collocations in German essays: from corpus-driven data to corpus-based materials *The Language Learning Journal, 43(1)***,** 110-127.

Kwary, D., Ratri, D. & Artha, A. 2017. Lexical bundles in journal articles across academic disciplines. *Indonesian Journal of Applied Linguistics, 7*(1)**,** 131-140.

Kwary, D. A. 2011. A hybrid method for determining technical vocabulary. *System, 39*(2)**,** 175-185.

Laufer, B. & Waldman, T. 2011. Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language learning, 61*(2)**,** 647-672.

Lea, M. R. & Street, B. V. 1998. Student writing in higher education: An academic literacies approach. *Studies in higher education, 23*(2)**,** 157-172.

Lee, D. & Chen, S. 2009. Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners *Journal of Second Language Writing, 18*(4)**,** 280-296.

Lee, E. 2004. A corpus-based analysis of the Korean EFL learners' use of conjunctive adverbials. *English Teaching, 59*(4)**,** 283-301.

Leech, G. 2002. The importance of reference corpora. *Hizkuntza-corpusak. Oraina eta geroa*.

Leech, G. 2007. New resources, or just better old ones? The Holy Grail of representativeness. *Corpus linguistics and the web.* Brill Rodopi.

Leedham, M. 2011. *A corpus-driven study of features of Chinese students' undergraduate writing in UK universities.* The Open University.

Leedham, M. & Cai, G. 2013. Besides… on the other hand: Using a corpus approach to explore the influence of teaching materials on Chinese students' use of linking adverbials. *Journal of Second Language Writing, 22*(4)**,** 374-389.

Lehecka, T. 2015. Collocation and colligation. *Handbook of pragmatics online.* Benjamins.

Lei, L. 2012. Linking adverbials in academic writing on applied linguistics by Chinese doctoral students. *Journal of English for Academic Purposes, 11*(3)**,** 267-275.

Leki, I. & Carson, J. 1994. Students' perceptions of EAP writing instruction and writing needs across the disciplines. *TESOL quarterly, 28*(1)**,** 81-101.

Leńko-Szymańska, A. 2008. Non-native or non-expert? The use of connectors in native and foreign language learners' texts. *Acquisition et interaction en langue étrangère,* (27)**,** 91-108.

Leone, P. 2010. General spoken language and school language: Key words and discourse patterns in history textbooks. *Keyness in texts.* John Benjamins.

Levon, E. 2010. Organizing and processing your data: The nuts and bolts of quantitative analyses. *Research methods in linguistics*, 68.

Lewis, M. 2000. *Language in the lexical approach*.

Li, J. & Schmitt, N. 2009. The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing, 18*(2)**,** 85-102.

Li, Z. & Volkov, A. 2017. " To Whom it May Concern": A Study on the Use of Lexical Bundles in Email Writing Tasks in an English Proficiency Test. *TESL Canada Journal, 34*(3)**,** 54-75.

Liu, C. & Chen, H. 2020. Analyzing the functions of lexical bundles in undergraduate academic lectures for pedagogical use. *English for Specific Purposes, 58***,** 122-137.

Liu, D. 2008. Linking adverbials: An across-register corpus study and its implications. *International journal of corpus linguistics, 13*(4)**,** 491-518.

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.

Lu, X. & Deng, J. 2019. With the rapid development: A contrastive analysis of lexical bundles in dissertation abstracts by Chinese and L1 English doctoral students. *Journal of English for Academic Purposes, 39***,** 21-36.

Luab, W., Leeb, S.-M. & Jhang, S.-E. 2017. Keyness in maritime institutional law texts.

Luzón, M. 2009. The use of we in a learner corpus of reports written by EFL Engineering students. *Journal of English for Academic Purposes, 8*(3)**,** 192-206.

Maire, C. 1999. The Freiburg-LOB Corpus of British English ('FLOB').

Marco, M. 2000. Collocational frameworks in medical research papers: A genre-based study. *English for specific purposes, 19*(1)**,** 63-86.

Martínez, I. 2005. Native and non-native writers' use of first person pronouns in the different sections of biology research articles in English. *Journal of Second Language Writing, 14*(3)**,** 174-190.

Mauranen, A. & Bondi, M. 2003. Evaluative language use in academic discourse. *Journal of English for Academic Purposes, 2*(4)**,** 269-271.

Mccann, T. 1989. Student argumentative writing knowledge and ability at three grade levels. *Research in the Teaching of English***,** 62-76.

Mccrostie, J. 2008a. Writer visibility in EFL learner academic writing: A corpus-based study. *32*(1)**,** 97-114.

Mccrostie, J. 2008b. Writer visibility in EFL learner academic writing: A corpus-based study. *Icame Journal, 32*(1)**,** 97-114.

Mcenery, A. & Wilson, A. 2001. *Corpus linguistics: an introduction*, Edinburgh University Press.

Mcenery, T., Mcenery, A., Xiao, R. & Tono, Y. 2006. *Corpus-based language studies: An advanced resource book*, Taylor & Francis.

Mcenery, T. & Wilson, A. 1996. *Corpus linguistics,* Edinburgh, Edinburgh Textbooks in Applied Linguistics.

Mcenery, T. & Xiao, R. 2011. What corpora can offer in language teaching and learning. *Handbook of research in second language teaching and learning, 2***,** 364-380.

Mchugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb), 22*(3)**,** 276-82.

Mei, W. S. 2006. Creating a contrastive rhetorical stance: Investigating the strategy of problematization in students' argumentation. *RELC journal, 37*(3)**,** 329-353.

Meunier, F. & Granger, S. 2007. *Phraseology in Foreign Language Learning and Teaching,* Amsterdam, John Benjamins Publishing.

Milton, J. 1998. Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. *In S. Granger (ed.) Learner English on Computer, 186-198. Longman, London and New York.*

Mirhosseini, S. 2017. Introduction: Qualitative research in language and literacy education. *Reflections on qualitative research in language and literacy education.* Springer.

Moon, R. 1998. *Fixed expressions and idioms in English: A corpus-based approach,* Oxford, Clarendon Press.

Moore, T. & Morton, J. 2005. Dimensions of difference: A comparison of university writing and IELTS writing. *Journal of English for Academic Purposes, 4*(1)**,** 43-66.

Moynie, J. 2018. *Lexical Bundles within English for Academic Purposes Written Teaching Materials: A Canadian Context.* Doctoral dissertation, Carleton University.

Muşlu, M. 2014. *Stance lexical bundles in academic L2 writing: A contrastive interlanguage analysis.* Unpublished doctoral dissertation, Çukurova University, Adana, Turkey.

Muşlu, M. 2018. Use of stance lexical bundles by Turkish and Japanese EFL learners and native English speakers in academic writing. *Gaziantep University Journal of Social Sciences, 17*(4)**,** 1319-1336.

Mutiara, R. 2018. Lexical bundles and keywords in psychology research articles. *Asian EFL Journal, 20*(7)**,** 135-142.

Natsukari, S. 2012. Use of I in essays by Japanese EFL learners. *34*(1)**,** 61-78.

Nattinger, J. & Decarrico, J. 1992. *Lexical phrases and language teaching*, Oxford University Press.

Neely, E. & Cortes, V. 2009. A little bit about: analyzing and teaching lexical bundles in academic lectures. *Language Value, 1*(1)**,** 17-38.

Neff-Van Aertselaer, J. & Dafouz-Milne, E. 2008. Argumentation patterns in different languages: An analysis of metadiscourse markers in English and Spanish texts. *Developing contrastive pragmatics interlanguage and cross-cultural perspectives***,** 87-102.

Nekrasova-Beker, T. & Becker, A. 2020. The use of lexical patterns in engineering. *Advances in Corpus-based Research on Academic Writing: Effects of discipline, register, and writer expertise, 95***,** 227.

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

Németh, N. & Kormos, J. 2001. Pragmatic aspects of task-performance: The case of argumentation. *Language Teaching Research, 5*(3)**,** 213-240.

Nesi, H. & Basturkmen, H. 2006. Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics, 11*(3)**,** 283-304.

Nesi, H. & Gardner, S. 2006. Variation in disciplinary culture: University tutors' views on assessed writing tasks. *British studies in applied linguistics, 21***,** 99.

Nesi, H., Gardner, S., Thompson, P., Wickens, P., Forsyth, R., Heuboeck, A. & Alsop, S. 2008. An investigation of genres of assessed writing in British higher education: full research report ESRC end of award report.

Nesselhauf, N. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied linguistics, 24*(2)**,** 223-242.

Nesselhauf, N. 2005. *Collocations in a learner corpus*, John Benjamins Amsterdam.

Nippold, M. & Ward-Lonergan, J. 2010. Argumentative writing in pre-adolescents: The role of verbal reasoning. *Child Language Teaching and Therapy, 26*(3)**,** 238-248.

Nkemleke, D. 2012. A corpus-based investigation of lexical bundles in students' dissertations in Cameroon. *Human & Social Science Series, 3*(1), 1 - 20.

North, B. & Jones, N. 2009. Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)-Further Material on Maintaining Standards across Languages, Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling. *Council of Europe, Strasbourg*.

Novita, H. & Kwary, D. 2018. Comparing the use of lexical bundles in Indonesian-English translation by student translators and professional translators. *Translation & Interpreting, 10*(1)**,** 53-74.

O'donnell, M., Römer, U. & Ellis, N. 2013. The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics, 18*(1)**,** 83-108.

O'keeffe, A., Mccarthy, M. & Carter, R. 2007. *From corpus to classroom: Language use and language teaching*, Cambridge University Press.

Oakey, D. 2002. Formulaic language in English academic writing. *Using corpora to explore linguistic variation, 9***,** 111-129.

Oakey, D. 2009. Fixed collocational patterns in isolexical and isotextual versions of a corpus. *Contemporary corpus linguistics***,** 140-58.

Ohlrogge, A. 2009. Formulaic expressions in intermediate EFL writing assessment. *Formulaic language, 2*, 375-86.

Oktavianti, I. N. & Adnan, A. 2020. A Corpus Study of Verbs in Opinion Articles of The Jakarta Post and the Relation with Text Characteristics. *3*(2), 108-117.

Oktavianti, I. N. & Ardianti, N. R. 2019. A corpus-based analysis of verbs in news section of The Jakarta Post: How frequency is related to text characteristics. *JOALL (Journal of Applied Linguistics and Literature), 4*(2), 203-214.

Omidian, T., Shahriari, H. & Siyanova-Chanturia, A. 2018. A cross-disciplinary investigation of multi-word expressions in the moves of research article abstracts. *Journal of English for academic purposes, 36*, 1-14.

Ortega, L. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics, 24*(4), 492-518.

Oshima, A. & Hogue, A. 2007. *Introduction to academic writing*, Pearson/Longman.

Öztürk, Y. 2014. *Lexical bunlde use of Turkish and native English writers: a corpus-based study.* Doctoral dissertation, Anadolu University, turkey.

Ozturk, Y. & Kose, G. 2016. Turkish and native English academic writers?: Use of lexical bundles. *Journal of language and linguistic studies, 12*(1), 149.

Palmer, T. M. 2016. *A corpus-driven, keyword-centered approach to lexical bundles in grade comments.* Doctoral dissertation, San Diego State University.

Pan, F. & Liu, C. 2019. Comparing L1-L2 differences in lexical bundles in student and expert writing. *Southern African Linguistics and Applied Language Studies, 37*(2), 142-157.

Pan, F., Reppen, R. & Biber, D. 2016. Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in Telecommunications research journals. *Journal of English for Academic Purposes, 21*, 60-71.

Pang, P. 2009. A study on the use of four-word lexical bundles in argumentative essays by Chinese English-majors: A comparative study based on WECCL and LOCNESS. *Teaching English in China, 32*(3), 25-4.

Pang, W. 2010. Lexical Bundles and the Construction of an Academic Voice: A Pedagogical Perspective. *Asian EFL Journal. Professional Teaching Articles, 47*.

Panthong, P. & Poonpon, K. 2020. Lexical bundles in Thai medical research articles. *Journal of Studies in the English Language, 15*(1), 59-106.

Paquot, M. 2008. Exemplification in learner writing: A cross-linguistic perspective. InF. Meunier & S. Granger (Eds.), Phraseology in foreign language learning and teaching 101-119.

Paquot, M. 2010. *Academic vocabulary in learner writing: From extraction to analysis*, Bloomsbury Publishing.

Paquot, M. 2013. Lexical bundles and L1 transfer effects. *International Journal of Corpus Linguistics, 18*(3), 391-417.

Paquot, M. & Granger, S. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32*(32), 130-149.

Park, Y. 2013. How Korean EFL students use conjunctive adverbials in argumentative writing. *ENGLISH TEACHING, 68*(4), 263-284.

Parkinson, J. & Musgrave, J. 2014. Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes, 14*, 48-59.

Pawley, A. & Syder, H. 1983. *Two puzzles for linguistic theory: Nativelike selection and nativelike fluency,* New York, Longman.

Pearson, W. 2021. A comparative study of lexical bundles in IELTS Writing Task 1 and 2 simulation essays and tertiary academic writing. *Journal of Academic Language and Learning, 15*(1), 27-52.

Pecorari, D. 2009. Formulaic language in biology: A topic-specific investigation. *Academic writing: At the interface of corpus and discourse, 91*, 105.

Perkins, D. N. 1985. Postprimary education has little impact on informal reasoning. *Journal of educational psychology, 77*(5), 562.

Peromingo, J. 2012. Grammatical collocations in the written production of Spanish university students. *REDUCA (Philology), 2*(1).

Peters, A. 1983. *The units of language acquisition*, CUP Archive.

Pojanapunya, P. & Todd, R. 2018. Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory, 14*(1)**,** 133-167.

Pourmusa, F. 2014. A Comparative Study of Lexical Bundles in Soft Science Articles Written by Native and Iranian Authors. *Journal of Applied Linguistics and Applied Literature: Dynamics and Advances, 2*(2)**,** 67-83.

Povolná, R. 2016. A cross cultural analysis of conjuncts as indicators of the interaction and negotiation of meaning in research articles. *Topics in Linguistics, 17*(1)**,** 45-63.

Powell, P. 2009. Retention and writing instruction: Implications for access and pedagogy. *College Composition and Communication***,** 664-682.

Qin, J. 2014. Use of formulaic bundles by non-native English graduate writers and published authors in applied linguistics. *System, 42***,** 220-231.

Qin, J. & Karabacak, E. 2010. The analysis of Toulmin elements in Chinese EFL university argumentative writing. *System, 38*(3)**,** 444-456.

Qiufang, W., Yanren, D. & Wenyu, W. 2003. Features of Oral Style in English Composition of Advanced Chinese EFL Learners: An Exploratory Study by Contrastive Learner Corpus Analysis. *Foreign Language Teaching and Research, 4***,** 268-74.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1985. A comprehensive grammar of contemporary English. London: Longman.

Rachmawati, S. & Susanti, Y. 2016. The use of transitions in the students argumentative essay. *Journal of English Teaching and Research, 1*(2)**,** 10-10.

Rafieyan, V. 2018. Role of knowledge of formulaic sequences in language proficiency: significance and ideal method of instruction. *Journal of English Teaching and Research, 3*(1)**,** 1-23.

Ramage, J. D., Bean, J. C. & Johnson, J. 2018. *Writing arguments: A rhetoric with readings*, Pearson.

Rayson, P., Berridge, D. & Francis, B. Extending the Cochran rule for the comparison of word frequencies between corpora. 7th International Conference on Statistical analysis of textual data (JADT 2004), 2004. 926-936.

Rayson, P. & Garside, R. Comparing corpora using frequency profiling. The workshop on comparing corpora, 2000. 1-6.

Reppen, R. & Olson, S. 2020. Lexical bundles across disciplines. *Academic Writing: Effects of discipline, register, and writer expertise, 95***,** 169.

Riley, J. & Reedy, D. 2005. Developing young children's thinking through learning to write argument. *Journal of Early Childhood Literacy, 5*(1)**,** 29-51.

Römer, U. 2009. English in academia: Does nativeness matter. *Anglistik: International Journal of English Studies, 20(2)***,** 89-100.

Ruan, Z. 2017. Lexical bundles in Chinese undergraduate academic writing at an English medium university. *RELC Journal, 48*(3)**,** 327-340.

Saito, K. & Liu, Y. 2021. Roles of collocation in L2 oral proficiency revisited: Different tasks, L1 vs. L2 raters, and cross-sectional vs. longitudinal analyses. *Journal of Early Childhood Literacy***,** 0267658320988055.

Salazar, D. 2012. Lexical bundles in Philippine and British scientific English. *Journal of Early Childhood Literacy, 41*(1).

Salazar, D. 2014. *Lexical bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching*, John Benjamins Publishing Company.

Salazar, L. & Joy, D. 2011. *Lexical bundles in scientific English: A corpus-based study of native and non-native writing.* Doctoral dissertation, Universitat de Barcelona.

Scheepers, R. 2014. *Lexical Levels and Formulaic Language: An Exploration of Undergraduate Students' Vocabulary and Written Production of Delexical Multiword Units.* Doctoral dissertation, University of South Africa.

Schmitt, N. 2004. *Formulaic sequences: Acquisition, processing, and use*, John Benjamins Publishing.

Schmitt, N. 2010. *Researching vocabulary: A vocabulary research manual*, Springer.

Schmitt, N., Dornyei, Z., Adolphs, S. & Durow, V. 2004. Knowledge and acquisition of formulaic sequences. *Formulaic sequences acquisition, processing and use*, 55-86.

Scott, M. 1997. PC analysis of key words—and key key words. *System, 25*(2)**,** 233-245.

Scott, M. 2008. Oxford wordsmith tools 5.0 Manual. Oxford: Oxford University Press.

Scott, M. 2010. Problems in investigating keyness, or clearing the undergrowth and marking out trails. *Keyness in texts***,** 43-57.

Scott, M. 2012. WordSmith Tools (Computer Software. Version 6.0). *Liverpool: Lexical Analysis Software*.

Scott, M. 2015. Wordsmith Tools Help. Liverpool: Lexical Analysis Software.

Scott, M. & Tribble, C. 2006. *Textual patterns: Key words and corpus analysis in language education*, John Benjamins Publishing.

Seale, C. 2008. Mapping the field of medical sociology: a comparative analysis of journals. *Sociol Health Illn, 30*(5)**,** 677-95.

Shea, M. 2009. A corpus-based study of adverbial connectors in learner text. *MSU Working, 1*(1).

Shechtman, Z. 1992. Interrater reliability of a single group assessment procedure administered in several educational settings. *Journal of Personnel Evaluation in Education, 6*(1)**,** 31-39.

Shin, Y. 2018. Lexical bundles in argumentative essays by native and nonnative English-speaking novice academic writers.

Shin, Y. 2019. Do native writers always have a head start over nonnative writers? The use of lexical bundles in college students' essays. *40***,** 1-14.

Shin, Y. K. & Kim, Y. 2017. Using lexical bundles to teach articles to L2 English learners of different proficiencies. *System, 69***,** 79-91.

Shirazizadeh, M. & Amirfazlian, R. 2021. Lexical bundles in theses, articles and textbooks of applied linguistics: Investigating intradisciplinary uniformity and variation. *System, 49***,** 100946.

Simpson-Vlach, R. & Ellis, N. C. 2010. An academic formulas list: New methods in phraseology research. *Applied linguistics, 31*(4)**,** 487-512.

Simpson, J. & Weinert, E. 2013. Oxford dictionary. *11*.

Simpson, R., Briggs, S., Ovens, J. & Swales, J. 1999. Michigan Corpus of Academic Spoken English (MICASE).

Simpson, R. & Mendis, D. 2003. A corpus-based study of idioms in academic speech. *37*(3)**,** 419-441.

Sinclair, J. 1991. *Corpus, concordance, collocation,* Oxford, UK, Oxford University Press.

Sinclair, J. 2004. Developing linguistic corpora: A guide to good practice corpus and text—basic principles, tuscan word centre.

Siyanova-Chanturia, A. 2015. Collocation in beginner learner writing: A longitudinal study. *System, 53***,** 148-160.

Siyanova-Chanturia, A. & Spina, S. 2020. Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study. *System, 70*(2)**,** 420-463.

Siyanova, A. & Schmitt, N. 2008. L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review, 64*(3)**,** 429-458.

Snow, C. E. & Uccelli, P. 2009. The challenge of academic language. *The Cambridge handbook of literacy, 112***,** 133.

Staples, S., Egbert, J., Biber, D. & Mcclair, A. 2013. Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for academic purposes, 12*(3)**,** 214-225.

Stengers, H., Boers, F., Housen, A. & Eyckmans, J. 2011. Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association?

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

Szudarski, P. 2017. *Corpus Linguistics for Vocabulary: A Guide for Research*, Taylor & Francis.

Taris, T. W. 2000. *A primer in longitudinal data analysis*, Sage.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L. P., Robson, R., Thabane, M., Giangregorio, L. & Goldsmith, C. H. 2010. A tutorial on pilot studies: the what, why and how. *BMC medical research methodology, 10*(1)**,** 1.

Thomas, J. & Short, M. 1996. *Using corpora for language research: studies in the honour of Geoffrey Leech*, Longman.

Tognini-Bonelli, E. 2001. *Corpus linguistics at work*, J. Benjamins Philadelphia, Amsterdam.

Tono, Y. Learner corpora: design, development and applications. Proceedings of the Corpus Linguistics 2003 conference, 2003. University Centre for Computer Corpus Research on Language Lancaster, 800-809.

Torrance, M. & Galbraith, D. 2006. The processing demands of writing. 67-80.

Tribble, C. 1999. *Writing difficult texts.* University of Lancaster.

Tribble, C. 2011. Revisiting apprentice texts. *Using lexical bundles to investigate expert and apprentice.*

Turabian, K. L. 2013. *A manual for writers of research papers, theses, and dissertations: Chicago style for students and researchers*, University of Chicago Press.

Ucar, S. 2017. A Corpus-based Study on the Use of Three-word Lexical Bundles in the Academic Writing by Native English and Turkish Non-native Writers. *English Language Teaching, 10*(12)**,** 28.

Ucles 2002. International English language testing system.: University of Cambridge Local Examinations Syndicate Cambridge.

Uysal, H. 2012. Argumentation across L1 and L2 Writing: Exploring Cultural Influences and Transfer Issues. *Vigo International Journal of Applied Linguistics,* (9)**,** 133-159.

Varghese, S. A. & Abraham, S. A. 1998. Undergraduates arguing a case. *Journal of Second Language Writing, 7*(3)**,** 287-306.

Vo, S. 2016. Use of Lexical Features in Non-native Academic Writing. *Corpus Linguistics Research, 2***,** 53-53.

Wang, M., Beal, D., Chan, D., Newman, D., Vancouver, J. & Vandenberg, R. 2017. Longitudinal research: A panel discussion on conceptual issues, research design, and statistical techniques. *Work, Aging and Retirement, 3*(1)**,** 1-24.

Wang, M. F. & Bakken, L. L. 2004. An academic writing needs assessment of English-as-a-second-language clinical investigators. *Journal of Continuing Education in the Health Professions, 24*(3)**,** 181-9.

Wei, L. & Moyer, M. G. 2009. *The Blackwell guide to research methods in bilingualism and multilingualism*, John Wiley & Sons.

Wei, Y. & Lei, L. 2011. lexical bundles in the academic writing of advance Chinese EFL Learners. *RELC Journal, 42(2)***,** 463-489.

Wijitsopon, R. 2019. Key multi-word expressions in Thai learner English argumentative essays. *Asian EFL Journal, 23*(6.1)**,** 115-141.

Wilcox, K. C. & Jeffery, J. V. 2014. Adolescents' writing in the content areas: National study results.

Wilson, A. 2013. Embracing Bayes factors for key item analysis in corpus linguistics.

Wingate, U. 2012. 'Argument!'helping students understand what essay writing is about. *Journal of English for academic purposes, 11*(2)**,** 145-154.

Wolfe, C. R. & Britt, M. A. 2008. The locus of the myside bias in written argumentation. *Journal of English for academic purposes, 14*(1)**,** 1-27.

Wolfe, C. R., Britt, M. A. & Butler, J. A. 2009. Argumentation schema and the myside bias in written argumentation. *Written Communication, 26*(2)**,** 183-209.

Wood, D. 2006. Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *Canadian Modern Language Review, 63*(1)**,** 13-33.

Wood, D. 2015. *Fundamentals of formulaic language: An introduction*, Bloomsbury Publishing.

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

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

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

Wray, A. 2008. *Formulaic Language: Pushing the Boundaries,* Oxford, Oxford University Press.

Wray, A. & Perkins, R. 2000. The functions of formulaic language: An integrated model. *Language and Communication, 20*, 1-28.

Yang, Y. 2017. Lexical bundles in argumentative and narrative writings by Chinese EFL learners. *International Journal of English Linguistics, 7*(3), 58-69.

Yoon, C. & Choi, J.-M. 2015a. Lexical bundles in Korean university students' EFL compositions: A comparative study of register and use. *16*(3), 47-69.

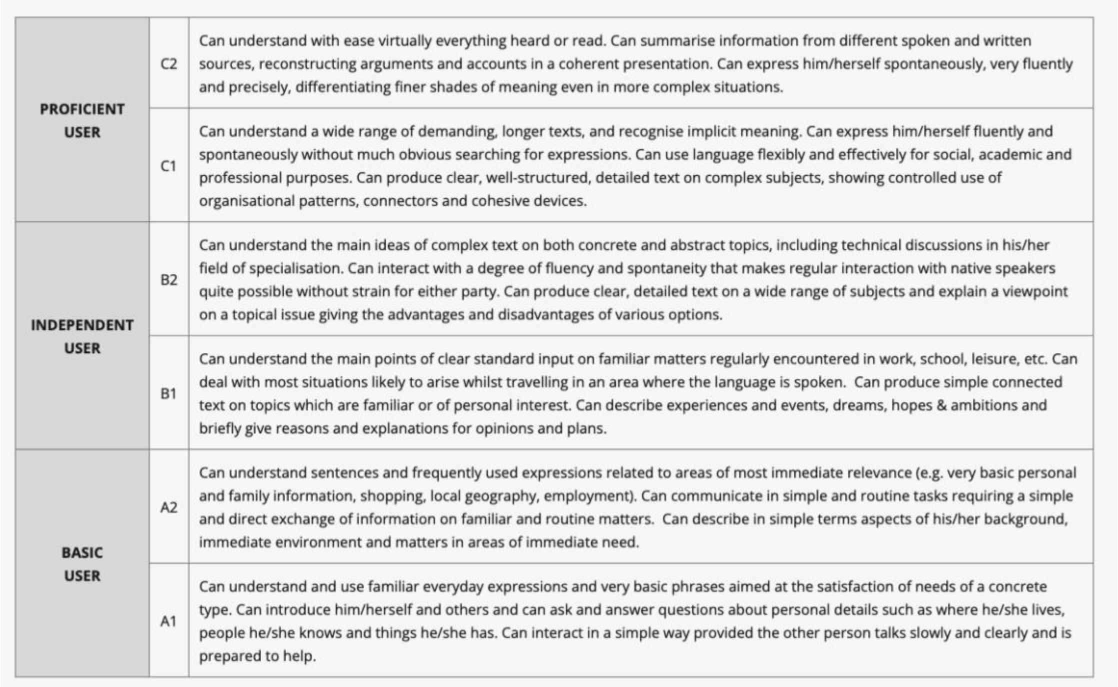
Yoon, C. & Choi, J. 2015b. Lexical bundles in Korean university students' EFL compositions: A comparative study of register and use. *Modern English Education, 16*(3), 47-69.

Youngdong, C. 2020. Linking Adverbials in Academic Writing on English Linguistics by Korean MA Students. *English studies, 40*.

Zhang, Y. & Mi, Y. 2010. Another look at the language difficulties of international students. *Journal of Studies in international Education, 14*(4), 371-388.

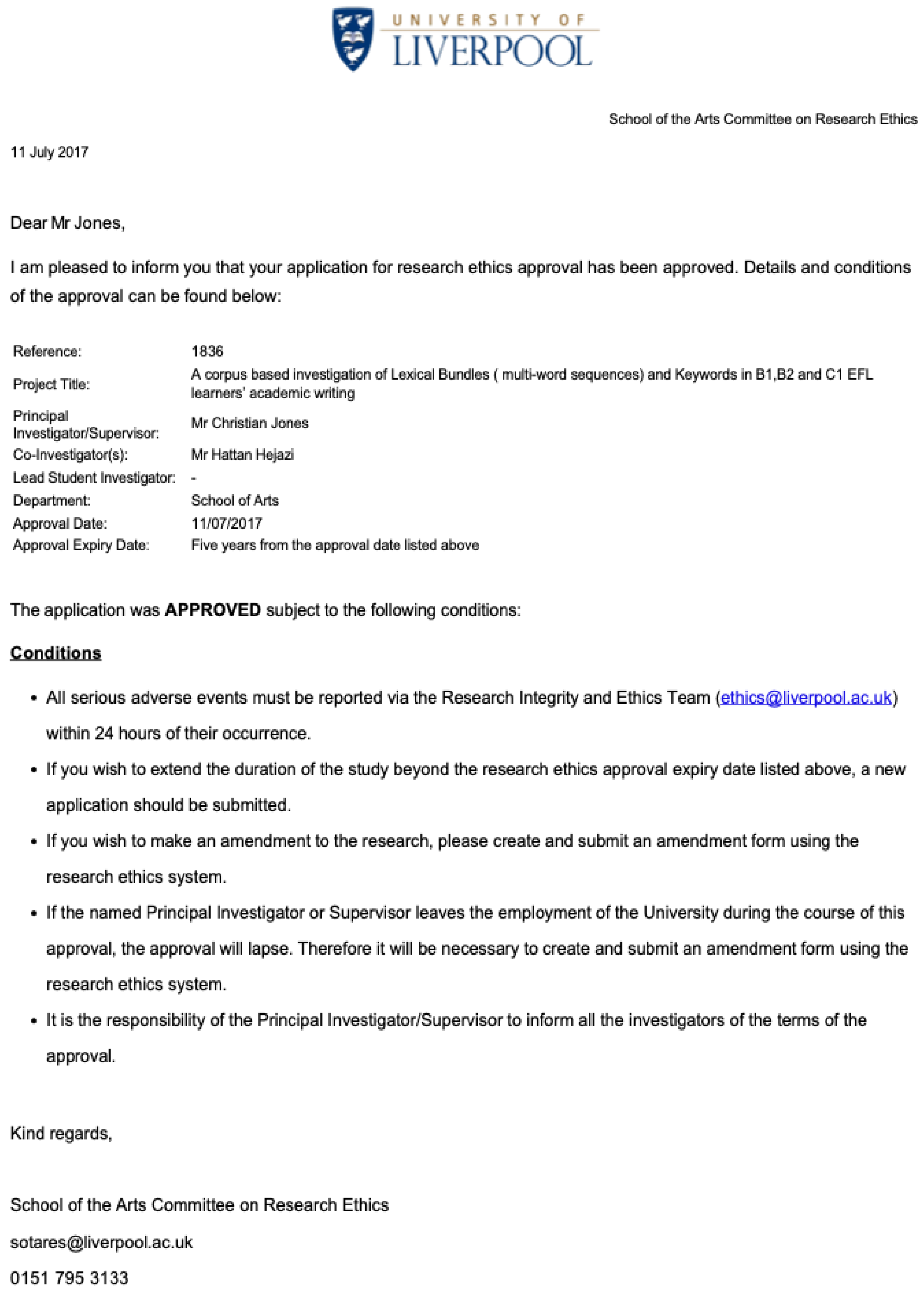
Zimmerman, C. B. 1997. Do reading and interactive vocabulary instruction make a difference? An empirical study. *TESOL quarterly, 31*(1), 121-140.

Appendix A



PROFICIENT USER	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
INDEPENDENT USER	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans.
BASIC USER	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Appendix B



UNIVERSITY OF LIVERPOOL

School of the Arts Committee on Research Ethics

11 July 2017

Dear Mr Jones,

I am pleased to inform you that your application for research ethics approval has been approved. Details and conditions of the approval can be found below:

Reference:	1836
Project Title:	A corpus based investigation of Lexical Bundles (multi-word sequences) and Keywords in B1,B2 and C1 EFL learners' academic writing
Principal Investigator/Supervisor:	Mr Christian Jones
Co-Investigator(s):	Mr Hattan Hejazi
Lead Student Investigator:	-
Department:	School of Arts
Approval Date:	11/07/2017
Approval Expiry Date:	Five years from the approval date listed above

The application was **APPROVED** subject to the following conditions:

Conditions

- All serious adverse events must be reported via the Research Integrity and Ethics Team (ethics@liverpool.ac.uk) within 24 hours of their occurrence.
- If you wish to extend the duration of the study beyond the research ethics approval expiry date listed above, a new application should be submitted.
- If you wish to make an amendment to the research, please create and submit an amendment form using the research ethics system.
- If the named Principal Investigator or Supervisor leaves the employment of the University during the course of this approval, the approval will lapse. Therefore it will be necessary to create and submit an amendment form using the research ethics system.
- It is the responsibility of the Principal Investigator/Supervisor to inform all the investigators of the terms of the approval.

Kind regards,

School of the Arts Committee on Research Ethics

sotares@liverpool.ac.uk

0151 795 3133

Appendix C

Hattan Hejazi
Department of Languages, School of Art
Supervisor: Dr. Christian Jones
University of Liverpool
Tel: +447479785599

Consent Form

Title of Project:
A corpus-based Investigation of Lexical Bundles and Keyness in B1, B2 and C1 ESL Learners' Academic Writing
Name of Researcher: Hattan Hejazi

Please initial box

1. I confirm that I have read and understood the information sheet for the above project and have had the opportunity to ask questions about the use of lexical bundles in the academic writing
2. I understand that my participation is voluntary and that
   I am free to withdraw at any time without giving any reason to the named researcher.
3. I understand that my responses will be used for analysis for this research project.
4. I understand that my responses will remain anonymous.
5. I agree to take part in the above research project.

Name of Participant	Date	Signature
Researcher	Date	Signature

To be signed and dated in presence of the participant

Appendix D

Table A. 1 3-word LBs- B1, B2 and C1

	B1		B2		C1	
	Bundles	Normalised frequncy	Bundles	Normalised frequncy	Bundles	Normalised frequncy
1	a lot of	91	a lot of	98	i think that	399
2	on the other	62	on the other	84	first of all	206
3	one of the	52	there are many	62	second of all	173
4	there are many	46	they do not	58	i believe that	150
5	it is not	44	first of all	54	i think it	121
6	first of all	42	point is that	48	it is a	113
7	i want to	42	i want to	48	to sum up	97
8	there is a	38	in this essay	42	on the other	93
9	in this essay	36	in the world	40	in order to	84
10	the use of	36	in order to	38	his or her	72
11	as well as	34	one of the	36	the opportunity to	64
12	it is a	34	some of the	34	in addition to	56
13	as a result	32	there is no	34	do not have	54
14	in order to	32	do not have	32	i do not	53
15	day by day	30	this essay will	32	i will give	53
16	most of the	30	we need to	32	to support my	53
17	of the world	30	in this world	30	i want to	51
18	that it is	30	there is a	30	in the following	51
19	i do not	28	as a result	28	a lot of	47
20	be able to	28	it is a	28	be able to	47
21	to become a	28	in my opinion	26	one of the	47
22	because it is	28	the use of	26	for several reasons	45
23	in my opinion	28	as well as	24	it is very	43
24	but it is	24	because it is	24	when i was	43
25	in the field	24	it is not	24	as well as	41
26	i had to	24	that it is	24	he or she	41
27	in the world	24	they are not	24	the one hand	41
28	to sum up	24	to sum up	24	they do not	41
29	want to become	24	be able to	22	aspect of this	39
30	we need to	24	because of the	22	does not have	39
31	we do not	22	due to the	22	i did not	39
32	do not have	22	in the past	22	which i will	39
33	point is that	22	is to become	22	a result of	37
34	some of the	22	it is the	22	in this case	37
35	we have to	22	the whole world	22	it is not	37
36	around the world	20	more and more	20	with each other	37
37	because of the	20	in the end	20	i am sure	35
38	i am going	20	are very different	20	to make a	35
39	is the most	20	i do not	20	brings many benefits	33
40	it is the	20	i think that	20	in the future	33
41	it would be	20	i will discuss	20	he does not	31
42	there is no	20	most of the	20	in this essay	31
43	to get a	20	we have to	20	there is no	31
44	to have a	20	with each other	20	point of view	29
45	the end of	18	would like to	20	a chance to	27
46	according to the	18	they have to	18	is a controversial	27
47	due to the	18	may not be	18	learn how to	27
48	i think that	18	is that the	18	there is a	27
49	in the future	18	it is also	18	at the same	25
50	is that it	18	of the world	18	bring many benefits	25
51	things that are	18	to become a	18	different kinds of	25
52	all over the	16	all the time	16	to have a	25
53	it is also	16	and they are	16	in this world	23
54	it is true	16	are aware of	16	role in our	23
55	it is very	16	day by day	16	the help of	23

-- Table continues on the next page --

Table A.1 *continued*

B1		B2		C1	
Bundles	**Normalised frequncy**	**Bundles**	**Normalised frequncy**	**Bundles**	**Normalised frequncy**
56 they are not	16	from each other	16	to find out	23
57 you have to	16	i believe that	16	to get a	23
58 the benefit of	14	in front of	16	a couple of	21
59 can be used	14	it can be	16	all over the	21
60 i am very	14	it would be	16	go to a	21
61 i have been	14	the process of	16	has its own	21
62 in front of	14	there are also	16	help them to	21
63 is going to	14	to each other	16	i have to	21
64 they do not	14	to go to	16	i prefer to	21
65 to be a	14	we do not	16	is open for	21
66 to improve my	14	a long time	14	it would be	21
67 we had a	14	there are no	14	range of interests	21
68 when they are	14	as long as	14	this kind of	21
69 you do not	14	because they are	14	are able to	19
70 large number of	12	but it is	14	huge amount of	19
71 point of view	12	each other and	14	i like to	19
72 i did not	12	is also causing	14	is the one	19
73 the advancement of	12	is the best	14	my point is	19
74 will not be	12	point of view	14	not want to	19
75 a number of	12	the fact that	14	the fact that	19
76 all of the	12	the number of	14	the following points	19
77 and i can	12	the same way	14	the other hand	19
78 and the way	12	there may be	14	those practical benefits	19
79 are able to	12	to know about	14	would have to	19
80 as compared to	12	to take a	14	and present my	18
81 at the beginning	12	we can say	14	and they have	18
82 for one thing	12	will look at	14	by use of	18
83 has also been	12	he has to	12	did not have	18
84 have to make	12	impact on the	12	great opportunity to	18
85 i believe that	12	according to the	12	have their own	18
86 i have to	12	do not get	12	in front of	18
87 i think it	12	do not know	12	it does not	18
88 i will discuss	12	if you are	12	it means that	18
89 i would like	12	in this way	12	most of the	18
90 in the same	12	is a very	12	should not be	18
91 is that the	12	is going to	12	the amount of	18
92 large amount of	12	is the most	12	they need to	18
93 make use of	12	over the world	12	to be a	18
94 should not be	12	so we should	12	we need to	18
95 the level of	12	that there are	12	and it is	16
96 there are some	12	that there is	12	because it is	16
97 they have to	12	the main reason	12	from my opinion	16
98 think it is	12	the same time	12	go to the	16
99 to get the	12	there are some	12	have a great	16
100 to know the	12	there are very	12	i am not	16
101 when i was	12	they should be	12	i will list	16
102 between these two	10	they spend their	12	in favour of	16
103 it should be	10	to say that	12	is the most	16
104 depends on the	10	to the world	12	it is like	16
105 we are not	10	will not be	12	it is the	16
106 can say that	10	look at the	10	they want to	16
107 is caused by	10	we want to	10	to learn more	16
108 a negative effect	10	take care of	10	what kind of	16
109 a long time	10	i think it	10	will have to	16
110 a wide range	10	a way of	10	a few years	14

-- Table continues on the next page --

Table A.1 *continued*

B1		B2		C1	
Bundles	**Normalised frequncy**	**Bundles**	**Normalised frequncy**	**Bundles**	**Normalised frequncy**
111 and many other	10	and it will	10	and gain more	14
112 as they are	10	and many more	10	and i am	14
113 can be a	10	and this is	10	be required to	14
114 different types of	10	around the world	10	because it will	14
115 does not mean	10	at the end	10	difficult to imagine	14
116 for me to	10	be in the	10	great impact on	14
117 for the betterment	10	can be seen	10	have a few	14
118 has its own	10	do not need	10	i base my	14
119 have access to	10	effect on the	10	i have a	14
120 i used to	10	his or her	10	i must confess	14
121 i will do	10	i have to	10	i think the	14
122 if they are	10	if it is	10	i will not	14
123 in the end	10	if you want	10	i would like	14
124 in the first	10	in other words	10	know each other	14
125 is also a	10	in the form	10	not care about	14
126 is not a	10	in their own	10	reasons why i	14
127 is not only	10	is also a	10	the best way	14
128 it can be	10	is caused by	10	the development of	14
129 it is more	10	it does not	10	the problem of	14
130 it was a	10	it is an	10	the real world	14
131 just because of	10	it is true	10	they will be	14
132 more and more	10	it is very	10	to each other	14
133 need to be	10	need to be	10	to make their	14
134 not want to	10	now i will	10	who think that	14
135 of this is	10	one another and	10	better chance to	12
136 part of the	10	so they are	10	but it is	12
137 some of them	10	that should be	10	can be taken	12
138 that i will	10	the amount of	10	focus on the	12
139 that is the	10	the impact of	10	great way to	12
140 that is why	10	the loss of	10	has a great	12
141 that there are	10	the name of	10	how to use	12
142 that there is	10	the need to	10	i can stand	12
143 the causes of	10	the world and	10	i think i	12
144 the impact of	10	there are several	10	in my opinion	12
145 the number of	10	there is also	10	is not so	12
146 the process of	10	this is not	10	is that a	12
147 there are more	10	this is the	10	is that it	12
148 there are no	10	to be a	10	it is really	12
149 this essay will	10	to have a	10	move from one	12
150 this is a	10	to see the	10	one more reason	12
151 this is not	10	vital role in	10	others think that	12
152 to do it	10	we must use	10	say that the	12
153 to do something	10	we talk about	10	some of the	12
154 to do the	10	you have to	10	that i will	12
155 to improve the	10	be found in	8	the ability to	12
156 to say that	10	that is not	8	the beginning of	12
157 what is happening	10	what is the	8	the issue about	12
158 which is not	10	have only one	8	the main reason	12
159 will be a	10	the world is	8	these practical benefits	12
160 would love to	10	acts as a	8	this is the	12
161 that i can	8	all of the	8	through their entire	12
162 more likely to	8	all of us	8	to add that	12
163 not that much	8	and in many	8	to gain more	12
164 know how to	8	and it is	8	to learn from	12
165 passage of time	8	and many other	8	to make our	12

-- Table continues on the next page --

Table A.1 *continued*

	B1		B2		C1	
	Bundles	**Normalised frequncy**	**Bundles**	**Normalised frequncy**	**Bundles**	**Normalised frequncy**
166	a great deal	8	and we must	8	to reach their	12
167	across the world	8	as compared to	8	what they like	12
168	after a while	8	as they are	8	a few centuries	10
169	and i am	8	as we know	8	a part of	10
170	and i have	8	available in the	8	and did not	10
171	and i will	8	but there are	8	and how to	10
172	and many more	8	but they have	8	are the best	10
173	and that is	8	can be very	8	avoid them next	10
174	and there are	8	does not like	8	be a great	10
175	and they are	8	due to which	8	be ready to	10
176	are a major	8	for many years	8	be spent on	10
177	argue that the	8	has become a	8	be up to	10
178	at that time	8	has its own	8	can spend more	10
179	being able to	8	have to be	8	day by day	10
180	but also the	8	have to face	8	each other and	10
181	but no one	8	i could not	8	essential for a	10
182	do not need	8	if they are	8	for a long	10
183	from each other	8	in addition to	8	for instance if	10
184	has become a	8	in case of	8	for instance my	10
185	helps us to	8	in conclusion i	8	have to do	10
186	i also have	8	in conclusion our	8	him or her	10
187	i am doing	8	in this respect	8	i could not	10
188	i am in	8	increase in the	8	i know that	10
189	i found it	8	instead of being	8	i think a	10
190	i hope i	8	is better than	8	if it is	10
191	i will have	8	is busy in	8	in a few	10
192	in addition to	8	is given by	8	in most cases	10
193	in conclusion i	8	is like a	8	in the same	10
194	in many ways	8	is much better	8	it can be	10
195	in other words	8	is not a	8	it is impossible	10
196	in our daily	8	is not always	8	it is rather	10
197	in spite of	8	it is my	8	it was a	10
198	in the last	8	it is necessary	8	learns how to	10
199	is a very	8	it is obvious	8	make conclusions and	10
200	is better for	8	it is often	8	more and more	10
201	known as the	8	it means that	8	need to know	10
202	means that the	8	it will be	8	of them are	10
203	more than one	8	large amount of	8	one of them	10
204	not only this	8	large number of	8	option has its	10
205	over the last	8	make sure that	8	reason for this	10
206	positive aspects of	8	of course it	8	should have the	10
207	such as a	8	pay attention to	8	should spend more	10
208	tell them that	8	problems such as	8	that they are	10
209	that can be	8	than ever before	8	the importance of	10
210	that i am	8	that it has	8	the number of	10
211	that i have	8	the development of	8	the reason why	10
212	that they are	8	the effect of	8	the same way	10
213	that we have	8	the quality of	8	the whole world	10
214	the fact that	8	that they are	8	there are many	10
215	the importance of	8	the increase in	8	there are some	10
216	the increase in	8	the increase of	8	they are not	10
217	the lack of	8	the reasons for	8	they like to	10
218	the real world	8	their tastes in	8	to arrange their	10
219	the result of	8	there are not	8	to learn new	10
220	the same as	8	there is something	8	to make the	10

-- Table continues on the next page --

Table A.1 *continued*

	B1		B2		C1	
	Bundles	**Normalised frequncy**	**Bundles**	**Normalised frequncy**	**Bundles**	**Normalised frequncy**
221	the world because	8	there is very	8	to spend their	10
222	there are not	8	there will be	8	we have to	10
223	there are several	8	these are the	8	what they want	10
224	there should be	8	they are both	8	with me that	10
225	they are a	8	they used to	8	a waste of	8
226	they are the	8	they want to	8	ability to think	8
227	they need to	8	this is a	8	access to the	8
228	this is because	8	this means that	8	according to their	8
229	this is the	8	time there are	8	almost impossible to	8
230	this kind of	8	to do so	8	along with the	8
231	to find out	8	to do something	8	also it is	8
232	to find a	8	to make the	8	and do not	8
233	to get to	8	to think about	8	and enjoy the	8
234	to make it	8	try to make	8	and even make	8
235	to reduce the	8	we see that	8	and it was	8
236	to take the	8	when i was	8	and make a	8
237	to think about	8	which is a	8	and try to	8
238	together in the	8	it can also	6	around the world	8
239	we also need	8	a major role	6	as a whole	8
240	we can enjoy	8	a variety of	6	as long as	8
241	we can give	8	a wide range	6	as they are	8
242	we have a	8	a world with	6	at this moment	8
243	we see that	8	able to see	6	be better prepared	8
244	we today have	8	about how to	6	be the best	8
245	we want to	8	all of these	6	because of the	8
246	what they are	8	all these problems	6	because of their	8
247	which is the	8	also there is	6	because they are	8
248	which means that	8	and also in	6	because they will	8
249	with the help	8	and do not	6	both of these	8
250	would not be	8	and try to	6	do many things	8
251	at the same	6	and we are	6	does not make	8
252	a chance to	6	another reason is	6	each of these	8
253	a collection of	6	are based on	6	feel more secure	8
254	a part of	6	are lack of	6	for the future	8
255	a sense of	6	are made of	6	from each other	8
256	a way that	6	are very simple	6	has some negative	8
257	about the world	6	as we all	6	have common interests	8
258	all the time	6	be used for	6	he wants to	8
259	also like to	6	but on the	6	helps them to	8
260	although it is	6	can be a	6	how to make	8
261	although there are	6	can be done	6	however others believe	8
262	an opportunity to	6	can learn about	6	i belief that	8
263	and at the	6	claim that the	6	i mentioned above	8
264	and do not	6	defined as the	6	i need to	8
265	and i also	6	different from the	6	i think many	8
266	and in a	6	does not have	6	if i had	8
267	and it is	6	due to this	6	if one wants	8
268	and more efficient	6	each and everything	6	in the case	8
269	and the best	6	each other through	6	in the world	8
270	and the world	6	for many reasons	6	is a perfect	8
271	and then i	6	for more than	6	is easier to	8
272	and try to	6	go to the	6	is going to	8
273	and what is	6	has a positive	6	is more enjoyable	8
274	another difference is	6	however the most	6	is not as	8
275	are available for	6	however there are	6	is that the	8

-- Table continues on the next page --

Table A.1 *continued*

B1		B2		C1	
Bundles	**Normalised frequncy**	**Bundles**	**Normalised frequncy**	**Bundles**	**Normalised frequncy**
276 are facing a	6	i am very	6	it is difficult	8
277 are going to	6	i feel that	6	it is my	8
278 are located in	6	i have seen	6	many things to	8
279 are not very	6	i think this	6	most likely will	8
280 as much as	6	i will examine	6	need to be	8
281 at the time	6	i will say	6	of a new	8
282 because i have	6	in addition the	6	of that importance	8
283 because it was	6	in the field	6	on the whole	8
284 because there are	6	in favour of	6	others prefer to	8
285 because there is	6	in recent years	6	spend most of	8
286 but in the	6	in some cases	6	take care of	8
287 but there are	6	in terms of	6	that can be	8
288 by reducing the	6	in the future	6	that will be	8
289 can see that	6	in the last	6	the issue whether	8
290 chance to know	6	in the right	6	the pace of	8
291 create a positive	6	in the way	6	the rest of	8
292 difficult for me	6	integral part of	6	the result of	8
293 do not get	6	is a big	6	the way of	8
294 due to which	6	is also not	6	there are plenty	8
295 especially with the	6	is also very	6	there are two	8
296 even though i	6	is available in	6	these two options	8
297 for instance i	6	is that they	6	they are needed	8
298 for me and	6	is that we	6	they become more	8
299 for the sake	6	is the only	6	they have to	8
300 for their own	6	is to be	6	they will not	8
301 from that point	6	is too much	6	to choose the	8
302 go to a	6	is very interesting	6	to do it	8
303 going to the	6	is very simple	6	to do something	8
304 has helped the	6	issue is that	6	to find a	8
305 have their own	6	it did not	6	to follow the	8
306 have to be	6	it has been	6	to learn and	8
307 have to spend	6	it is clear	6	to listen to	8
308 he used to	6	it is in	6	to make new	8
309 how to use	6	it is more	6	to meet new	8
310 however i tend	6	it is probably	6	to save some	8
311 however there are	6	it is really	6	to the next	8
312 i am interested	6	it seems to	6	we are not	8
313 i am not	6	just a few	6	what is happening	8
314 i believe it	6	led to an	6	when he can	8
315 i decided to	6	looking at the	6	who want to	8
316 i had no	6	many of these	6	would not be	8
317 i know that	6	may be a	6	a bunch of	6
318 i really like	6	much of their	6	a large amount	6
319 i think i	6	my best to	6	a little bit	6
320 i was not	6	no doubt that	6	a way of	6
321 i will make	6	no one can	6	about each other	6
322 i will try	6	not have a	6	about what is	6
323 in a productive	6	now if we	6	all mentioned above	6
324 in addition the	6	now it is	6	all of them	6
325 in each area	6	of being an	6	all things done	6
326 in fact i	6	of them are	6	also has some	6
327 in my view	6	out of the	6	and does not	6
328 in the long	6	owing to the	6	and make the	6
329 in the middle	6	problem is that	6	and of cause	6
330 in which i	6	reduced to none	6	and then i	6

-- Table continues on the next page --

Table A.1 *continued*

B1		B2		C1	
Bundles	**Normalised frequncy**	**Bundles**	**Normalised frequncy**	**Bundles**	**Normalised frequncy**
331 is better than	6	respect for the	6	and they are	6
332 is full of	6	should be taken	6	another part of	6
333 is in the	6	should have the	6	are very different	6
334 is not available	6	so as to	6	as far as	6
335 is not enough	6	so that i	6	as for me	6
336 is the main	6	solve their problems	6	as soon as	6
337 is thrown in	6	some of them	6	at that time	6
338 is to have	6	such as the	6	believe that the	6
339 it also has	6	take action to	6	benefits in the	6
340 it does not	6	that cannot be	6	break limits and	6
341 it has a	6	that i have	6	bring only benefits	6
342 it is easy	6	that is why	6	brought many benefits	6
343 it is much	6	that they can	6	be proud of	6
344 it is often	6	that they cannot	6	be satisfied with	6
345 it is quite	6	that you have	6	be supportive and	6
346 it is up	6	the condition of	6	can do better	6
347 it will be	6	the importance of	6	can have more	6
348 large amounts of	6	the kind of	6	can say that	6
349 learning about the	6	the most obvious	6	cannot afford to	6
350 led to the	6	the need for	6	constantly improve their	6
351 most of our	6	the problems of	6	create a new	6
352 no matter what	6	the role of	6	depends on a	6
353 no one can	6	the safety of	6	did not know	6
354 of it like	6	the world which	6	different types of	6
355 of the final	6	there are more	6	do not feel	6
356 one of my	6	there are other	6	does not feel	6
357 rate of the	6	there are so	6	does not give	6
358 role in our	6	there is not	6	explain bellow i	6
359 seems to be	6	there was a	6	for a while	6
360 should have to	6	these things have	6	get used to	6
361 side of this	6	they are a	6	great amount of	6
362 so i can	6	they are also	6	great means of	6
363 so we should	6	they can do	6	has many benefits	6
364 some of these	6	thing is that	6	has nothing to	6
365 take care of	6	think about the	6	have to go	6
366 that no one	6	this can be	6	he can chart	6
367 that they have	6	this has a	6	he is interested	6
368 the amount of	6	this type of	6	he must be	6
369 the first step	6	this world and	6	he was not	6
370 the first time	6	to achieve the	6	help each other	6
371 the most obvious	6	to achieve their	6	him as a	6
372 the need for	6	to all of	6	how to save	6
373 the possibility of	6	to all the	6	however i must	6
374 the purpose of	6	to be successful	6	i am going	6
375 the rate of	6	to begin with	6	i can do	6
376 the time i	6	to do in	6	i can state	6
377 the world and	6	to find a	6	i cannot say	6
378 the world is	6	to form a	6	i state my	6
379 the world they	6	to make sure	6	i think both	6
380 then you will	6	to spend a	6	i think these	6
381 there are different	6	to the next	6	i was required	6
382 there is always	6	to this problem	6	i would be	6
383 there is an	6	used for testing	6	imagine that a	6
384 these kinds of	6	we all know	6	impact on the	6
385 they have been	6	we all need	6	in addition i	6

-- Table continues on the next page --

Table A.1 *continued*

	B1		B2		C1	
	Bundles	**Normalised frequncy**	**Bundles**	**Normalised frequncy**	**Bundles**	**Normalised frequncy**
386	they have no	6	we are th	6	in its turn	6
387	thing that is	6	we should start	6	in many ways	6
388	think that if	6	what to do	6	in some cases	6
389	this advancement in	6	when it is	6	in the next	6
390	this means that	6	which is very	6	is great because	6
391	throughout the world	6	who are not	6	is like a	6
392	to achieve their	6	you can find	6	is no longer	6
393	to be an	6	you know that	6	is not a	6
394	to be in	6	you need to	6	is not the	6
395	to be more	6			is on the	6
396	to become more	6			is that i	6
397	to create a	6			is that they	6
398	to figure out	6			is very busy	6
399	to go to	6			it can save	6
400	to help the	6			it gives the	6
401	to keep in	6			it is also	6
402	to learn more	6			it is an	6
403	to look for	6			it is much	6
404	to make a	6			it is worth	6
405	to reach the	6			it must be	6
406	to the situation	6			it was the	6
407	today relates to	6			looking for a	6
408	use of the	6			looking for the	6
409	used in the	6			make a decision	6
410	very popular in	6			many beautiful moments	6
411	was a little	6			more about the	6
412	was able to	6			more interesting and	6
413	we can do	6			move on to	6
414	we get from	6			much attention to	6
415	we have developed	6			much better than	6
416	we learned how	6			of all he	6
417	we will be	6			of the world	6
418	went back to	6			one can gain	6
419	what we have	6			one more thing	6
420	whether it is	6			ones such as	6
421	which are not	6			options have their	6
422	which is a	6			otherwise if a	6
423	which we could	6			pass down their	6
424	why it is	6			ready to help	6
425	will always be	6			same amount of	6
426	will say why	6			second reason for	6
427	with each other	6			sense of humour	6
428					she or he	6
429					so if a	6
430					so it is	6
431					so when they	6
432					some kind of	6
433					some of them	6
434					support this idea	6
435					that i am	6
436					that i can	6
437					that is why	6
438					that make our	6
439					that the first	6
440					that there are	6

-- Table continues on the next page --

Table A.1 *continued*

	B1		B2		C1	
	Bundles	**Normalised frequncy**	**Bundles**	**Normalised frequncy**	**Bundles**	**Normalised frequncy**
441					that we need	6
442					the better one	6
443					the end of	6
444					the idea about	6
445					the majority of	6
446					the most obviously	6
447					the reason behind	6
448					the right to	6
449					they are in	6
450					they can choose	6
451					they did not	6
452					they have more	6
453					they should have	6
454					they try to	6
455					they will have	6
456					they with great	6
457					things such as	6
458					this way of	6
459					to be more	6
460					to choose from	6
461					to choose what	6
462					to compete with	6
463					to deal with	6
464					to decrease the	6
465					to do so	6
466					to do things	6
467					to get to	6
468					to learn about	6
469					to meet their	6
470					to prepare for	6
471					to see the	6
472					to spend some	6
473					to stay at	6
474					to understand the	6
475					try to keep	6
476					very interesting and	6
477					way to get	6
478					we go to	6
479					we need more	6
480					what he is	6
481					what it is	6
482					when it comes	6
483					who do not	6
484					who succeeded in	6
485					will be more	6
486					will be the	6
487					will most likely	6

Appendix E

Table B. 1 Shared Bundles across ESL learners' sub-corpora.

Shared bundles	B1	B2	C1	Shared bundles	B1	B2	C1
a lot of	91	98	47	it would be	20	16	21
and do not	6	6	8	most of the	30	20	18
and it is	6	8	16	need to be	10	10	8
and they are	8	16	6	of the world	30	18	6
and try to	6	6	8	point of view	12	14	29
around the world	20	10	8	some of them	10	6	6
as they are	10	8	8	take care of	6	10	8
as well as	34	24	41	that is why	10	6	6
be able to	28	22	47	that there are	10	12	6
because it is	28	24	16	that they are	8	8	10
because of the	20	22	8	the amount of	6	10	18
but it is	24	14	12	the fact that	8	14	19
day by day	30	16	10	the importance of	8	6	10
do not have	22	32	54	there are many	46	62	10
first of all	42	54	206	there are some	12	12	10
from each other	8	16	8	there is a	38	30	27
has its own	10	8	21	there is no	20	6	31
I believe that	12	16	150	they are not	16	24	10
I do not	28	20	53	they do not	14	58	41
I have to	12	10	21	they have to	12	18	8
I think it	12	10	121	this is the	8	10	12
I think that	18	20	399	to be a	14	10	18
I want to	42	48	51	to do something	10	8	8
in addition to	8	8	56	to find a	8	6	8
in front of	14	16	18	to have a	20	10	25
in my opinion	28	26	12	to sum up	24	24	97
in order to	32	38	84	we have to	22	20	10
in the future	18	6	33	we need to	24	32	18
in the world	24	40	8	when I was	12	18	34
in this essay	36	42	31	with each other	6	20	37
is going to	14	12	8	all over the world	10	3	10
is not a	10	8	6	as a result of	12	5	8
is that the	12	6	6	at the same time	6	3	10
is the most	20	12	16	do not want to	6	3	8
it can be	10	16	10	first of all I	6	3	10
it does not	6	10	18	I bdo not think	12	3	6
it is a	34	10	113	I would like to	10	5	16
it is also	16	18	6	in this essay I	32	16	42
it is not	44	24	37	is one of the	10	5	16
it is the	20	22	16	on the other hand	60	29	78
it is very	16	10	43	one of the most	16	8	16

Appendix F

Table C. 1 Shared Bundles across ESL learners' sub-corpora and the BAWE (i.e., reference corpus)

Rrank	BAWE bundles	Normalised	B1/100,000	100,000	100,000
1	a collection of	2	6	0	0
2	a great deal	2	8	0	0
3	a long time	2	10	14	0
4	a lot of	5	91	98	47
5	a number of	18	12	0	0
6	a part of	6	6	0	10
7	a sense of	23	6	0	0
8	according to the	6	18	12	8
9	all of the	5	12	8	0
10	although it is	3	6	0	0
11	and it is	9	6	8	16
12	and they are	2	8	16	6
13	argue that the	2	8	0	0
14	as a result	16	32	28	0
15	as much as	2	6	0	0
16	as they are	6	10	8	8
17	as well as	28	34	24	41
18	at that time	3	8	0	6
19	at the time	6	6	0	0
20	be able to	14	28	22	47
21	because it is	6	28	24	16
22	because of the	6	20	22	8
23	but also the	2	8	0	0
24	but in the	3	6	0	0
25	but it is	6	24	14	12
26	can be used	7	14	0	0
27	different types of	3	10	0	6
28	do not have	4	22	32	54
29	due to the	23	18	22	0
30	have to be	4	6	8	0
31	however there are	2	6	6	0
32	i am not	2	6	0	16
33	i decided to	3	6	0	0
34	i did not	2	12	0	39
35	i had to	3	24	0	0
36	i have to	2	12	10	21
37	i think that	3	18	20	399
38	in addition the	3	6	6	0
39	in addition to	3	8	8	56
40	in front of	4	14	16	18
41	in order to	35	32	38	84
42	in other words	7	8	10	0
43	in spite of	4	8	0	0
45	in the end	2	10	20	0
46	in the first	15	10	0	0
47	in the future	3	18	6	33
48	in the last	4	8	6	0

-- Table continues on the next page --

Table C.1 *continued*

Rrank	BAWE bundles	Normalised	B1/100,000	100,000	100,000
49	in the middle	4	6	0	0
50	in the same	10	12	0	10
51	in the world	6	24	40	8
52	is a very	3	8	12	0
53	is also a	6	10	10	0
54	is going to	3	14	12	8
55	is in the	5	6	0	0
56	is not a	10	10	8	6
57	is not only	6	10	0	0
58	is that it	4	18	0	12
59	is that the	7	12	18	8
60	is the most	3	20	12	16
61	it can be	14	10	16	0
62	it does not	6	6	10	18
63	it is a	15	34	28	113
64	it is also	10	16	18	6
65	it is more	2	10	6	0
66	it is not	19	44	24	37
67	it is often	2	6	8	0
68	it is the	16	20	22	16
69	it is very	2	16	10	43
70	it should be	3	10	0	0
71	it was a	3	10	0	10
72	it will be	3	6	8	0
73	it would be	7	20	16	21
74	known as the	4	6	0	0
75	led to the	2	6	0	0
76	more and more	2	10	20	10
77	more likely to	7	8	0	0
78	more than one	4	8	0	0
79	most of the	7	30	20	18
80	need to be	6	10	10	8
81	of the world	6	30	18	6
82	on the other	19	62	84	0
83	one of the	26	52	36	47
84	part of the	21	10	0	0
85	point of view	8	12	14	29
86	seems to be	6	6	0	0
87	should not be	3	12	0	18
88	some of the	10	22	34	12
89	such as a	2	8	0	0
90	that can be	6	8	0	8
91	that i have	3	8	6	0
92	that is the	2	10	0	0
93	that it is	22	30	24	0
94	that there are	9	10	12	6
95	that there is	18	10	12	0

-- Table continues on the next page --

Table C.1 *continued*

Rrank	BAWE bundles	Normalised	B1/100,000	100,000	100,000
96	that they are	8	10	12	6
97	that they have	3	6	0	0
98	the amount of	3	6	10	18
99	the end of	24	18	0	6
100	the fact that	35	8	14	19
101	the importance of	20	8	6	10
102	the lack of	8	8	0	0
103	the need for	4	6	6	0
104	the number of	8	10	14	10
105	the possibility of	3	6	0	0
106	the process of	8	10	16	0
107	the purpose of	6	6	0	0
108	the real world	3	8	0	14
109	the same as	2	8	0	0
110	the use of	35	36	26	0
111	there are many	7	46	62	10
112	there are some	2	12	12	10
113	there is a	26	38	30	27
114	there is an	4	6	0	0
115	there is no	24	20	22	31
116	they are not	3	16	24	10
117	they do not	6	14	58	41
118	this essay will	6	10	32	0
119	this is a	11	10	8	0
120	this is because	5	8	0	0
121	this is not	7	10	10	0
122	this is the	9	8	10	12
123	this kind of	3	8	0	21
124	to be a	15	14	10	18
125	to be an	2	6	0	0
126	to be in	3	6	0	0
127	to be more	6	6	0	6
128	to create a	9	6	0	0
129	to find a	3	8	6	8
130	to find out	3	8	0	23
131	to have a	4	20	10	25
132	to make a	4	6	0	35
133	to make it	3	8	0	0
134	to say that	5	10	12	0
135	to sum up	2	24	24	97
136	used in the	9	6	0	0
137	was able to	3	6	0	0
138	we do not	2	22	16	0
139	what they are	2	8	0	0
140	when they are	3	14	0	0
141	whether it is	2	6	0	0
142	which is a	4	6	8	0

-- Table continues on the next page --

Table C.1 *continued*

Rrank	BAWE bundles	Normalised	B1/100,000	100,000	100,000
143	which is not	3	10	0	0
145	which is the	3	8	0	0
146	why it is	2	6	0	0
147	with each other	4	6	20	37
148	would not be	4	8	0	8
149	a variety of	8	0	6	0
150	a way of	3	0	10	6
151	acts as a	2	0	8	0
152	all of these	4	0	6	0
153	and this is	5	0	10	0
154	as long as	2	0	14	8
155	because they are	2	0	14	8
156	can be seen	20	0	10	0
157	defined as the	2	0	6	0
158	do not know	2	0	12	0
159	does not have	3	0	6	39
160	each other and	2	0	14	10
161	effect on the	4	0	10	0
162	his or her	2	0	10	72
163	i feel that	2	0	6	0
164	i will examine	2	0	6	0
165	if it is	2	0	10	10
166	in favour of	4	0	6	16
167	in terms of	23	0	6	0
168	in the past	4	0	22	0
169	in the way	6	0	6	0
170	in their own	2	0	10	0
171	in this way	9	0	12	0
172	is not always	2	0	8	0
173	is that they	2	0	6	6
174	is the only	2	0	6	0
175	is to be	4	0	22	0
176	it has been	13	0	6	0
177	it is an	5	0	10	6
178	it is clear	9	0	6	0
179	it is in	3	0	6	0
180	it is necessary	5	0	8	0
181	it seems to	2	0	6	0
182	looking at the	6	0	6	0
183	may be a	3	0	6	0
184	may not be	4	0	18	0
185	out of the	6	0	6	0
186	so as to	2	0	6	0
187	such as the	17	0	6	0
189	that it has	2	0	8	0
190	that they can	3	0	6	0
191	the development of	10	0	6	14

-- Table continues on the next page --

Table C.1 *continued*

Rrank	BAWE bundles	Normalised	B1/100,000	100,000	100,000
192	the effect of	7	0	8	0
193	the loss of	3	0	10	0
194	the need to	4	0	10	0
195	the role of	16	0	6	0
196	there are also	3	0	16	0
197	there is also	6	0	10	0
198	there is not	2	0	6	0
199	there was a	6	0	6	0
200	this can be	8	0	6	0
201	this type of	4	0	6	0
202	to become a	2	0	18	0
203	to do so	3	0	8	6
204	to each other	5	0	16	14
205	to make the	6	0	8	10
206	to see the	2	0	10	6
207	will look at	2	0	14	0
208	would like to	3	0	20	0
209	according to their	2	0	0	8
210	all of them	2	0	0	6
211	along with the	2	0	0	8
212	are able to	10	0	0	19
213	as a whole	8	0	0	8
214	as far as	5	0	0	6
215	both of these	2	0	0	8
216	focus on the	7	0	0	12
217	he does not	4	0	0	31
218	i have a	3	0	0	14
219	in the following	5	0	0	51
220	in the next	3	0	0	6
221	in this case	11	0	0	37
222	is no longer	5	0	0	6
223	is not as	2	0	0	8
224	is not the	6	0	0	6
225	is on the	2	0	0	6
226	it must be	2	0	0	6
227	it was the	3	0	0	6
228	of a new	2	0	0	8
229	some kind of	3	0	0	6
230	that the first	2	0	0	6
231	the ability to	6	0	0	12
232	the beginning of	14	0	0	12
233	the help of	3	0	0	23
234	the majority of	10	0	0	6
235	the opportunity to	2	0	0	64
236	the rest of	6	0	0	8
237	there are two	4	0	0	8
238	they are in	2	0	0	6

-- Table continues on the next page --

Table C.1 *continued*

Rrank	BAWE bundles	Normalised	B1/100,000	100,000	100,000
239	to deal with	2	0	0	6
240	to understand the	4	0	0	6
241	what he is	2	0	0	6

Appendix G

Table D. 1 Keybundle-B1

KeyBundles	Keyness	KeyBundles	Keyness
a lot of	130.76	to sum up	29.9
first of all	102.16	to get the	29.19
i want to	102.16	they have to	29.19
day by day	72.97	to know the	29.19
i do not	68.11	when i was	29.19
in my opinion	68.11	will not be	29.19
want to become	58.38	i believe that	29.19
in the field	58.38	have to make	29.19
we need to	58.38	has also been	29.19
point is that	53.51	i think it	29.19
we have to	53.51	i will discuss	29.19
to get a	48.65	and i can	29.19
i am going	48.65	for one thing	29.19
around the world	48.65	as compared to	29.19
things that are	43.78	the advancement of	29.19
all over the	38.92	make use of	29.19
you have to	38.92	large number of	29.19
there are many	38.81	large amount of	29.19
to become a	34.42	in the field of	58.38
i am very	34.05	first of all it	43.78
the benefit of	34.05	i am going to	43.78
you do not	34.05	in this essay i	37.65
i have been	34.05	anywhere in the world	29.19
to improve my	34.05	a second point is	29.19
we had a	34.05	i do not think	29.19
		increasing day by day	29.19

Table D. 2. Keybundle-B2

KeyBundles	Log_L	Key word	Log_L
a lot of	143.88	in a good	34.25
first of all	132.11	in cities are	34.25
point is that	117.43	in villages people	34.25
i want to	117.43	in the society	34.25
global warming is	88.07	is the best	34.25
people do not	83.18	is also causing	34.25
the people of	83.18	they are aware	34.25
we need to	78.29	there may be	34.25
social media is	73.39	there are no	34.25
in this world	73.39	do not have	33.47
aim of life	68.5	to sum up	30.21
they do not	66.74	the ozone layer	29.36
there are many	64.65	will not be	29.36
in my opinion	63.61	there are very	29.36
of social media	63.61	to the world	29.36
of global warming	58.71	they are very	29.36
essay i will	58.18	third point is	29.36
the whole world	53.82	they should be	29.36
village life is	53.82	way of life	29.36
my father is	53.82	they spend their	29.36
is to become	53.82	for their children	29.36
to become an	53.82	he has to	29.36
my mother is	53.82	do not want	29.36
on the other	50.13	for and against	29.36
the temperature of	48.93	impact on the	29.36
we have to	48.93	in my country	29.36
are very different	48.93	i used to	29.36
social media has	48.93	if you are	29.36
i do not	48.93	another point is	29.36
i will discuss	48.93	arguments for and	29.36
this essay i	46.01	and city life	29.36
the other hand	44.91	animal testing is	29.36
are a lot	44.04	do not get	29.36
they have to	44.04	do not think	29.36
aim in life	44.04	available in cities	29.36
on social media	44.04	can communicate with	29.36
the social media	44.04	over the world	29.36
become an engineer	44.04	so we should	29.36
in my life	44.04	of the country	29.36
the people who	39.14	of the village	29.36
life is to	39.14	the global warming	29.36
of village life	39.14	the main reason	29.36
people who are	39.14	temperature of the	29.36
people of cities	39.14	the children will	29.36
temperature of earth	39.14	my aim of	29.36
stage of the	39.14	my parents are	29.36
of the process	39.14	is very important	29.36
in our society	39.14	my aim in	29.36
are aware of	39.14	not want to	29.36
to go to	39.14	of life is	29.36
day by day	39.14	my sister and	29.36
i believe that	39.14	not be able	29.36
from each other	39.14	in this essay i	56.57
a third point	39.14	on the other hand	44.91
all the time	39.14	they are aware of	34.25

-- Table continues on the next page --

Table D.2 *continued*

KeyBundles	**Log_L**	**Key word**	**Log_L**
in the world	35.99	this essay will examine	34.25
in this essay	35	a second point is	34.25
their children to	34.25	different from each other	34.25
different from each	34.25	another point is that	29.36
essay will examine	34.25	there are a lot	39.14
to take a	34.25	a third point is	29.36
people of the	34.25	this essay i will	59.87
want to do	34.25	stage of the process	39.14
we can say	34.25	are a lot of	39.14
a second point	34.25	the people of the	34.25
can say that	34.25	second point is that	34.25
people of villages	34.25	to become an engineer	29.36
city life is	34.25	third point is that	29.36
second point is	34.25	arguments for and against	29.36
to know about	34.25	essay i will discuss	29.36
in cities there	34.25	my aim of life	29.36
life is very	34.25	my aim in life	29.36
		life is to become	29.36

Table D. 3. Keybundle-C1

KeyBundles	Log_L	Key word	Log_L
i think that	890.21	difficult to imagine	33.86
first of all	512.77	a few years	33.86
second of all	430.54	reasons why i	33.86
i believe that	372.49	i will not	33.86
i think it	299.92	i must confess	33.86
on the other	232.2	he does not	31.67
to sum up	189.67	what they like	29.02
i do not	130.61	to reach their	29.02
i will give	130.61	to gain more	29.02
to support my	130.61	to add that	29.02
i want to	125.77	through their entire	29.02
his or her	124.11	to make our	29.02
for several reasons	111.26	to learn from	29.02
the opportunity to	109.68	has a great	29.02
when i was	106.42	how to use	29.02
the one hand	101.59	move from one	29.02
he or she	101.59	one more reason	29.02
aspect of this	96.75	great way to	29.02
which i will	96.75	is not so	29.02
it is a	95.23	i think i	29.02
i am sure	87.07	in my opinion	29.02
in addition to	84.7	i can stand	29.02
brings many benefits	82.24	is that a	29.02
do not have	75.57	it is really	29.02
a chance to	67.72	that i will	29.02
is a controversial	67.72	these practical benefits	29.02
it is very	66.47	the issue about	29.02
different kinds of	62.89	the main reason	29.02
bring many benefits	62.89	others think that	29.02
in the following	58.54	better chance to	29.02
i did not	58.48	say that the	29.02
in this world	58.05	can be taken	29.02
to get a	58.05	i think it is	256.39
role in our	58.05	to sum up i	222.52
does not have	53.29	on the other hand	198.34
help them to	53.21	in the following paragraphs	125.77
i prefer to	53.21	i think that the	125.77
is open for	53.21	reasons to support my	111.26
range of interests	53.21	in conclusion i think	101.59
has its own	53.21	reasons which i will	96.75
all over the	53.21	from the one hand	96.75
go to a	53.21	when i was a	91.91
a couple of	53.21	first of all i	87.07
a lot of	49.54	i am sure that	82.24
those practical benefits	48.37	second of all i	82.24
would have to	48.37	aspect of this is	72.56
the following points	48.37	have the opportunity to	67.72
is the one	48.37	i will give my	67.72
not want to	48.37	i think that every	67.72
my point is	48.37	my point of view	67.72
huge amount of	48.37	in addition to those	62.89
i like to	48.37	i think that it	62.89
it means that	43.54	is open for debate	53.21
did not have	43.54	in addition to these	53.21
have their own	43.54	does not have to	53.21

-- Table continues on the next page --

Table D.3 *continued*

KeyBundles	Log_L	Key word	Log_L
we need to	43.54	there is no doubt	48.37
by use of	43.54	my point is that	48.37
and present my	43.54	on the following points	48.37
they need to	43.54	an essential role in	48.37
great opportunity to	43.54	will be able to	43.54
and they have	43.54	first of all a	43.54
with each other	41.6	of all it is	43.54
in the future	40.8	is a controversial one	43.54
they do not	38.97	however i believe that	43.54
from my opinion	38.7	is the one that	43.54
they want to	38.7	all over the world	43.54
to learn more	38.7	not be able to	38.7
have a great	38.7	a great opportunity to	38.7
it is like	38.7	a huge amount of	38.7
what kind of	38.7	in order to succeed	38.7
i will list	38.7	i did not like	38.7
will have to	38.7	can bring many benefits	33.86
go to the	38.7	and present my view	33.86
to make a	38.23	from the other hand	33.86
learn how to	36.99	i think that a	33.86
to make their	33.86	have a chance to	33.86
who think that	33.86	however i think that	33.86
i base my	33.86	will help them to	33.86
because it will	33.86	i think that this	33.86
the best way	33.86	to summarize i think	33.86
be required to	33.86	in this essay i	33.77
i think the	33.86	is one of the	31.52
know each other	33.86	does not want to	29.02
the problem of	33.86	a great impact on	29.02
and i am	33.86	a great way to	29.02
not care about	33.86	to learn more about	29.02
have a few	33.86	believe that it is	29.02
they will be	33.86	the best way to	29.02
great impact on	33.86	second of all a	29.02
and gain more	33.86	in order to get	29.02