A LONGITUDINAL CORPUS STUDY OF LEXICAL BUNDLES IN STUDENTS' WRITTEN AND SPOKEN NARRATIVES

SHARON SANTHIA JOHN

FACULTY OF LANGUAGES AND LINGUISTICS
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

A LONGITUDINAL CORPUS STUDY OF LEXICAL BUNDLES IN STUDENTS' WRITTEN AND SPOKEN NARRATIVES

SHARON SANTHIA JOHN

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ENGLISH AS A SECOND LANGUAGE

FACULTY OF LANGUAGES AND LINGUISTICS
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Sharon Santhia A/P John

Matric No: TGB150003

Name of Degree: Master of English as a Second Language

Title of ~~Project Paper/Research Report~~/Dissertation/~~Thesis~~ ("this Work"):

A Longitudinal Corpus Study of Lexical Bundles in Students' Written and Spoken Narratives

Field of Study: Corpus Linguistics

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation: Supervisor

A LONGITUDINAL CORPUS STUDY OF LEXICAL BUNDLES IN STUDENTS' WRITTEN AND SPOKEN NARRATIVES

ABSTRACT

Phraseology in language use is said to be at the heart of language description (Sinclair, 1991; Hunston, 2002). Over the past 25 years there has been an upsurge in studies investigating phraseology in language use with corpus linguistics method and tools (Sinclair, 1991; Hunston 2002; Paquot & Granger, 2012). Yet, there is a lack of phraseological studies focusing on secondary school students of English (Ebeling & Hasselgård, 2015a). This study investigates the use of four-word lexical bundles based on structural and functional analysis in the written and spoken narrative texts of 42 students over a period of six months. The findings revealed that the use of lexical bundles in students' written and spoken corpora seem to decrease over time. Structurally, the written and spoken narrative texts are dominated by verb phrase-based bundles followed by noun phrase/prepositional phrase-based bundles while functionally, referential expressions are most commonly used in the written and spoken narrative texts followed by topic-oriented expressions. The substantial use of referential expressions and minimal use of stance and discourse organizing bundles in the written and spoken narrative texts, despite the difference in the modes of production may indicate the possible requirement of the narrative genre that is descriptive in nature. Taken together the overall findings, complexity, inconsistency and dynamicity are observed within the written and spoken language of students as well as between the written and spoken language where divergent developmental paths are noted in both language use. The nature of language development is also observed to include developing towards specificity and a matter of choice of the students in making use of

bundles with different structural forms for the same function in their written and spoken narrative texts.

Keywords: phraseology, lexical bundles, longitudinal learner corpus, narrative texts

KAJIAN KORPUS LONGITUDINAL TENTANG IKATAN LEKSIKAL DALAM NARATIF BERTULIS DAN LISAN PELAJAR

ABSTRAK

Frasaologi dalam penggunaan bahasa dikatakan memainkan peranan yang penting dalam kajian deskripsi bahasa (Sinclair, 1991; Hunston, 2002). Sepanjang 25 tahun terdapat peningkatan dalam kajian yang mengkaji frasa dalam penggunaan bahasa dengan menggunakan kaedah dan alatan linguistik korpus (Sinclair, 1991; Hunston 2002; Paquot & Granger, 2012). Walau bagaimanapun, kajian frasa yang memberi tumpuan kepada pelajar sekolah menengah yang mempelajari bahasa Inggeris (Ebeling & Hasselgård, 2015a) didapati agak minimal. Oleh itu, kajian ini bertujuan untuk mengkaji penggunaan ikatan leksikal berdasarkan analisis struktur dan fungsi dalam teks naratif bertulis dan bertutur oleh 42 orang pelajar dalam tempoh enam bulan. Dapatan kajian menunjukkan bahawa penggunaan ikatan leksikal dalam teks bertulis dan bertutur pelajar seolah-olah berkurangan dari masa ke masa. Dari segi struktural, penulisan dan pertuturan dikuasai oleh ikatan berasaskan frasa kata kerja yang diikuti dengan ikatan berasaskan frasa kata nama/preposisi manakala dari segi fungsinya, ungkapan referensi yang kerap-kali digunakan dalam teks bertulis dan bertutur diikuti oleh ungkapan berorientasikan topik. Penggunaan substansial ekspresi referensi dan penggunaan minimal ungkapan pendirian dan wacana penganjuran dalam teks naratif bertulis dan lisan mungkin menunjukkan ciri jenis naratif yang bersifat deskriptif, walaupun mod penggunaan bahasa berbeza. Secara keseluruhannya, kerumitan, dan dinamik diperhatikan dalam tulisan dan lisan pelajar serta antara bahasa bertulis dan lisan di mana corak perkembangan yang berbeza dilihat dalam kedua-dua penggunaan bahasa. Corak perkembangan bahasa juga diperhatikan termasuk perkembangan ke arah

pengkhususan dan pilihan pelajar dalam menggunakan ikatan leksikal dengan struktur yang berbeza untuk fungsi yang sama dalam teks naratif bertulis dan lisan.

Kata kunci: frasaologi, ikatan leksikal, korpus longitudinal, teks naratif

ACKNOWLEDGEMENTS

I am taking this opportunity to thank the people who supported and guided me throughout this journey. I owe it to all of you.

My parents, my father, John Kovilpillai, my mother, Federick Mary Stanislaus, and my sister, Sherlin Santhia John who were with me throughout this journey supporting me in the completion of my master studies, especially, my sister who tirelessly listens to my research endeavours till date. My paternal extended family, my seven aunties, Mary Santha Kovilpillai, late Lily Kovilpillai, Kamala Kovilpillai, Joyce Kovilpillai, Angel Kovilpillai, Christy Kovilpillai, Alice Kovilpillai and their spouses, especially, my godparents, Sasayah Rajagopal and Mary Santha Kovilpillai for their constant support. My maternal extended family, my two uncles, Stephen Alaxice Rozario Stanislaus, Ferliex Stanislaus, my aunty, Evensia Mary Stanislaus and their spouses for their constant support. My cousins who are my first best friends, Dr. Stanis Sutharsan Das, Sylvia Sutharsana, Aaron Anish Arivalagan, Adrian Sasayah and Mercy Shulamite Nyanamani. I am truly thankful to each and every one of you for your support and presence. All of you are living examples of sheer hard work and diligence.

The person who believed in me, my supervisor, Dr. Chau Meng Huat whom I look up to for inspiration, whose guidance, knowledge and support goes beyond this research journey and has taught me important life lessons. I am eternally grateful for such an amazing supervisor. I, indeed, consider this a blessing. His words of wisdom have impacted me deeply resonating to my heart and soul. I do not have enough words to thank you.

Pastor Tan Chee Kiang who has been a good mentor, whose prayers have lifted my spirit in times of emotional breakdowns.

My best friends, Darshini Jeyasimman and Kannigaa Markundu for always lending me your shoulders when the journey got tougher. You both inspire me to work harder everyday.

My good friends, Karima Ibrahim and Kuan Jie Ling for accepting to be the inter-raters of this study, for supporting me and encouraging me in times of need.

The 42 students of Sungai Tiram secondary school who consented to participate in my study together with the English teachers who gave me their full cooperation.

All praises to God Almighty for His inexhaustible grace and favour in my life. Without Him this would be impossible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AdjP	:	Adjective phrase
AdvP	:	Adverb phrase
BMELC	:	Business and Management English Language Learner Corpus
BNC	:	British national corpus
CALES	:	Corpus Archive of Learner English Sabah-Sarawak
CIA	:	Contrastive interlanguage analysis
CL	:	Corpus linguistics
DC	:	Dependent clause
DDL	:	Data-driven learning
EA	:	Error analysis
EFL	:	English as a foreign language
ELC	:	Engineering Lecture Corpus
EMAS	:	English of Malaysian School Students
FS	:	Formulaic sequence
ICLE	:	International corpus of learner English
LB	:	Lexical bundle
LCR	:	Learner corpus research
LINDSEI	:	Louvain international database of spoken English interlanguage
L1	:	First language
L2	:	Second language
MACLE	:	Malaysian Corpus of Learner English
MUET	:	Malaysian University English Test
MWE	:	Multi-word expression
MWU	:	Multi-word unit

NNS	:	Non-native speaker
NP	:	Noun phrase
NS	:	Native speaker
PP	:	Prepositional phrase
RQ	:	Research question
RWC	:	Recurrent word combination
SLA	:	Second language acquisition
TL	:	Target language
VP	:	Verb phrase

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Introduction of the study

Over the past 25 years there has been an upsurge in studies investigating phraseology in language use aided by CL method and tools (e.g., Sinclair, 1991; Altenberg, 1998; De Cock, 1998, 2004; Granger, 1998a; Howarth, 1998; Moon, 1998; Biber, Johansson, Leech, Conrad & Finegan, 1999; Hunston, 2002; Biber, Conrad & Cortes, 2004; Conrad & Biber, 2005; Chau, 2008; Ellis, Simpson-Vlach & Maynard, 2008; Hyland, 2008a, 2008b; Chen & Baker, 2010, 2016; Crossley & Salsbury, 2011; Ädel & Erman 2012; Paquot, 2013; Staples, Egbert, Biber & McClair, 2013; Leńko-Szymańska, 2014; Elturki & Salsbury, 2015; Allan, 2016; Pan, Reppen & Biber, 2016; Wang, 2017). A great measure of contribution is owed to Sinclair (1991) for pioneering and developing the area of corpus-assisted lexicography (Stubbs, 2008, 2009). Although phraseology has only gained its rightful status as an academic discipline in linguistics relatively recently (Sinclair, 1991; Ebeling & Hasselgård, 2015a), its history goes back to the early 20th century. Scholars like Jespersen (1924) and Firth (1957) have discussed the idea of phrases and word combinations in language use in the first half of the century. Firth (1957, p. 190) states, "…each word when used in a new context is a new word" which indicates that the act of associating meaning to single word units may not always be appropriate.

The focus then shifted due to the widespread impact of Chomskyan tradition that advocated a rule-governed approach to language processing (Ellis, 2008). Grammar was cut-off from lexis, performance and social usage which reduced the importance of studying phraseology of the language (Ellis, 2008). As a result, meaning was usually associated to single word units. Vocabulary learning was highly reliant on acquiring individual words. After some time, the association of meaning to single word units and

it being placed into the slots grammar makes available (i.e., the open-choice principle) was strongly refuted by Sinclair (1991) as a rare occasion. On the contrary, he posited that meanings are dependent on the phrases rather than single word units (i.e., the idiom principle). To quote Sinclair (1991, p. 108):

> By far the majority of text is made of the occurrence of common words in common patterns, or in slight variants of those common patterns. Most everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up text. This is totally obscured by the procedures of conventional grammar.

Sinclair's (1991) ideas on phraseology directly challenged the Chomskyan approach to studying language competence instead of language performance of learners (Granger, 1998b). Language study needed to be descriptive. The intuition based interpretations of language were prescriptive. They tended to disregard typical and less noticeable preferred phrases in language as they did not fit into the rule-governing approach (Sinclair 1991; Hunston, 2002). However, corpus-based evidence aided in describing language as it is by not allowing the researcher's intuitions to override the data (Granger, 1998b; Granger, Gilquin & Meunier, 2015). The use of learner corpora facilitated in studying SLA mechanisms using large sets of natural data which was not possible in the field before the revolution of CL. As a result, learner corpus studies dealing with phraseology of language started to flood in.

Researchers have also found that language use consist of substantial use of FS (Pawley & Syder, 1983; Sinclair, 1991; Wray & Perkins, 2000). This has led to the interest in investigating 'co-occurence' (e.g., collocation, phrasal verb) and 'recurrence' (e.g., bigrams, LBs) of FS in the texts of students of English using corpus tools (Paquot & Granger, 2012). There has been an increase in the number of studies on LBs in recent years (e.g., Biber et al., 1999, 2004; Cortes, 2004; Conrad & Biber, 2005; Biber & Barbieri, 2007; Shirato & Stapleton, 2007; Chau, 2008; Hyland, 2008a, 2008b; Chen & Baker, 2010, 2016; Crossley & Salsbury, 2011; Wei & Lei, 2011; Ädel & Erman 2012;

Staples, Egbert, Biber & McClair, 2013; Bestgen & Granger, 2014; Granger, 2014; Ong & Yuen 2014, 2015; Allan, 2016; Ruan, 2016; Pan, Reppen & Biber, 2016; Wang, 2017). Along these studies, the current study aims to investigate the use of LBs in students' written and spoken narrative texts over time. This study also aims to inform some notable gaps in the literature of LCR field. First, learner corpus studies on phraseology have been very much focused on advanced, adult learners leading to a lack of phraseological studies of secondary school students (Ebeling & Hasselgård, 2015a) (see Chau 2008, 2015; Leńko-Szymańska, 2014 for exceptions). Second, extensive learner corpus studies have dealt with the written language of learners whereas the investigation of the spoken language is relatively lesser (O'Keeffe, McCarthy & Carter, 2007; Adolphs, & Knight, 2010; Paquot & Granger, 2012; Granger et al., 2015). Third, the cross-sectional research design has become a norm in the field as opposed to the longitudinal design (Ellis & Barkhuizen, 2005; Chau, 2012; Granger et al., 2015). Longitudinal learner corpus studies are fewer in number (e.g., Crossley & Salsbury, 2011; Chau, 2015; Elturki & Salsbury, 2015) in comparison to cross-sectional learner corpus studies. Therefore, this longitudinal learner corpus study aims to investigate the use of four-word LBs in the written and spoken narrative texts of school students over a period of six months although ideally, a longer period of time would allow for more interesting observations. This is followed by structural and functional analysis of the bundles found in the written and spoken narrative texts of the students.

1.2 Background of the study

It is almost undeniable that the heartbeat of SLA research field is to want to know how a language is acquired, learned and even developed by the learner. SLA, without a doubt has lived through glorious 50 years and continues beaming in the 21st century (Ortega, 2013) of its knowledge about human language. It is also a known fact that there

are no straightforward answers to language acquisition and development processes of the learner because language in itself is like a living organism just like humans. This phenomenon further adds to the complexities of studying language use. In the 19th century some of the many great forefathers of linguistics such as Franz Bopp (1827), August Pott (1833) and Guiliano Bonfante (1946) have engaged in an unending debate of whether or not to equate language to the laws of biology, botany and zoology prompted by Charles Darwin's book titled *Origin of Species* (1859) (Sampson, 1980).

Turning back in time, in the 18th century when the systemization of the English language was prevalent, many researchers believed in codifying and preserving the language (Barber, Beal & Shaw, 2009) to be passed on to generations. In some ways this has led to the belief that any difference from the original form is impure. The impact of this view is still felt in SLA given the 50 years of growth the field has experienced where learner language is still compared to the yardstick of NS standard (Cook 1992, 2012). During the early 1960s, researchers in SLA were focused on the extent to which the learners deviated from the NS language (Ellis, 1985). Learner language features that did not conform to NS standard were identified as errors. The era of 1990s witnessed a shift towards socially and ecologically grounded theories of knowledge (Kramsch & Whiteside, 2007) (see Chapter 2 for a detailed discussion on the growth and changes in the field of SLA). Researchers began to question the deficit perspective placed upon learner language. Advocates of bilingualism and multilingualism formed a distinction between the bilingual or multilingual learner and monolingual NS (Cook, 1992, 2012). Through the complexity theory, Larsen-Freeman (1997, 2006) posited the need to treat learner language as a separate system from the NS system. This view is further supported by researchers such as Cook (2012), Garcia (2014) and Chau (2015). It is the researcher's intention (along with this body of

research) to treat learner language in its own right and eventually make sense of the nature of language development.

1.3 Aim of the study

This learner corpus study aims to investigate the use of four-word LBs in the written and spoken narrative texts of secondary school students over a period of six months. It also aims to examine the structures and functions of the bundles found in the written and spoken narrative texts of these students. The nature of language development is then observed based on the use, structural and functional analysis of bundles found in the written and spoken narrative texts.

1.4 Research questions

In line with the aim of the study, the researcher intends to answer three RQs that are as follows:

1. What are the most frequent four-word LBs that occur in the written and spoken narrative texts of the students over time?
2. What are the structures and functions of the four-word LBs and to what extent do the LBs found in the written narrative texts differ from those found in the spoken narrative texts?
3. How might the changes in the use of LBs observed over time explain the nature of language development?

1.5 Significance of the study

The present study seeks to contribute to the fields of LCR, phraseology and SLA research. First, this study is a first study in LCR that examines a longitudinal corpus of both the written and spoken narratives of 42 students of English. Second, since there is a

lack of longitudinal studies in LCR (Ellis & Barkhuizen, 2005; Chau, 2012; Granger et al., 2015), the present study adds to LCR by making use of longitudinal data to track language development. Third, it also contributes to the literature by studying the use of phraseology among secondary school students of English as little is known about how these students make use of LBs in their written and spoken language (Ebeling & Hasselgård, 2015a). Furthermore, this learner corpus study is one of the very few studies that methodologically treat learner language as an independent system and not as a substandard of an idealized norm. Thus, it contributes to the field of SLA research.

1.6 Scope of the study

Unlike past learner corpus studies that looked at phraseology in the language of adult, advanced learners, this study investigates the use, structures and functions of four-word LBs in the written and spoken narrative texts of 16-year-old students of English. There are 252 written and spoken narrative texts in total, contributed by the same group of 42 Malaysian Secondary Four students from a national type secondary school over a period of six months.

1.7 Conclusion

As noted earlier, this learner corpus study aims to investigate the use of LBs in the written and spoken narrative texts of secondary school students of English over time. In addition to that, the structures and functions of bundles found in the narrative texts are examined. The nature of language development is observed based on the use, structural and functional analysis of bundles found in the written and spoken narrative texts.

In the next chapter, a review on a range of past studies relevant to the present study is provided. This is followed by a detailed discussion on the methodology used in Chapter 3 and the findings and discussion of the study in Chapter 4. Finally, in Chapter 5 a

conclusion is provided with the implications and limitations of the present study as well as suggestions for future research.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

As noted in Chapter 1, this study aims to (1) investigate the use of LBs over time, (2) the structures and functions of bundles, and (3) the nature of language development. Hence, the present study covers three broad research fields: (1) CL, the subfield LCR, (2) phraseology, and (3) SLA. The insights drawn from these three research fields work collectively to address the concerns of the present study. In this chapter, a brief overview of the growth and change of these research fields is provided together with a review and discussion on a range of past studies done in the respective research fields relevant to the present study.

2.2 Corpus linguistics and learner corpus research

Through the initiative of Sinclair (1991), CL gave a new dimension to language whereby language was beginning to look a lot different from what it seemed to be previously. Corpus investigation techniques were introduced to provide objective evidence by processing 'raw' texts (Sinclair, 1991). This was very different from the conventional methods in SLA research. Language mechanisms were studied using data yielded from controlled environments and language interpretations were mainly based on the intuitions of the researcher which were said to be manipulated and prescribed (Granger, 1998b; Granger et al., 2015). The use of manipulated data to study language use is refuted by Sinclair (1991). He states, "[o]ne does not study all of botany by making artificial flowers" (p. 6). Instead, he advocates the need for objective evidence to study language use. Moreover, the intuition-based interpretations of language were prone to disregard typical and less noticeable preferred phrases that existed in language use because they did not fit into the rule-governing approach (Sinclair 1991; Kennedy,

1998; Hunston, 2002). CL methods aid in describing language as it is without having the researcher's intuitions to override the data obscuring the insights that the data can provide about language use (Granger, 1998b).

The LCR, one of the branches of CL is a growing yet well-known field for its contributions to studying language use of learners using computer-assisted methods over the past 25 years (O'Keeffe, 2007; Granger et al., 2015). LCR is sought after for its twofold advantages. First, it permits the investigation of learner language in a naturally occurring state as in the classroom without manipulation or control imposed. Second, it aids in processing large sets of data samples using computer-assisted tools (Bonelli, 2010; Granger et al., 2015). In the past, SLA researchers could only deal with limited data samples due to manual analysis. Another unresolved challenge in SLA is to bridge the gap between SLA research community and the teachers in solving the classroom issues. This is because the nature of data used was rather artificial and the findings of the studies were not directly applicable in the classroom (Ellis, 1997; Granger et al., 2015). LCR, however, is applied orientated whereby it does not just stop at providing solutions to inform research practices but also provides practical solutions that are implementable in the classroom (Chau, 2012, 2015; Granger et al., 2015).

The first two pioneering learner English corpora in the European context are known as the Longman Learners' Corpus and the ICLE (Granger, 2003; Paquot & Granger, 2012). The ICLE comprises written texts of learners from 16 different mother-tongue backgrounds. Its spoken counterpart, the LINDSEI is relatively smaller in size. It is made up of oral data yielded from learners of 11 mother-tongue backgrounds (Paquot & Granger, 2012). These two corpora are smaller in size in comparison to the BNC or the Bank of English yet they function as a solid empirical base for SLA research (Granger, 2003). On the other hand, in the Malaysian context, corpus research in English language dates back to the 1990s (Hajar, 2014). The pioneering corpus project was on developing

a Malay language corpus that was initiated in the early 1980s. To date, the development of learner English corpora and corpus-based studies of English language outweigh those on the Malay language (Hajar, 2014). There has also been a rise in the development of learner English corpora in Malaysia (Hajar, 2014; Siti & Hajar, 2014). Some of the notable Malaysian learner English corpora include the EMAS corpus, CALES, MACLE and the genre-specific learner corpora such as BMELC and ELC (Hajar, 2014; Siti & Hajar, 2014) which are used for research purposes.

One of the initial corpus-based studies on learner language in Malaysia was done by Arshad (2004). In his study, Arshad (2004) made use of the EMAS corpus comprising written essays of about 800 students from three different age groups (i.e., 11 years old, 13 years old and 16 years old). He studied the students' language development using cross-sectional data by examining their language production as well as vocabulary sophistication and range. The results showed some form of increase in the language production and vocabulary use of all three age groups. Chau (2008) conducted a pseudo-longitudinal study using the written data of Malaysian Secondary One learners of English (i.e., 13 years old) which was part of the EMAS corpus project. He investigated the development of phraseological competence of these students in his study. Chau's (2008) study confirmed the view of dynamism in language development as the results revealed that learners produced basic verb + noun sequences at the beginning level then proceeded with an overflow of the sequences, and then moved on to more sophisticated use of the sequences. This process was noted as a dynamic process where the learner reorganizes his/her linguistic repertoire in the course of language development. Apart from that, researchers have made use of learner corpora to investigate a wide range of issues faced by learners of English in the country. These studies include investigating spelling errors of L2 learners using the CALES (e.g., Botley & Dillah, 2007), studying the collocational competence among undergraduate

law students (e.g., Kamariah & Su'ad, 2011), conducting a comparative study to investigate compliment patterns in the writing of Malay ESL students and NS (Paramasivam & Atieh, 2017) to name a few.

Learner corpus studies are also carried out to examine various types of linguistic and grammatical features, to test hypotheses and theories of SLA and to study phraseology in learner language and so on. Different types of methodologies have been employed to study the mechanisms of SLA such as CIA, combination of learner corpus and experimental method, comparisons between L2 learner data and NS data (i.e., L2 vs. L1) as well as between two different L2 data (i.e., L2 vs. L2) (Granger, 2003; Paquot & Granger, 2012; Callies, 2015). In this study, corpus investigating techniques are used to investigate the use of LBs in the written and spoken narrative texts of students (see Chapter 3 for a detailed explanation on the methodology used). In the next section, an overview of the area of phraseology is provided with a review on a range of learner corpus studies dealing with phraseology.

2.3 Phraseology

Phraseology in language use in the western tradition is said to be highly influenced by the developments of Russian phraseology (Cowie, 1998a). The scholars, H. E. Palmer and A. S. Hornby are acknowledged as the founding fathers of EFL lexicography who have paved the path for significant growth of the field (Cowie, 1998a). As highlighted in the first chapter, the idea of phraseology in language use was evident in the first half of the 20th century which then lost its focus when Chomsky's idea on ruled-governed approach to language began to prevail. The Chomskyan tradition advocated general grammatical rules and principles of Universal Grammar which abandoned the importance of phraseology in language use (Ellis, 2008). As a result, traditionally, language acquisition was based on learning syntactic rules.

It is only in the late 20th century Sinclair's (1991) groundbreaking discoveries precipitated a major shift in the area of phraseology highlighting the importance of phrases in language use. Some of his major arguments directly challenged Chomsky's ideas on the rule-governed approach to language acquisition. According to Hunston (2002, p. 138):

> Sinclair (1991) puts phraseology at the heart of language description, arguing that the tendency of words to occur in preferred sequences has three important consequences which offer a challenge to the current views about language:
>
> - There is no distinction between pattern and meaning;
> - Language has two principles of organization: the idiom principle and the open-choice principle;
> - There is no distinction between lexis and grammar.

Sinclair (1991) puts forth the view that everyday language use is made up of preferred sequences of words and these preferred sequences of words (i.e., phrases) are the carriers of meaning rather than individual words. He illustrates this phenomenon using two principles in which he states that language operates more often according to the idiom principle and less often according to the open-choice principle. For instance, the hearer or reader understands the meaning of a phrase from the phrase itself rather than from the individual word made available by grammatical slots. He also challenges the conventional idea of distinction between lexis and grammar by arguing that there is no crucial difference between both (Hunston, 2002). It is also argued that through the observation of the patterns attached to all lexical items, grammar can be formed. Sinclair's (1991) views gave a new perspective to language study which then resulted in an increase of phraseological studies in the areas of CL and LCR. The studies dealing with phraseology using learner corpora have not only informed the learning and teaching processes but have also challenged the conventional ideas about language providing new insights to be explored further in the research realm.

The unsystematic terminologies and arbitrary characteristics to identify phraseological units have added to the complexity in studying phraseology in language use (Cowie, 1998b; Ebeling & Hasselgård, 2015a). After all, [p]hraseolog is a fuzzy part of language (Altenberg, 1998, p. 101). Wray and Perkins (2000) argue that there are about 40 terms used to refer to the different types of FS. To illustrate, the terms used by researchers to refer to different types of FS include collocation (Firth, 1957; Sinclair, 1991), prefabricated patterns (Hakuta, 1974), memorized sentences and lexicalized stems (Pawley & Syder, 1983), lexical phrases (Nattinger & De Carrico, 1992), recurrent word-combinations (Altenberg, 1998), prefabs (Granger, 1998a), and LBs (Biber et al., 1999). MWU such as idioms (e.g., *out of the blue*), proverbs (e.g., *beauty is only skin deep*) and similes (e.g., *as white as snow*) are said to be fixed, idiomatic and semantically opaque or transparent sequences. LBs (e.g., *in the case of*) and collocation (e.g., *heavy rain*) are said to be fixed and semantically transparent sequences. Idioms, also, referred as 'colourful' sequences (Granger, 2014) have been widely studied in the past (Howarth, 1998; Paquot & Granger, 2008) due to their infrequent usage which gives a proficient status to the language user. However, there is a need for substantial contextual and pragmatic analysis to understand the meaning of these sequences (Wray & Perkins, 2000). On the other hand, FS that are fixed and semantically transparent (i.e., LBs) are usually dismissed as insignificant sequences probably because they are commonly found in the writing and speech of the language user. Nonetheless, the very ubiquitous nature of this FS has attracted the attention of researchers like Biber et al. (1999). Biber et al. (1999) found that LBs are relatively common than idioms in registers (i.e., conversation and academic prose). For instance, bundles like *in the case of* and *do you want me to* occurred at least 20 times per million words in comparison to idioms like *slap in the face* and *kick the bucket* which occurred less than 5 times per

million words in the two registers. These idioms were found to be even more less in registers like conversation (Biber et al., 1999).

Apart from that, researchers have also found that everyday language use comprise substantial use of FS (Pawley & Syder, 1983; Sinclair, 1991; Wray & Perkins, 2000; Biber et al., 1991, 2004). It is almost undeniable that language users do rely on phrases when they write or speak. For example, they say, *a very good morning* less so, *a much great morning*, *well done* not as much of, *well finished*, *all the best* and rarely, *all the great*. These instances give an impression that words do have preferred sequences and readers or hearers understand the meaning of these phrases from the phrases itself. It is highly unlikely for people to make use of novel and creative language in their everyday communication. If language users did so then there would be a great deal of effort spent attempting to interpret the intended message. This by no means intends to undermine the ingenious, creative thoughts showcased by great poets and writers of the century through high-flown, elaborate language. The main goal of language users is to communicate through writing and speech in order to convey the intended message. In that endeavour language becomes an instrument that bridges the communicative process between both parties. Hunston (2010) argues that there are a lot of repetitions involved when someone writes or speaks the language without planning them consciously and these repeated words then become patterns. This has also attracted researchers to study recurrent word sequences such as LBs in language. The investigation of LBs in the written and/or spoken language of students of English shows an increase over the years (Greaves & Warren, 2010; Paquot & Granger, 2012; Granger, 2014). In the following section, a review on a range of studies investigating LBs in the written and spoken language of students is presented.

2.4 Lexical bundles

Over the past two decades there has been a rise in studies investigating LBs with the use of corpus tools. Biber et al. (1999) first coined the term LBs in the *Longman Grammar of Spoken and Written English*. LBs are defined as "…sequences of words that most commonly co-occur in a register" (Biber et al., 1999, p. 989). These sequences are structurally incomplete units (Biber et al., 1999, 2004). Several terms used to refer to LBs (i.e., sequences that are fixed and continuous) include clusters (Hyland, 2008a), prefabs (Granger, 1998a) and RWC (Altenberg, 1998). LBs have caught the attention of many researchers for reasons such as their frequent occurrence in language use, specific discourse functions in text organization, semantically transparent property that aids in minimizing the processing and decoding effort as well as for fluency purpose (Pawley & Syder, 1983; Wray & Perkin, 2000; Biber et al., 2004; Conklin & Schmitt, 2008, 2012). LBs are disregarded by traditional linguistic research for two main reasons. First, LBs are semantically transparent units and thus are discounted by researchers who consider idiomaticity a necessity for formulaic language (Biber et al., 2004; Conrad & Biber, 2005). Second, they are made of up clausal (e.g., *it is possible to*) and phrasal (e.g., *at the beginning of*) fragments that are not complete structural units (Ädel & Erman, 2012). LBs differ from the grammatical items recognized by tradition linguistic research (Biber et al., 2004; Conrad & Biber, 2005). Despite its non-idiomatic and structurally incomplete properties, it has been found that LBs function to bridge two clauses in speech (e.g., *I want to know*) and two phrases (e.g., *in the case of*) in writing (Biber et al., 2004; Biber & Barbieri, 2007).

The literature suggests that there has been a good number of studies on LBs in various areas of CL and LCR (see Hyland, 2012 for a detailed review on LBs in academic discourse). These areas include a wide range of registers (e.g., Biber et al., 1999, 2004; Conrad & Biber, 2005; Biber & Barbieri, 2007), genres and/or disciplines

(e.g., Cortes, 2004; Hyland, 2008a, 2008b; Allan, 2016; Pan, Reppen & Biber, 2016; Wang, 2017). A body of research has also looked at the use of LBs in native writing in comparison to non-native writing (i.e., L1-English vs. L2-English) (e.g., Granger, 1998a; Ellis, Simpson-Vlach & Maynard, 2008; Chen & Baker, 2010, 2016 (Chinese); Wei & Lei, 2011 (Chinese); Ädel & Erman 2012 (Swedish); Paquot, 2013 (French); Staples, Egbert, Biber & McClair, 2013; Ebeling & Hasselgård, 2015b (Norwegian)) and speech (e.g., Altenberg, 1998; De Cock, 1998, 2004 (French); Shirato & Stapleton, 2007). LBs have been studied in non-native varieties (i.e., L2 vs. L2) (e.g., Huang, 2015) as well. Researchers have also investigated the developmental processes of LBs in student writing (e.g., Chau, 2008; Bestgen & Granger, 2014; Ruan, 2016), speech (e.g., Crossley & Salsbury, 2011) and in both written and spoken language using longitudinal data (e.g., Elturki & Salsbury, 2015). Granger (2014) conducted a study to examine the use of LBs in two languages, English and French. Researchers have carried out quite a few studies on LBs in the Malaysian context as well. Some of the LB studies that has been done in Malaysia include Chan, Hadi and Tan's (2014) study that examined LBs in group discussions of university students, Ong and Yuen's (2014, 2015) studies that investigated the use of LBs in MUET reading texts as well as Hadi and Chan's (2014) study on LBs in university lectures. Given the laying out of various types of LBs studies conducted in the past, the categorization above may not be as direct as it seems to be as there may be overlaps of studies fitting into more than one category.

Glimpsing through the history of LBs studies, the very first study using corpus method was probably conducted by Altenberg (1998) using the London-Lund Corpus in which he investigated three-word recurring sequences in English. Subsequently, Biber et al. (1999) investigated four-word, five-word and six-word LBs in two registers, conversation and academic prose. It was found that conversation contained more LBs than academic prose. A structural taxonomy was developed in Biber et al. (1999)

comprising 12 different structural patterns in academic prose and 14 different structural patterns in conversation. Most of the bundles in conversation were made up of pronominal subject followed by VP (e.g., *I don't know why*) and the beginning of a complement clause (e.g., *I thought that was*). The bundles found in the academic prose were made up of NP (e.g., *the nature of the*) and PP (e.g., *as a result of*). The bundles in conversation consisted of the beginning of a main clause followed by the beginning of an embedded complement clause. In contrast, the bundles in academic prose were nominal rather than clausal bundles. It was concluded that most LBs in conversation tend to be building blocks for verbal and clausal structural units whereas the bundles in academic prose are building blocks for extended NP or PP.

Following Biber et al. (1999), a series of studies as extensions of this study were conducted. One of the studies is Biber et al. (2004) which explored the structures and functions of LBs in two university registers, textbooks and classroom teaching. Biber et al. (2004) compared the findings of their study to the findings of the previous study by Biber et al. (1999). A revised structural taxonomy comprising three main structural categories was developed in this study: (1) VP-based bundles, (2) DC-based bundles and (3) NP and PP-based bundles. Along that, a preliminary functional taxonomy was developed. Three main functional categories were identified: (1) stance expressions, (2) discourse organizers and (3) referential expressions. The findings revealed that bundles used in classroom teaching were similar to conversation despite the fact that classroom teaching was pre-planned. Surprisingly, classroom teaching had the most LBs compared to the other three registers. It was expected that classroom teaching would be more literary. But classroom teaching contained both conversational and literate bundles as a consequence of its reliance on face-to-face interaction that needed speech production to be processed on the spot. The identification of structural categories revealed that these bundles had strong grammatical correlates that help bridge sentences. The functional

characteristics of these bundles showed that they hold important discourse functions that are distinctive according to registers. The bundles used in the spoken register (i.e., conversation) were dominated by stance expressions whereas the written registers (i.e., textbooks and academic prose) were dominated by referential expressions. Unexpectedly most of the bundles in the spoken register, classroom teaching functioned as stance expressions and referential expressions having a combination of both oral and literate bundles.

Conrad and Biber (2005) investigated the use of three-word and four-word bundles across two varieties of English language in two registers, conversation (British English) and academic prose (American and British English). The findings revealed that there were more bundles used in conversation (i.e., 28%) than academic prose (i.e., 20%). Conrad and Biber (2005) claimed that although LBs did not cover a major part of words in both registers, they carry important discourse functions.

The study by Biber and Barbieri (2007), an extension of the past study by Biber et al. (2004) examined the use and functions of four-word bundles in spoken and written university registers like management registers (i.e., written course management and class management talk), instructional registers (i.e., textbooks and classroom teaching), student advising (i.e., office hours), institutional registers (i.e., institutional writing and service encounters) and student-student academic interactions (i.e., study groups). Biber and Barbieri (2007) found that the written register, course management contained the most LB types in comparison to all the other registers. As for the spoken registers, service encounters and class management talk contained the most bundle types. Classroom teaching ranked as the third highest register among all the registers for bundle types used. The finding here contradicts the earlier findings by Biber et al. (2004) in which classroom teaching contained the most bundle types. Additionally, the use of bundles in institutional writing was just as much as the use of bundles in the

spoken registers. The findings here challenge the findings in the past which showed that LBs are relatively common in spoken register than in written register (Biber et al., 1999). Biber and Barbieri (2007) argue that the use of LBs is not only dependent on the general spoken or written differences. But it is also highly influenced by the communicative purpose which determines the extent to which a speaker or writer depends on bundles. In terms of the functional distribution of these bundles, stance bundles were widely used in all the spoken university registers compared to other functional categories. Service encounters made use of the most stance bundles compared to other spoken university registers. This is because stance bundles are said to be a general characteristic of spoken university registers. On the other hand, as for the written registers, stance bundles were most commonly used in course management only whereby institutional writing was dominated by referential bundles.

Biber et al. (2004) and Conrad and Biber (2005) argue for the theoretical status of LBs as having an important role in constructing discourse. They claim that these units should be seen as a basic linguistic construct which are different from the traditional linguistic features. Although LB studies take on a frequency-driven approach where frequency becomes the deciding criteria, it is claimed that LBs should not be discounted as unimportant sequences (Biber et al., 2004) (see Chapter 3 for a detailed discussion on the identification of bundles). This is because LBs can be interpreted in terms of structure and function. Even though they do not fit into the grammatical structures acknowledged by traditional linguistic research, most LBs are made up of well-defined structural correlates. For instance, the structures of bundles can function as structural 'frames' followed by a 'slot' which provide readers with the knowledge to interpret information (Biber et al., 2004; Biber & Barbieri, 2007).

Another prominent figure in this area of MWU is Hyland (2008a) who refers to LBs as academic clusters and extended collocations. Hyland (2008a) investigated the use of

four-word clusters in terms of their forms, structures and functions in three corpora of research articles, masters and doctoral dissertations. He went on to explore how these clusters differed across three different academic genres. Hyland's (2008a) study is different from the ones in the past as it fills in the gap in the literature by looking at specific use of clusters across different academic genres, identifying the similarities and differences in all three academic genres. The findings of this study support the findings of previous studies by Cortes (2004) as well as Scott and Tribble (2006) which claimed that there are variations in the frequency of form, structure and function of clusters used in student and expert writing.

Pan, Reppen and Biber (2016) took on a disciplinary perspective to examine LBs. They examined the structural patterns and functional characteristics of four-word LBs used by L1-English versus Chinese L2-English academic professionals in their written texts for Telecommunications journals. The results revealed that there were 55 four-word bundles in TELE-EN corpus and 71 bundles in TELE-CH corpus. About 24 bundles were shared by both groups of writers. Three bundles used by Chinese L2 writers did not occur in the NS corpus and this is said to be the result of translation from Chinese language to English. It is inferred that L2 writers heavily rely on the use of LBs compared to L1 writers. In terms of the structural types of bundles used, it was found that L1 writing contained more phrasal bundles (i.e., NP and PP-based bundles) whereas the L2 English writing contained more use of clausal bundles (i.e., VP-based bundles). This study also supports the hypothesis by Biber, Gray and Poonpon (2011) in which it is stated that academic writers go through a developmental progress from making use of clausal bundles to phrasal bundles. Functionally, text-oriented bundles were widely used in both corpora whereas stance bundles were found to be least used in both corpora.

In addition to the studies discussed above, two initially established corpus studies in the area of phraseology are by Moon (1998) and Granger (1998b). Moon (1998)

investigated the correlations between frequency, form, idiom type and discourse functions of phrasal lexemes using an 18 million word corpus known as the Oxford Hector Pilot Corpus. Phrasal lexemes are phraseological units ranging from "…fixed and semi-fixed complex items which dictionaries in the Anglo-American tradition classify and treat as 'phrases' or 'idioms'…" (Moon, 1998, p.79). These sequences were classified into three categories which are 'anomalous collocations' (i.e., closely related to 'restricted collocations'), 'formulae' (i.e., simple formulae, sayings, proverbs, and similes) and 'metaphors' (i.e., transparent, semi-transparent, or opaque metaphors). It was found that 70% of phrasal lexemes occurred less than one in a million words. The metaphorical expressions had frequencies lesser than one per million words. However, 'anomalous collocations' were found to be very common expressions. Simple formulae accounted for 70% of the Hector Corpus. There were no metaphors that occurred more frequently than fifty times per million words. Metaphors which occurred were not pure idioms as well. The corpus data revealed that only a few literal equivalents of metaphorical expressions were found which contradicted the conventional assumption that true idioms ought to have literal referents. About 5% of the phrasal lexemes were polysemous. The frequent polysemous phrasal lexemes were *give way*, *in line* and *take care* – the different uses of meaning were linked to different forms and collocations. About 40% of phrasal lexemes did not have fixed forms.

Granger (1998b) examined the use of prefabricated language (i.e., collocations and formulae) in advance French speaking EFL learner writing in comparison to NS writing. The NS corpus used for the study comprises parts of three corpora: the Louvain essay corpus, the students essay component of the International Corpus of English and the Belles Lettres category of the Lancaster-Oslo-Bergen corpus. The NNS corpus is a sub-corpus of the ICLE. In terms of collocations, Granger (1998b) examined the use of intensifying adverbs (i.e., amplifiers ending in –ly) (e.g., *although this feeling is*

perfectly natural). They consisted of collocations from restricted collocability (e.g., *bitterly cold*) to wide collocability (e.g., *completely different/new/free*). She found that NS writing contained more amplifiers than NNS writing. The learners overused two amplifiers (i.e., *completely*, *totally*) and underused one amplifier (i.e., *highly*). This is said to be due to direct translation from the learners' L1 (French). It was noted that learners tend to use collocational pairs that are uncommon among NS which suggest that they have an underdeveloped sense of salience and difficulty in identifying collocations. In terms of formulae, Granger (1998b) focused on 'sentence-builders', phrases that are known as macro-organizers in the learner's text. She examined formulae consisting of two discourse frames: (1) passive frame, 'it + (modal) + passive verb (of saying/thinking) + that-clause' (e.g., *it is said/thought that…; it can be claimed/assumed that…*), and (2) active frame, 'I or we/one/you (generalized pronoun) + (modal) + active verb (of saying/thinking) + that-clause' (e.g., *I maintain/claim that…; we can see/one could say that…*). The results revealed that NNS made similar use of passive structures as the NS but they overused the active structures. Granger (1998b) inferred that learners cling on the limited fixed phrases that they feel confident using because of their restricted repertoire in English. Based on the results, it is said that the use of prefabs as well as learners' acquisition process are strongly influenced by their L1.

Furthermore, a good number of learner corpus studies have investigated the use, structural patterns and functional characteristics of LBs in the written or spoken language of NNS in comparison to NS in the European and Asian settings. Among the notable studies is the study by Chen and Baker (2010). They investigated the use of four-word LBs in terms of their structure and function by conducting a three-way comparison between L1-English and Chinese L2-English student academic writing to native expert writing in published research articles. The researchers found that the NNS

and NS student writing displayed similar use of LBs. VP-based bundles and discourse organizers were more commonly found in NNS and NS student writing in comparison to native expert writing. The native students made use of a more cautious language but the L2 writing displayed preference for particular idiomatic expressions and connectors. The L2 students also over-generalized some bundle types. Based on the findings, it was claimed that the use of formulaic expressions tend to increase with writing proficiency. Chen and Baker's (2010) findings are contrary to Hyland's (2008a) findings which showed that clusters were more commonly used by postgraduate students than professional writers. Hyland (2008a) indicated that less proficient students are more likely to rely on formulaic expressions to exhibit their competence in academic discourse than proficient writers. One possible reason for this contradiction to occur is said to be because Hyland (2008a) did not remove context-related bundles as well as overlapping bundles which were removed by Chen and Baker (2010).

Ädel and Erman (2012) studied the use of four-word LBs in undergraduate Swedish EFL learner writing comparing it to NS writing. The functions of these LBs found in both corpora were analysed as well. This study is amongst the first to investigate the use of LBs in undergraduate EFL setting in the European context. The researchers hypothesized that NNS students would produce fewer bundles (i.e., overall frequency) and lesser varied bundles (i.e., bundle types). Ädel and Erman's (2012) study confirmed the hypothesis formed as NS writing contained a relatively wider range of bundles in comparison to NNS writing accounting for 130 bundles and 60 bundles respectively. It was also found that 22% of bundles were shared by both groups. This finding here is similar to the finding of Chen and Baker (2010). However, Chen and Baker (2010) claimed that bundles used in both native student writing and non-native student writing were similar but this was not the case in Ädel and Erman's (2012) study. They also found that both groups relied more on referential expressions accounting for 47% and

45% of the overall bundles respectively. Non-native students used discourse organizing bundles more than native students which accounted for 27% and 22% of the overall bundles respectively.

Ebeling and Hasselgård (2015b) looked at three-grams and four-grams in the written texts of Norwegian learners of English and NS of English across two academic disciplines (i.e., linguistics and business). They investigated the saliency and functions of n-grams used by both groups. Similar to Chen and Baker (2010) and Ädel and Erman (2012), this study compared the use and discourse functions of n-grams between learner writing and NS writing. This study adopted the functional framework by Moon (1998) which includes three main categories: ideational or informational, interpersonal and textual. Modifications were made to the framework following Halliday's metafunction. The functional analysis revealed that both NS and learners from the linguistics discipline made high use of informational n-grams than interpersonal and textual n-grams. However, NS writing contained a greater use of informational n-grams than learner writing. The second most used n-grams were the organizational n-grams that were relatively lesser in NS writing than learner writing. Notably, no situational n-grams were found in learner writing. On the other hand, in the business discipline, learners made use of more informational n-grams than NS. Situational n-grams were not found in NS writing. Both disciplines only shared 6% of the n-grams yielded which were interpersonal and textual n-grams. Learners used fewer modalizing and evaluating n-grams than their counterparts in both disciplines. This study showed statistically important differences between disciplines than the NNS and NS comparison. The result of this study is quite similar to those in the past which clearly suggest that n-grams are discipline specific.

Discussed above are some of the significant learner corpus studies that have examined the use of LBs in NNS writing in comparison to NS writing. Now, a review

on the studies that have dealt with LBs in the spoken language of learners of English is presented below. As highlighted earlier in Chapter 1, phraseology in learner speech is an interesting area of research which has brought about many phraseological studies although not as much as studies that has dealt with learner writing (O'Keeffe et al., 2007; Adolphs & Knight, 2010; Paquot & Granger, 2012; Granger et al., 2015). Hakuta (1974) is one of the initial studies that investigated prefabricated patterns in the speech of a five year old Japanese child over 60 weeks. In this longitudinal study, three prefabricated patterns were analysed: (1) the use of copula, (2) *do you* segment used in questions and (3) *how to* segment in *how*-questions. Some interesting discoveries of the study include the strategy of learning through memorization of segments without the knowledge of the internal structure of the segments of speech. These patterns were said to be employed by the learner at the initial stage as a prop before building the foundation in the language learnt. Copula sentences were made up of about half her speech in the first month however, reduced from the second month onwards to 20% eventually. An interesting interplay between form and function was noted by Hakuta (1974). The learner made use of the rigid form *these are* to express plurality. She made use of this form sometimes in singular noun sentences as well. Moreover, the learner also produced correct utterances of the segment *do you* and then moved on to use *how*-question form which disintegrated over time. It was found that she made inverted forms of this type resulting in incorrect forms which again suggested that she did not just depend on what was heard from her peers.

Another significant corpus study which looked at phraseology in spoken language is by Altenberg (1998) who investigated the grammatical and functional aspects of RWC using the London-Lund Corpus of Spoken English. The findings of this study revealed that RWC that were extracted ranged between three to five words and it was concluded that RWC in speech appeared to be fairly short. In terms of the grammatical types, these

sequences were categorized into three categories: full clauses (i.e., independent and dependent), clause constituents (i.e., multiple and single) and incomplete phrases. The clause constituents accounted for 56% of the phrases in comparison to the other two grammatical categories. Incomplete phrases and full clauses were made up of 14% and 10% of the phrases respectively. As for the independent clauses, they were categorized into three functional categories: responses, epistemic tags and metaquestions. The most used independent clauses were responses which indicated the interactive nature of spoken discourse. The epistemic tags (e.g., *I don't know, I'm not sure*) functioned as modal comment clauses. The metaquestion reflected difficulties of encoding in spontaneous speech. These phraseological units were semantically transparent. Only a few of the sequences were syntactically fixed. These expressions were said to be restricted to particular speech situations. Altenberg (1998) claims that RWC are conventionalized language. They are widespread and have various functions in the spoken language. Most of these sequences are free constructs and lexicalized units rather than completely fixed sequences which complicate the distinction between lexis and grammar. He mentions that speakers who are engaged in spontaneous interaction retrieve expressions from a large stock of RWC to convey their intended message and thus seldom make use of completely fixed sequences as observed through the findings.

Apart from that, De Cock (1998) is one of the initial studies which investigated formulae in the speech of adult French EFL learners in comparison to NS speech. The researcher examined two-word to five-word formulae in the spoken language of both groups. In the past, SLA researchers dealt with limited spoken learner data samples. This study is one of the firsts that deals with large spoken data samples with the aid of computer-assisted techniques. The NNS corpus used for this study is the LINDSEI that constitutes 25 transcripts of informal interviews whereas the NS corpus comprises 25 transcripts of informal interviews. In this study, the researcher lays out rigid criteria as

an attempt to improvise on the identification of formulaic expressions. De Cock (1998) filters the automatically extracted sequences following three phases: the form filter, the function filter and the formulaic filter. This is done to examine the validity of formulaic expressions extracted rather than simply concluding that the formulae extracted are all important. Not all automatically extracted RWC are said to be formulaic as some sequences do not carry any pragmatic or discourse-structuring functions (e.g., *in the*, *it was it was*) which is why De Cock (1998) argues for the need of a manual filtration process.

Following that, De Cock (2004) explored the use and functions of two-word to six-word sequences in the spoken texts of advanced French EFL learners in comparison to the spoken texts of NS of English. It was found that NNS speech had more use of two-word to five-word sequence types but less six-word sequence types compared to NS speech. The findings on the speech tokens showed that learners overused two-word to six-word sequences. However, after the removal of repeated and hesitation sequences, the findings revealed that learners underused two-word to six-word sequence types compared to NS. The results here were in line with Altenberg's claim where the length of these sequences was inversely linked to the frequency of the sequences. The NNS corpus contained three to four times more repeats and hesitations than NS corpus. Markers of vagueness were significantly underused by NNS in their speech. It was inferred that there were less interaction and involvement in spoken language of learners compared to the spoken language of NS.

Shirato and Stapleton (2007) investigated the use of conversational vocabulary in the speech of Japanese EFL learners and NS speech data. Single word units and MWE were examined in the spoken data of NNS and NS. The MWE were classified according to four different functional categories: (1) discourse markers, (2) vagueness and approximation, (3) indirect forms of face and politeness and (4) hedging. The findings

showed that the NNS relied less on these MWE compared to NS. Some of these expressions that carried important functions were not even found in the speech of learners. For instance, the discourse markers (e.g., *you know*, *I mean*) and vagueness and approximation items (e.g., *and something like that*, *a couple of*) that were common in NS speech did not occur in NNS speech. MWE functioning as face and politeness items (e.g., *do you think*, *I don't know if/whether*) only occurred once in the learner data. The hedging markers were commonly used by NS and learners. However, the learners used hedging markers (e.g., *sort of*) that preceded only NP whereas the NS used it preceding nouns, adverbs, verbs, PP and adjectives. Shirato and Stapleton (2007) also discussed the potential strategies of teaching clusters in the EFL contexts.

Some important cross-sectional studies on LBs in the written and spoken data of learners were reviewed above. As stated in the previous chapter, extensive LBs studies are cross-sectional in design leading to a lack of longitudinal studies investigating LBs in the field (Ellis & Barkhuizen, 2005; Chau, 2012; Granger et al., 2015). Some of the longitudinal studies that have examined LBs include Bestgen and Granger (2014) and Ruan (2016) in learner writing, Crossley and Salsbury (2011) in learner speech and Elturki and Salsbury (2015) in both learner writing and speech.

Bestgen and Granger (2014) and Ruan (2016) took on a longitudinal approach to study the development of LBs in L2 learner writing. Bestgen and Granger (2014) observed the phraseological competence of L2 learners by examining the quality and quantity of bigrams in L2 learner writing in comparison to a reference corpus using the CollGram technique, MI score and *t*-score. The findings showed that bigrams that had the top *t*-scores were made up of frequent grammatical words such as prepositions, pronouns, determiners and auxiliaries as well as high-frequency lexical verbs (e.g., *think, get, want, say*). Moreover, there were many bigrams that were made up of preposition + determiner (e.g., *of the*, *in the*) or pronoun + verb (e.g., *it was, he was*). As

for the MI score, the top-scoring bigrams were made up of less frequent words in which most of them comprised noun + noun sequences (e.g., *rocket launchers, personality traits*) or adjective + noun sequences (e.g., *acid rain, alcoholic beverage*). The lowest-scoring bigrams identified using the MI score consisted of inaccurate combinations of grammatical words (e.g., *a out, there are*). 70 out of 200 bigrams found in the learner corpus that did not exist in the reference corpus were grammatically possible sequences. The researchers also provided pedagogical implications on the role of phraseology in the development of L2 writing and language teaching.

Next, Ruan (2016) studied the developmental patterns of LBs in Chinese L2 learner academic writing at four points between Year One to Year Four of their studies. Ruan (2016) found that more different bundles were used by the learners in Year 4 Final Year Project dissertations than the earlier points although there were fluctuations in the use of bundles in-between. The average token frequency of occurrence of the LBs in the written texts were said to decrease from Year One to Year Four. The researcher highlighted that the average token frequency of bundle types lessened as learners advanced to higher level in their studies. Similar to Hyland (2008a), Chen and Baker (2010) and Ädel and Erman (2012) who found two sequences, *at the same time* and *on the other hand* as the most frequent bundles used in academic writing, this study also noted the frequent occurrence of these bundles in the writing of learners over the years. Structurally, the learners first made use of VP-based bundles, NP-based bundles and eventually PP-based bundles. Functionally, the learners' written texts contained a high use of discourse organizing bundles and a low use of stance expressions. Ruan (2016) highlighted that the learners tended to rely on repeated use of a narrow range of bundles in their academic writing at their lower level of studies. However, they made use of more LBs types as they progressed to the higher level of studies. It was inferred that

learners' level of academic studies might have a great influence in the range of LBs used in their academic writing.

Crossley and Salsbury (2011) conducted a longitudinal corpus study to investigate the development of bigram accuracy in the spoken data of six L2 learners of English over a year. They compared the use of bigrams in learner data to the use of bigrams in NS data. They also studied how the frequency of LBs used changed over time. Crossley and Salsbury (2011) hypothesized that as L2 learners tend to produce NS-like language when they obtain lexical proficiency. It was found that three learners showed significant connections between the Test of English as a Foreign Language scores and time spent learning English. As learners became proficient in English their bigram accuracy increased as well. The results revealed that the L2 learners showed an increase in the accuracy of bigrams over time. The findings also demonstrated that learners produced bigrams that developed over time which paralleled with the frequency of production of NS. The learners made use of more common bigrams and less use of uncommon sequences. The findings of this study also support the notion that words are not acquired individually but through having the knowledge of word combinations.

Elturki and Salsbury (2015) investigated the use of three-word and four-word FS and the attitudinal information (i.e., semantic prosody) attached to these sequences in the spoken and written discourse of an individual learner of English. This case study involved a single participant named Eun-Hui from Korea who is in America for her undergraduate studies whose speech and writing were observed for a year. Elturki and Salsbury (2015) also investigated the strength of the relationship between the FS and the semantic prosodies. The results revealed that the spoken discourse of the learner had more FS than the written discourse. This was said to be because of the demands of conversation in spoken language. In trimester 3, it was found that the spoken data contained significant links of FS to prosodies compared to the written data. It was

inferred that this scenario was due to her written discourse becoming more impersonal and less attitudinal. In the written and spoken discourse, the most frequent sequences contained negative prosodies such as expressing deficit, disapproval and obligation. It was also found that the sequences that were linked to the prosodic information were unique to a particular learner and also closely related to his or her developing identity of the language being acquired. Elturki and Salsbury (2015) also demonstrated the benefit of using longitudinal data to study the development of these sequences used by the same learner over a period of time which is not possible with cross-sectional data.

As mentioned above, among the few corpus-based studies that investigate LBs in the Malaysian context is the study by Chan, Hadi and Tan's (2014). The researchers looked at the frequency, structures and functions of three-word and four-word LBs in a corpus comprising 20 group discussion transcripts collected from Malaysian undergraduate university students. The findings revealed that the most frequent three-word and four-word bundles that topped the lists were *I think that* and *I agree with you* respectively. The structural analysis revealed that the bundles were dominated by VP, NP and PP-based bundles which showed that the students made use of more phrasal bundles than clausal bundles whereas the functional analysis revealed that majority of the bundles used were referential bundles which was followed by stance bundles. This study suggests that LBs should be taught to less proficient ESL/EFL learners to enable them to engage in group discussions more effectively.

Another study is by Ong and Yuen (2015) that investigated the functional types of LBs found in MUET reading passages from two main disciplines (i.e., arts and science). Ong and Yuen (2015) found that despite the differences in the number of LBs found in both disciplines, there were quite a lot of similar LBs found in both disciplines. It was also found that the arts-based texts contained more participant-oriented bundles whereas the science-based texts contained more research-based bundles. Similar to Ebeling and

Hasselgård's (2015b) study, the findings of this study confirmed that LBs are discipline specific.

Most of the studies listed above have focused on LBs in the written and spoken language of advanced, adult learners instead of in the written and spoken language of secondary school students. Less is known about the use of LBs among students (see Leńko-Szymańska, 2014 for an exception). Even less is known about the use of LBs in the narrative texts of secondary school students especially in the Malaysian context (see Chau, 2008 for an exception). Therefore, the present study addresses this concern by examining the use, structures and functions of LBs in the written and spoken narrative texts of secondary school students over time. Furthermore, as reviewed in the earlier sections, LCR dealing with cross-sectional data has become a norm resulting in a deficiency of longitudinal learner corpus studies that investigate phraseology in language. The present study studies the use, structures and functions of LBs in a longitudinal corpus to ultimately understand the nature of language development of the students. In the subsequent section, the growth of the field of SLA, theories and notions pertaining to the learner and language development are discussed.

2.5 Second language acquisition

Having reached an extensive 50 years span, the field of SLA is well-known for its contributions even in the 21st century (Ortega, 2013). SLA research has evolved so much today and this is witnessed in the major shifts of ideas in the field especially pertaining to the learner and learner language. The three major developmental stages of SLA are said to be in the 1950s, 1960s and 1970s (Smith, 1994). These three important stages mark the evolvement of perspectives on language acquisition and development that is from the behaviourist approach to the mentalist approach and to the sociolinguistic approach respectively (Johnson, 2008).

During the 1950s, the behaviourist approach advocated the view that language is acquired through habit formation and learner errors could be avoided with continuous practice (Johnson, 2008). In the 1960s onwards, the behaviourist views in Skinner's *Verbal Behaviour* (1957) were strongly refuted by Chomsky (Johnson, 2008). Chomsky argued that language acquisition does not take place through habit formation but from rules formation (Larsen-Freeman & Long, 1991). His arguments included the fact that language acquisition is the result of the mental processes of the learner which he describes as the 'language acquisition device'. As opposed to the behaviourist approach which claims that the environment has substantial effect on language learning rather than the learner himself, the mentalists believe that language learning process primarily involves the learner and very little contribution of the environment. Gradually, Chomky's observations proved that the Contrastive Analysis (CA) Hypothesis which lies within the behaviourist approach was of no much value as learner errors were far more than just L1 interference in L2 learning (Mitchell, Myles, & Marsden, 2013). Following the mentalist approach three pioneering works concerning learner system were Corder (1967) (i.e., transitional system), Nemser (1971) (i.e., approximative system) and Selinker (1972) (i.e., interlanguage (IL)). Corder (1967) argued that learners acquire and develop in a language according to their internal system known as the 'built-in syllabus' despite external input given. Selinker (1972) introduced the notion of L2 learner's IL as a systematic and dynamic system used by learners to create patterns from the source and TL.

Although researchers began to view learner language as a separate system, the idea of errors in learner language proved that learner language production was still measured to NS language norms. These anti-behaviourist perspectives claimed that errors should not be seen from a negative lens but are pathway to understand how the learner acquires and develops in the language. The need to study learner error in its own right led to the

discovery of the EA. EA is the systematic investigation of L2 learner's errors which was pioneered by Corder (1967) (Mitchell et al., 2013). After some time, researchers debated that EA still emphasized on what learners did wrong and not on what they did right. EA also failed to provide solutions to the complications learners had in SLA classroom (Larsen-Freeman & Long, 1991; Mitchell et al., 2013). Other theories that advocated the innate biological endowment include Krashen's Monitor Model, Long's Interaction Hypothesis and Swain's Output Hypothesis. Schumann's Acculturation Model emphasized on the influence of social setting in L2 acquisition (Mitchell et al., 2013).

In 1990s onwards, SLA began experiencing a whole new shift of perspective (i.e., from the behaviourist, mentalist and to the emergentist approach). Larsen-Freeman (1997) first posited the chaos/complexity theory in which she argued that SLA process is complex, dynamic and nonlinear in juxtaposition to the traditional assumption that learner language development is a fixed and linear process. The complexity theory is an umbrella term that is used to include chaos theory, dynamical theory and complex system theory (Larsen-Freeman & Cameron, 2008). She also challenged the idea of viewing TL as the end-state of SLA process where the learner is believed to have completely acquired the language once he has conformed to the NS norms. Instead, she stated that development in SLA may not necessarily be fixed towards one direction; it may also encompass instability and change (Larsen-Freeman, 1997, 2006). Larsen-Freeman (2006) investigated the developmental process of the written and oral production of five Chinese learners of English over six months to understand the complex system of SLA. She observed the production of the same group of learners by administering the same task at all four points in time. She inferred that when the entire production of learners over time is taken into account it might seem that there are no changes taking place. But when the focus is narrowed to one subject at one point in time

there are variations noted in learner language production over time. This suggests that the system is not fixed. Larsen-freeman (2006, p. 612) explains,

> I will use an image from chaos/complexity theory, that of a fractal […]. A fractal is a geometric figure that exhibits self-similarity at different levels of scale or magnifications. One important dimension of self-similarity is that it suggests that even when things appear to be static at one level, there may be continuous dynamics within the system at another level, just as we might not see a tree growing if we watch it, even for a long time, yet we know, in fact, that at another level of scale, there is a great deal of growth taking place.

Challenging the early theories in the field which saw language learning process as transmitting knowledge, researchers advocating the sociocultural theory (e.g., Nieto, 2009) saw the learner as a product of his social and cultural identities as well as his political background who is actively involved in the process. Researchers like Ellis, O' Donnell & Romer (2013) focused on usage-based theory. Bilingualism and multilingualism started to become the focus of researchers which further complicated the comparison of learner language to NS system. Cook (1992, 2012) proposed the term 'multicompetence' to explain the state of mind of a speaker who speaks more than one language. He argued that the bilingual or multilingual learner (i.e., L2 learner) is completely different from the monolingual NS. Creese and Blackledge (2010), Garcia (2014) and Li (2017) demonstrated the concept of 'translanguaging' in the language learning process to illustrate the interdependence of knowledge across languages by bilinguals and multilinguals. Researchers also questioned that idea of L2 learner errors (Larsen-Freeman, 2011). Ortega (2013) depicted the complexities in studying human language by illustrating the interconnections between eight different types of acquisition or loss of language in the language acquisition process. Chau (2015) investigated the language development of Malaysian secondary school students over time with the use of a longitudinal corpus and a cross-sectional corpus. In his study, he discussed about the idea of 'languaculturing' where he argued that the language user ought to be seen as a languaculturing being who is engaged in a dynamic meaning making process. The

studies noted above further challenged the idea of comparing of learner language to NS standard in SLA research.

Today to some extent although not completely SLA research field is slowly but steadily heading towards a direction in liberating the language learner from the homogenous NS language norms. As discussed above, the comparison of learner language to the yardstick of TL is highly debated as it portrays learners as inadequate beings and empty vessels who fail to 'pull themselves up by their bootstraps' or 'join the mainstream' to be on par with NS (Cook, 1992; Nieto, 2009). The learner is compelled to give up his own language, literacy and culture at the expense of learning another person's language, literacy and culture (Nieto, 2009). The treatment of learner language in its own right is important because the comparison of it to the idealized NS standard in many ways has not only demeaned the language production of the learner, but has also ripped his identify by requiring him to forget and put aside the collective life experiences gained throughout his life just when he comes to the English language classroom. This is because wanting one to forget his other language experiences is equal to erasing his identity. This statement in no way intends to discount the works that adopt the target-like perspective to examine the learner language. This kind of studies is indeed needed for students who intend to become like NS in their language use. As far as this study is concerned, it is the researcher's intention (along with the studies highlighted above) to treat learner language in its own right.

2.6 Conclusion

The present study in some ways differs from the past learner corpus studies as highlighted in the first chapter as well as in the current chapter. First, learner corpus studies dealing with the spoken language of learners are considerably fewer than the learner corpus studies dealing with the written language of learners. As an attempt to fill

in the gap, this study looks at both the written and spoken language of students of English. Second, a great number of studies in the past have investigated phraseology in the language of adult, advanced learners of English. Little is known about the phraseological use of secondary school students of English. The present study investigates the use of LBs in the written and spoken narrative texts of 42 secondary school students of English over time. Third, unlike the most of past studies that made use of cross-sectional data to study the developmental processes of language or FS, in this study a longitudinal corpus is compiled to observe the nature of language development of the students based on the use, structures and functions of bundles found in the written and spoken narrative texts of the students. In the next chapter, a detailed account of the methodology used for the study is provided.

CHAPTER 3: METHODOLOGY

3.1 Introduction

CL method and tools have opened a whole new possibility of descriptive language study as discussed in Chapter 2. A growing body of learner corpus studies have employed methodologies such as CIA and combination of experimental method (Callies, 2015) as well as EA to investigate phraseology in learner language (Lu, 2010). However, the corpus-driven approach is said to be the most heuristic method to investigate continuous sequences of words (i.e., LBs) in language use (De Cock, 2004; Biber, 2009).

This study takes on a corpus-driven approach using frequency data to investigate the LBs in the written and spoken narrative texts of the students over time. The corpus-driven approach is argued to be more inductive than the corpus-based approach (Biber, 2009). The former allows linguistic items to emerge from the analysis itself. The latter, however, is employed to investigate the use of the pre-defined linguistic features (see Biber, 2009 for a detailed explanation on corpus-driven and corpus-based approaches to study FS). The use of frequency data is more likely to unveil significant patterns which may have been otherwise obstructed by intuition-based interpretation or grammatical categories recognized by traditional linguistic research (Biber et al., 2004; Conrad & Biber, 2005; Hyland, 2008a; Conrad, 2010). Time to time, the validity of frequency data has been questioned. To some extent, it is agreeable that quantity may not always mean quality and further in-depth analysis is required to distinguish the quality of the sequences yielded (Huang, 2015). In this study, frequency data acts as the deciding factor of sequences that qualify as LBs in order to investigate the use of these sequences as building blocks in the written and spoken narrative texts of students. The longitudinal research design (i.e., observing the language use of the same group of students over

time) is adopted as it aids in identifying the developmental patterns in the use of LBs over time. As opposed to the cross-sectional data, the longitudinal data is seen as the most appropriate source to observe developmental patterns (Cortes, 2004; Larsen-Freeman, 2006; Chau, 2012, 2015). This study comprises quantitative and qualitative data analysis. The quantitative analysis includes counting the raw frequency of use as well as the normalized frequency of use of LBs at three points over time. The qualitative analysis covers close text observation to identify the structures and functions of bundles used in written and spoken narrative texts.

3.2 Participants

The participants of the study are 42 students of Sungai Tiram secondary school situated in Ulu Tiram, Johor, Malaysia, whose written and spoken narrative texts form the written corpus and the spoken corpus of this study respectively. The participants comprise 16 males and 26 females. Initially, there were 46 students. However, the number of students decreased to 42 over time due to absenteeism. These students are from the first two classes of Form Four. They are aged 16 years old. Out of the 42 students, 40 of them are Malays and 2 of them are Indians. It is understood through the statements of the students that their mother tongues are Malay and Tamil respectively. They are studying the English language as a second language in the school. As pointed out in Chapter 1, prior studies on phraseology have primarily focused on adult learners rather than secondary school students (Ebeling & Hasselgård, 2015a). Thus, as an attempt to fill in the gap, this study investigates the use of LBs in the language of secondary school students of English.

3.3 Corpus design

The two corpora developed for this study consist of the written and spoken narrative texts of students of English collected at three points in time across six months. The written and spoken corpus components used for the present study are part of a larger corpus project compiled by the researcher that spans 12 months. The written corpus consists of 126 written narrative texts or 53,658 words (tokens). The spoken corpus consists of 126 spoken narrative texts or 40,082 words (tokens). The written and spoken data were collected at three points in time: April 2017 (i.e., Time 1), July 2017 (i.e., Time 2) and October 2017 (i.e., Time 3). Each text is coded ranging between 01-42, 42 indicating the total number of participants. The written narrative texts are coded with 'W' and the spoken narrative texts with 'S' to indicate the differences between written and spoken narrative texts respectively. This is followed by the code of the participant. For example, W01 and S01 indicate the written and spoken narrative texts of the same participant. The coding of data also involves reference to the three different data collection points across six months. The data collected at Time 1, Time 2 and Time 3 are referred as 'a', 'b' and 'c' respectively. For instance, the written narrative texts of two different students collected at the same point in time are referred as W01a and W02a, the codes W01 and W02 indicating the written narrative texts of two different students and 'a' referring to the data collected at Time 1 (i.e., April 2017). On the other hand, the written narrative texts of the same student collected at two different points in time are coded as W01a and W01c, where 'a' represents April 2017 and 'c' represents October 2017. The similar coding method is applied for the spoken narrative texts as well.

The students were required to complete two tasks, one written and one spoken. Following the common Sijil Pelajaran Malaysia (SPM) English examination format 'Section B: Continuous Writing', the students were required to write a narrative ending

with the line: "*It was the happiest day of my life."* The instruction for the students was to write not fewer than 350 words. The researcher selected the written and spoken task that was familiar to students in order to create an environment similar to the language classroom. This measure was taken to avoid collecting data from a highly controlled setting which may not closely reflect the nature of language use in the language classroom (Granger et al., 2015). The narratives were all written within one hour without the aid of reference materials. As for the spoken task, they were required to speak on the same topic given for the written task. The spoken task was completed after two weeks from the written task was completed. A three months gap was given between the first point and second point of data collection and another three months gap was given between the second point and third point of data collection, all totalling up to six months period. Ortega and Iberri-Shea (2005) note that past longitudinal studies in SLA research field usually span between three months to six years. They also state that there are no clear cut answers to the optimal length of observation for a longitudinal study. But one common goal of longitudinal studies is to observe change over time. This study takes on a six months period of observation of the written and spoken language of students. It is believed that investigating the use of LBs in learner language across six months would provide some interesting observations on the developmental patterns in the language use of the students. Moreover, Larsen-Freeman (2006) has shown that it is possible to examine the developmental patterns of the written and spoken language of students within a six months period of observation.

A repeated-task design is used in this study (i.e., the same topic is given for the written and spoken tasks at all three points in time) to facilitate comparability in the use of LBs from one point in time to another (Larsen-Freeman, 2006; Chau, 2015). This is because different task types would possibly demand the use of different linguistic features disrupting the comparability in the use of LBs over time. Furthermore, the

focus of this study is to observe the use of LBs as building blocks in learner language and not to examine the effects of different task types on the use of bundles. Larsen-Freeman (2006, p. 595) highlights the benefit of using the repeated-task design:

> …using the same task several times was one way of dealing with the fact that 'even subtle differences in a task can affect performance profoundly' (Thelen and Corbetta 2002: 61), leaving unanswered the question of whether the subject has control over the language resources or not. I wanted to be able to look at performance variability that might be an 'important harbinger of change, or indeed the manifestation of the very process of change' (Thelen and Corbetta 2002: 61), not variable performance that could be due to differences in tasks or contexts.

One of the caveats of administrating the repeated-task design is that students might run out of ideas when they are required to complete the same task over time. Recalling Sinclair's statement to 'trust the text', in this study, the written sub-corpora and spoken sub-corpora only expanded over time (refer to Tables 3.1 and 3.2 below). This shows that use of the repeated-task design did not exhaust the production of students over time.

The written data were typed and saved as electronic texts whereas the spoken data were recorded, transcribed and saved as electronic texts for the analysis. The sample narrative texts taken from the written and spoken corpora are presented below. The written and spoken narrative texts are coded as 'W02a' and 'S02a' respectively, the former indicating the written narrative text of student 02 produced at Time 1(a) and the latter, indicating the spoken narrative text of the same student produced at Time 1(a):

Sample written and spoken texts of student 02 at Time 1 (*a*)

W02a

Go to holiday

Last year, my grandmother was planing go te holiday at the beach with my family and cousin. After finish the school, my family and I went to the beach at 2.00 p.m. After arrive the beach, I saw my cousin also went to holiday at the beach. Air at the beach very fresh.

At the 3.00 p.m ; I ate for to play football with my brothers and cousins. I don't know my friends at there. My friend also went to holiday at the beach. I am so happy because my frrend at there. My brothers and I were playing football with my cousins. They are so expect when they were playing football. I feel so happy because I saw my brothers and cousin so happy when they were playing the football yet my team lose. When my friend arrive to play with me.. When we tired, we stopped to play the football and we were eating the chickens. We were washing at the beach. During I was, my cousin taught how right to swim. I was learning to swim but first time I learn to swim it is so difficult yet it is not dificult if I try to swim and never give up. My brothers were playing the ball at the beach and like to brother me when I was wash but I know what my brother do it is joke.

After that, I was eating again because I very hungry. I saw my brother ate very fast and ate very much. I saw my family and cousin smile because they are very happy. After that, my family and I were arranging the things to back home but I saw my brother was in a rush. I must help my brother to arrange the things. My father was so angry because my brother was in a rush. After finish arrange, we was back home. My cousin very finish when they were arranging the things to back home and not in a rush.

During journey to back home, I saw the many sport cars. They were drivering very fast and I saw my favourite car is Lancer Evo. I feel very happy because I can hung out with my family go to holiday. It was happiest day of my life.

S02a

Last year my grandmother first person to plan holiday at Desaru. When I back home when finish the class my family and I go the beach. My cousins are also go holiday at the beach. It time to year when come at the Desaru my brothers and I plan to play footballs and with my cousins. At there my friends also to holiday at the beach. I playing football with them. I see they're very happy and I feel very happy. But they're very expert to playing football. I cannot to playing with them. Once their wins playing football with me I so happy when I see their smile because after that I watch with my family. After that my father is cook the chickens. I eat then my brothers eat the very much because their very hungry after playing long time because their tired after playing football. After finish that I too continue to playing game with them. I see they are very happy because they can release tension playing football. I very happy because I can release my tension to back to school. After finish all I ready to back home. I see my brothers very fast because they really want to sleep at home because are very tired and then went to back home I very happy because my brothers are playing quite friendly very happy because their can hang out with me and it's was happiest days of my life.

The details of the written and spoken corpora, with the number of students, texts, word tokens and types are provided in Tables 3.1 and 3.2 below.

Table 3.1: Written corpus

Time	No. of students	No. of texts	No. of word tokens	No. of word types
1 (April 2017)	42	42	14,288	1,730
2 (July 2017)	42	42	19,138	2,129
3 (October 2017)	42	42	20, 232	2,211
Total	42	126	53, 658	6,070

Table 3.2: Spoken corpus

Time	No. of students	No. of texts	No. of word tokens	No. of word types
1 (April 2017)	42	42	11,344	1,285
2 (July 2017)	42	42	12, 643	1,376
3 (October 2017)	42	42	16, 095	1,607
Total	42	126	40, 082	4,268

3.3.1 Challenges faced during the compilation of the spoken corpus

The compilation of the spoken corpus on top of the compilation of the written corpus was a very challenging task. The researcher faced a number of difficulties in collecting the spoken data. First, the researcher was only allowed to administer the spoken tasks during the English periods. As a result, the collection of the spoken data of 42 students at one point lasted for 3 separate days. This resulted in a methodological concern as there was a possibility for students who completed the spoken task in the subsequent days to have prepared their speech. However, it was understood through the statements of the students they did not pre-plan their speech at any point in time. Second, the collection of the spoken data was a lengthy process which consumed a lot of time especially in terms of data transcription. A minimum of 45 minutes was required to

transcribe a five-minute speech of a student. Third, the researcher had to make some important decisions during the transcription of the spoken data. The researcher decided to transcribe the spoken data without strictly adhering to the spoken transcription conventions. Strict pronunciation conventions were not considered when transcribing the data. Short and long pauses and 'errs' made by students were not included in the transcription as well. The students also reconstructed sentences a few times in their speech. The final sentence uttered after several attempts of reconstruction was taken as the intended sentence. To illustrate, the pause and reconstruction of sentence that were omitted in the speech of student 01 at Time 1 are shown below (in underlines):

> We suppose to leave at home ahh early but my father could not find the key. […] It was [pause] hap it was happy at there. (S01a)

This decision was made as the main purpose of collecting the spoken data is to understand the use of LBs as building blocks in the students' speech rather than to investigate the features of speech of these students. Apart from that, words that were not clear in the speech were indicated using the mark-up convention '<unclear>' for the readability of the software. The spoken data was transcribed according to the standard spelling convention. All the texts in the corpora were kept untreated for the conventionally acknowledged grammatical errors (for the written and spoken narrative texts) and spelling errors (for the written narrative texts) in order to retain the originality of the texts. This is in line with the stand of the study to treat learner language as a distinct system from the NS system. These are some of the challenges faced during the compilation of the spoken corpus.

3.4 Identification of lexical bundles

There are a few criteria used for the identification of LBs such as the frequency cut-off, dispersion and the length of bundles (see Chen & Baker, 2010 for a detailed

explanation on the operationalization of LBs). The frequency criterion (i.e., the minimum number of times a bundle should occur in a corpus) and range or dispersion criterion (i.e., the minimum number of occurrence of a bundle in different texts in a corpus) are the two most important criteria used to decide whether or not a sequence qualifies as a bundle. Researchers argue that the frequency and dispersion criteria used to identify LBs are rather arbitrary (Biber et al., 1999, 2004; Conrad & Biber, 2005; Biber & Barbieri, 2007). The identification criteria of LBs also vary according to the corpus size used for a study. For instance, Biber et al. (1999) included bundles that occurred 10 times in a million words, Biber et al. (2004) set the cut-off of 40 times per million words and Hyland (2008a) counted in clusters that occurred 20 times per million words. The dispersion criterion is also considered important because it guards against individual writing or speaking style (Biber et al., 2004; Conrad & Biber, 2005; Chen & Baker, 2010). For example, Biber et al. (1999, 2004) included bundles that occurred in at least 5 different texts whereas Hyland (2008a) included clusters that occurred in at least 10% of the texts. Considering the sizes of the corpora used for this study which are rather small, sequences that recur two times in at least two different texts are identified as LBs for the analysis.

This study looks at the use of four-word bundles in the written and spoken narrative texts of students. The length of LBs usually ranges from two-word to six-word bundles. But four-word bundles are by far the most researched length because four-word bundles carry wider range of structures and functions in comparison to three-word bundles (Hyland, 2008a). They are also 10 times more common than five-word bundles (Cortes, 2004). Four-word bundles are said to be within a manageable size for manual analysis (Chen & Baker, 2010).

3.5 Ethical considerations

Appropriate ethical measures were taken into account when the study was conducted. First, the researcher obtained permission to conduct the study in the intended school from the government departments, Kementerian Pendidikan Malaysia and Jabatan Pendidikan Negeri Johor. Permissions to have access to the school and participants were obtained from the school authority. Consent of participants and their parents or guardians was also obtained before collecting the data. The right to withdraw from the study at any stage was explained to the participants prior to data collection. Data generated from the participants who withdrew from the study were obliterated. Participants' anonymity was kept by assigning codes instead of their real names.

3.6 Procedure of data analysis

3.6.1 Research question 1: The identification of lexical bundles in the written and spoken corpora

The four-word LBs in the written and spoken corpora were extracted according to the frequency and dispersion criteria set (i.e., bundles that occur twice in at least two different texts) using 'n-Gram/Cluster' list in the AntConc (Anthony, 2014) concordance programme. The raw frequency of the bundle types and the overall frequency of bundles were normalized to occurrence per 1000 words to facilitate comparability between the sub-corpora (Biber & Barbieri, 2007). The findings on the normalized frequencies of bundle types and overall bundles used are presented in Chapter 4.

In contrast to the studies by Biber et al. (2004), Chen and Baker (2010) and Ädel and Erman (2012), topic-dependent bundles and overlapping bundles were not removed from the list of four-word LBs extracted from both corpora. Biber et al. (2004), Chen and Baker (2010) and Ädel and Erman (2012) excluded topic-dependent bundles in their

analysis to avoid the idiosyncrasies introduced by the topics. However, De Cock (2004) and Hyland (2008a) included topic-dependent bundles in their studies. Similarly, the researcher decided to retain the topic-dependent bundles since the data used for the study is based on one topic. In addition to that, the aim of this study is to observe the use of bundles as building blocks of learner language development rather than to merely examine the discourse functions of LBs. Hence, a decision was made to include topic-dependent bundles. Overlapping bundles too were included for the analysis. For example, close text analysis revealed that the bundles, *happiest day of my* (W01a) and *the happiest day of* (W01a) were parts of the five-word bundle, *the happiest day of my* (W01a). These two bundles were presented as two separate bundles in the analysis. A concern arose that this might affect the normalized frequency count of the bundles. The occurrence of overlapping bundles were not restricted to Time 1 but were also noted at Time 2 and Time 3 in the written and spoken sub-corpora resulting in a consistent observation.

Contractions (e.g., *can't*, *don't*, *didn't*) were counted as single words in the analysis. Unlike past studies, bundles that incorporate punctuation marks were included in the analysis as well. For instance, bundles containing a comma (e.g., *after that, my family*), apostrophe (e.g., *I got 6a's in*), quotation mark (e.g., *"it was the happiest*) and full stop (e.g., *day of my life.*) were included. Bundles that incorporate parts of two different sentences (e.g., *go back home. We*) were included in the structural analysis but were excluded in the functional analysis. These bundles were identified as intersentential bundles under Pattern 6a in the modified structural framework (refer to Table 3.3 below). In the study by Biber and Barbieri (2007), sequences that spanned punctuation marks and a turn boundary were not treated as LBs as they were considered as interrupted sequences. Biber and Barbieri (2007) also suggested that future researchers should consider investigating bundles that incorporate punctuation marks. Thus, in this

study, the researcher included sequences with punctuation marks and sequences incorporating parts of two different sentences in order to avoid restricting LBs to only uninterrupted sequences.

3.6.2 Research question 2:

3.6.2.1 The identification of the structures of lexical bundles

The structures and functions of LBs found in the written and spoken corpora were identified following Biber et al.'s (1999, 2004) structural and functional frameworks. Modifications were made to the existing frameworks due to differences in the corpora (i.e., Biber et al.'s (1999, 2004) corpora consist of conversation, classroom teaching, academic prose and textbooks whereas the corpora built for the present study consist of narrative texts). Given the differences in the aim of the study, corpus size and representativeness between Biber et al. (1999, 2004) and the present study, it was essential to do a close text observation to identify the structures and functions of bundles found in this study.

The structural classification of LBs was done by the researcher followed by 20% of the bundles extracted from the corpora which were classified by two inter-raters to ensure the reliability of the findings. The researcher and inter-raters reached about 90% of agreement in which the remaining 10% of differences were discussed and agreed upon. Based on the structural analysis, six main structural categories were identified: (1) VP fragments, (2) DC fragments, (3) NP and PP fragments, (4) AdjP fragments, (5) AdvP fragments and (6) intersentential bundles. The complete modified structural framework is presented in Table 3.3.

Table 3.3: Modified structural framework of lexical bundles

1.	**Lexical bundles that incorporate verb phrase (VP) fragments**
1a.	(connector +) 1st/3rd person pronoun + VP fragment Example bundles: *and I went to, she went to the, it was the happiest*
1b.	(connector +) pronoun/Noun phrase + VP fragment Example bundles: *because this is my, family and I went, a big smile plastered*
1c.	Copula be + Noun phrase/Adjective phrase Example bundles: *was the happiest day, are a lot of, am very happy because*
1d.	(connector +) VP fragment: Example bundles: *and go to the, packed all the thing, cooked my favourite food*
1e.	Preposition phrase + VP fragment Example bundles: *in my life was, by car and arrive, of my life was*
1f.	Adjective phrase (with VP fragment) Example bundles: *I was so nervous, we are so happy, excited because this is*
2.	**Lexical bundles that incorporate dependent clause fragments**
2a.	(connector +) 1st/3rd person pronoun + dependent clause fragment (Example bundles: *and I go to, he want to go, I was going to*
2b.	WH-clause fragments: Example bundles: *audience were wonder who, when I arrive at, don't know how to*
2c.	(connector +) *to*-clause fragments: Example bundles: *a chance to see, and get ready to, very excited to go*
2d.	(VP +) *That*-clause fragments: Example bundles: *told us that we*, *think that it was, that we go to*
3.	**Lexical bundles that incorporate noun phrase and prepositional phrase fragments**
3a.	(connector +) Noun phrase with *of*-phrase fragment: Example bundles: *a lot of food, day of my life, the end of the*
3b.	Noun phrase with other post-modifier fragment: Example bundles: *living room with my, happiest day in my, the first day at*
3c.	Other noun phrase expressions: Example bundles: *my family and I*, *days and one night*, *last year my family*
3d.	(connector +) Prepositional phrase expressions: Example bundles: *and on the evening*, *at the same time, in front of the*
4.	**Lexical bundles that incorporate adjective phrase fragments**
4a.	(connector +) Adjective phrase expressions: Example bundles: *and I so happy, so nervous because I, very fast and I*
5.	**Lexical bundles that incorporate adverb phrase fragments**
5a.	(connector +)Adverb phrase expressions: Example bundles: *other than that we*, *back to the hotel*, *and after that we*
6.	**Lexical bundles that cross sentence boundaries**
6a.	Intersentential bundles: Example bundles: *(with my family. We), (are very happy. After)*, *(me. It was a)*

These structural categories identified are referred to below.

A total of six subcategories of use of VP fragments were identified:

1a. (connector +) personal pronoun is followed by a VP fragment

1a. (1) (connector +) 1st person pronoun is followed by a VP fragment

e.g., Last year, my family and I go to the beach at Tanjung Balau.

1a. (2) 3rd person pronoun is followed by a VP fragment

e.g., It was the happiest day of my life.

or

1a. (3) 1st person pronoun + adverb is followed by a VP fragment

e.g., I also have a happiest moment in my life.

1b. (connector +) pronoun or NP is followed by a VP fragment

1b. (1) pronoun is followed by a VP fragment

e.g., Five students was very excellent in pt3 at my school and this is my turn to take my result pt3.

or

1b. (2) NP is followed by a VP fragment

e.g., During the school holiday, my family and I went visited Cameron Highlands.

1c. copula *be* is followed by NP or AdjP

1c. (1) copula *be* is followed by NP

e.g., It was the happiest day of my life.

or

1c. (2) copula *be* is followed by AdjP

e.g., I am very happy because i got 5A in PT3.

1d. (connector +) VP fragment

e.g., I am very happy because I can spend time with my family members.

1e. PP is followed by a VP fragment

e.g., There are a lots of happy momments in my life was happened but I have a few memories that I cant forget.

1f. AdjP with VP fragment

e.g., We arrived at 12 o' clock and I was very excited because I can keep the experience as a memory in my life.

Bundles that incorporate DC fragments represent four subcategories:

2a. (connector +) personal pronoun is followed by a DC fragment

2a. (1) 1st person pronoun is followed by a DC fragment

e.g., On the end of November in 2013, I was anxious and excited because I was going to get my UPSR result on that day.

or

2a. (2) 3nd person pronoun is followed by a DC fragment

e.g., Two years ago, my father was plan want to go to Cambodia because he want to visit her mother and her father there.

2b. WH-clause fragment

e.g., Every audience were wonder, who the lucky person that get straight A's.

2c. (connector +) *to*-clause fragments

2c. (1) *to*-clause fragment

e.g., After we spent about an hour there, we decided to go to the green tea farm too.

2c. (2) NP is followed by a *to*-clause fragment

e.g., I would call the festival as the happiest day of my life cause I got a chance to see my cousin Naren from Japan.

2c. (3) VP is followed by a *to*-clause fragment

e.g., We took some rest and get ready to go to SMK Temin Baru school for having a dinner and ice-breaking with the SMK Temin Baru students.

2c. (4) AdjP is followed by a *to*-clause fragment

e.g., I was very excited to go for Langkawi because I never go Langkawi before.

These patterns were only found in the spoken narrative texts of the students:

2c. (5) PP is followed by a *to*-clause fragment

e.g., So my teacher told to me to create a design. (S24a)

or

2c. (6) NP + VP is followed by a *to*-clause fragment

e.g., So, she cried and my father asked me to buy ice-cream and give to her.

2d. (VP +) *that*-clause fragments

2d. (1) *that*-clause fragment

e.g., I thought that I will get bad result as on the trial UPSR, I only got 2As 2Bs 1C.

or

2d. (2) VP is followed by a *that*-clause fragment

e.g., I am very sure that the activities will be more fun if my mother join together.

Bundles with NP and PP fragments consist of four subcategories:

3a. (connector +) NP with *of*-phrase fragment

e.g., So yes, I really think that it was the happiest day of my life.

3b. NP with other post-modifier fragments

3b. (1) NP with PP fragment

e.g., After we finish do activities, we take a rest and take a shower at the shower room.

or

3b. (2) NP with relative pronoun *that*

e.g., The Place that we went is Legoland.

3c. other NP expressions

3c. (1) NP fragment

e.g., After that, my mother and aunty chatting at the kitchen meanwhile my sister and I watched Running Man together with our cousins together in the living room.

or

3c. (2) (connector +) NP fragment

e.g., After that, my brother asked me to help him to bring out the thing that my father asked.

3d. (connector +) PP expressions

3d. (1) PP fragment

e.g., On the next day, we packed our things to back home.

or

This pattern was only found in the spoken narrative texts of the students:

3d. (2) (connector +) PP fragment

e.g., We take some rest and on the evening we played some games like badminton.

The use of AdjP and AdvP fragments were identified and grouped into two different categories:

4a. (connector +) AdjP expressions

4a. (1) AdjP fragment

e.g., I very happy because he say he want to buy my design.

or

4a. (2) (connector +) AdjP fragment

e.g., My cousin and I very happy and they swim at the swimming pool and that place no many people and we enjoyed.

5a. (connector +) AdvP expressions

5a. (1) AdvP fragment

e.g., First and foremost, we booked 4 comfortable homestays.

or

5a. (2) (connector +) AdvP fragment

e.g., After we all see the football match we all having a dinner and after that we check in hotel back.

Bundles that cross sentence boundaries were grouped as intersentential bundles:

6a. Intersentential bundles

e.g., My brother help me to put our bags in the car. We started our journey in the early morning.

The information on the distribution of structural categories of four-word LBs in the written and spoken corpora is provided in the following chapter. The findings on the structures of LBs in the written corpus were compared to the findings of structures of LBs in the spoken corpus to find out how similarly or differently they were used in both corpora.

3.6.2.2 The challenges in identifying the structures of lexical bundles

It is important to note that the identification of the structures of LBs found in both corpora was not straightforward. As highlighted earlier, although Biber et al.'s (1999, 2004) structural framework was adopted in this study there was a need to do modifications to the existing framework due to the differences in the corpora. The researcher also faced further complications when classifying bundles with learner language features that were not present in Biber et al. (1999, 2004). As can be seen, there are two separate AdjP categories allocated for LBs with AdjP fragments (i.e., see Table 3.3: Pattern (1f) AdjP with VP fragment (e.g., *I was very happy*) and Pattern (4a) AdjP without VP fragment (e.g., *I very happy because*)). To illustrate, sequences incorporating copula *be* followed by AdjP as in, *I was very happy* would be conventionally identified as AdjP with VP fragment. But AdjP without a copula *be* in the case of the bundle, *I very happy because* does not come under the same category of AdjP with VP fragment due to the absence of a verb. Therefore, bundles with the structure, AdjP without VP fragment (e.g., *I very happy because*) are referred as innovative forms that display learner language features. As highlighted earlier, the present study methodologically treats learner language as an independent system whereby it does not measure the language use of students to the NS system. The term 'conventional forms' is used to refer to bundles with structures that are recognized by traditional grammar whereas the term 'innovative forms' is used instead of 'errors' to

refer to bundles with structures that do not fit into the conventionally acknowledged grammatical structures.

A challenge arises when the analyst has to decide if she would interpret these two LB types (i.e., *I was very happy* and *I very happy because*) from a structural or a semantic perspective. If she was to adopt the structural perspective, two different structures represented by two different bundle types are to be identified. However, if she adopts the semantic perspective, the meaning of the bundles becomes the focus and therefore both bundle types would be identified as one category representing the same meaning. The analyst makes a difficult decision to interpret these two bundle types structurally whereby, *I was very happy* and *I very happy because* represent two different structures as in, Pattern 1f and Pattern 4a respectively. The structural perspective is adopted because the aim of the study is to identify the structures of the LBs found in their written and spoken narrative texts. From a semantic perspective, grouping *I was very happy* (i.e., AdjP with a verb) and *I very happy because* (i.e., AdjP without a verb) under one category would further complicate the analysis of the use of VP-based, DC-based and NP/PP-based bundles. This is because VP-based and DC-based bundles are clausal bundles (i.e., sequences that incorporate a verb component) and NP/PP-based bundles are phrasal bundles (i.e., sequences that do not incorporate a verb component) (see Biber et al.'s (2004) structural framework for explanation). It was also well thought that the interpretation of bundles semantically would be somewhat similar to the identification of the functions of bundles which is to be analysed in the following section. Therefore, the researcher decided not to perform two similar analysis. There is no right or wrong answer. But noteworthy is the fact that this decision has been consistently applied to other instances of use as well.

The next decision to make is what is to be treated as a connector? Coordinators or coordinating conjunctions (i.e., *and*, *but*, *or*) are usually known as connectors (Kennedy,

2003). Biber et al. (2004) treated 'and' and 'well' as connectors. In this study, there are many more instances of word fragments apart from 'and' and 'well' which occurred in the beginning of VP, DC, NP/PP, AdjP and AdvP fragments that were not present in Biber et al. (1999, 2004). The analyst has a difficult decision to make in order to classify these word fragments. The analyst decided to treat coordinating conjunctions, subordinating conjunctions, linking adverbials and circumstantial adverbials occurring in the beginning of VP, DC, NP/PP, AdjP and AdvP fragments found in the written and spoken corpora as connectors. There is possibly no right or wrong answer but the analyst has consistently applied this decision to all instances of use in the written and spoken narrative texts.

The connectors found in the written corpus are provided first followed by the connectors found in the spoken corpus. The coordinating conjunctions (1) that occurred in the written narrative texts are *and* and *but* which connect two sentences at the same clause level:

(1) My family and I were so happy *and* It was the happiest day of my life. (W37a)
But, I do not like to pack my clothes as I will confused and it takes time too. (W13a)
We go together to buy a ticket, at the airport my sister *and* I was very excited because this is my first time travel to Cambodia. (W38c)

The subordinating conjunctions (2) *after*, *because*, *before*, *eventhough*, *when*, and *while* connect two sentences of different clause levels (i.e., independent and dependent clauses):

(2) *After* we arrived at the hotel, My father check-in to the room. (W26a)
This is *because* I got 7A's in my PT3 results. (W41b)
Before we go to the beach, my sister and I helped my mother to packing clothes, food and others things. (W15c)

Next, linking adverbials (3) *after that*, *lastly* and *then* are used to connect the ideas expressed in order to ensure cohesion of the narrative:

(3) *After that*, we went to the lavender garden.(W42a)
Lastly, my family and go back to hotel and packed our things, take some rest. (W39a)
Then, we go to kayak for snorkelling at the far place with my sister. (W36c)

The fourth type of connector, circumstantial adverbials (4), *first time*, *last week*, *last year*, *the activities*, *the morning*, *the night*, *school holiday*, and *sunny da*y are part of the main clause which are used to provide information on questions like when, where and what in the narrative:

(4) *Last week*, My family and I went for camping at Desaru Beach. (W21a)
Last year, my family members celebrated the day at Tanjung Balau Beach. (W09a)
At *the morning*, my mother cooked some food to eat at the beach. (W25b)

Furthermore, other noun fragments in the written corpus that occurred in the beginning of VP, DC, NP/PP, AdjP and AdvP fragments were treated as connectors as well. For instance, in *Last year, my family and I had an exciting day at Sunway Lagoon* (W29a), the word, *year* is part of the NP, *last year*. Since, *last year* is treated as a connector (see example 4, line 2 above), the fragment *year* is also considered as a connector. This decision has been consistently applied to all instances of other noun fragments (i.e., *day*, *holiday*, *hour*, *old*, *that*, *time*, *week*, *years old*) as in, *After 5 day we go to airpot and take a flight from UK to Klia* (W26c).

The connectors identified in the spoken corpus occurring in the beginning of VP, DC, NP/PP, AdjP and AdvP fragments include the coordinating conjunction (5) *and*, the subordinating conjunctions (6) *after*, *because*, *before*, *even though* and *when*, the linking adverbials (7) *and after*, *and then*, *after that*, *then* and *so* and the circumstantial

adverbials (8) *after dinner*, *next day*, *last day*, *last week*, *last weekend*, *last year*, *that day*, *the evening*, *the night* and *then after*. These four types of connectors found in the spoken corpus are referred to below.

(5) We take some rest *and* on the evening we played some games like badminton. (S13a)
After two weeks my sister *and* I go to my brother house to discuss about the party again. (S14c)
Two years ago my family *and* I went to the Tioman Island. (S36c).

(6) *After* we arrive at my uncle home we take rest. (S21b)
Before we went to the Cameron Highland we pack our clothes into the bag and my brother was so excited to go Cameron Highland. (S42b)
So when we move here *even though* we are in the same country our accents is a bit different because of our states. (S10c).

(7) *And then* we go to cowboy show. (S06a)
After that my father and my mother and my family and I go to dinner. (S26a)
And after we finish eat we go to playing in the water. (S25c)

(8) *Last weekend* I go to holiday with my family. (S06a)
Next day we go to Johor Premium. (S22b)
And *then after* we arrive at Cambodia my aunty was pick me and my family to my grandmother's house. (S38c)

Other noun fragments that were found in the spoken corpus that were treated as connectors are *day*, *evening*, *holiday*, *home*, *melaka*, *night*, *park*, *that*, *year* and *years ago*. For example, the word, *day* which is part of the NP *next day* as in, *On the next day we go to A'Famosa resort which is a very known place and interesting* (S06b) was considered as a connector.

Apart from that, the analyst has another important decision to make in order to classify bundles that begin with a single word adverb (e.g., *back to the hotel*) and

bundles with connective function that are not part of VP, DC, NP/PP and AdjP fragments (e.g., *other than that we*). A difficult decision was made to allocate a separate category (see Table 3.3: Pattern 5a) in order to group bundles that are made up of only adverbials. The LBs grouped as AdvP expressions constitute single word adverbs or adverb phrases that modify sentences adverbially as well as adverbials with linking function.

3.6.2.3 The identification of the functions of lexical bundles

The functional classification of LBs was done by the researcher. Biber et al.'s (2004) functional framework was loosely adopted in this study to which modifications were made due to differences in the corpora as mentioned above. About 20% of the bundles extracted from the corpora were classified by the inter-raters to ensure the reliability of the findings. The researcher and inter-raters reached about 90% of agreement. The remaining 10% of differences were discussed and agreed upon. Based on the functional analysis, five main functional categories of bundles were identified: (1) stance expressions, (2) discourse organizers, (3) referential expressions, (4) topic-oriented expressions, and (5) bundles with special conversational functions. The complete modified functional framework is presented in Table 3.4.

Table 3.4: Modified functional framework of lexical bundles

Functional categories
1. Stance expressions **A. Epistemic stance** • Comments on the knowledge status (certain, uncertain, probable or possible) of the information in the following proposition. • Example bundles: *I think it was, I am very sure, I hope we can* **B. Attitudinal/modality stance** • Expresses the writer, speaker or character(s)'s attitude towards actions or events described in the following proposition. o B1) desire • Expresses desire or wish to perform an action. Example bundles: *because I want to*, *want to go to*, *and he want to* o B2) Obligation/directive • Expresses obligation or direction to perform an action. • Example bundles: *have to go to, brother asked me to, because we have to* o B3) Intention/prediction • Expresses intention to perform an action or prediction of a future action. Example bundles: *can spend time with*, *I can't wait to, because he will go*
2. Discourse organizers **A. Transition** • Serves to indicate transition of the events in the written or spoken narrative texts. • Example bundles: *after that my family, other than that we*, *first and foremost we*
3. Referential expressions **A. Identification/focus** • Serves to highlight on a significant event, animate, inanimate or abstract entity. • Example bundles: *happiest day in my*, *my family and I*, *fresh air at the* **B. Place reference** • Serves to point to a place setting or location. • Example bundles:, *go back to hotel, go to the beach, in the living room* **C. Time reference** • Serves to point to a time setting. • Example bundles: *at the same time*, *in the morning I, at 8.00 a.m. I* **D. Quantity specification** • Serves to specify a quantity or an amount. • Example bundles: *buy some things for*, *a lot of food*, *some food to eat*

Table 3.4, continued

Functional categories
4. Topic-oriented expressions **A. Depiction of action/state** • Serve as expressions to indicate an action or state of the writer, speaker or character(s) as well as the state of an inanimate entity. • Example bundles: *we go to the*, *my friends and I were*, *I saw my mother* **B. Depiction of feelings/emotions** • Expresses the feelings or emotions of the writer, speaker or character(s). • Example bundles: *I was very excited*, *I am so happy*, *so nervous because I* **C. Elaboration/clarification** • Serves to elaborate or clarify the main content of a sentence or the prior sentence. • Example bundles: *it is because my*, *to go to my*, *even though we were*
5. Special conversational functions **A. Reporting** • Serves to report or inform about something that has been done. • Example bundles: *he told me that*, *I said to my*, *my father said to*

These functions of bundles are referred to below with examples from the written and spoken narrative texts respectively. First, two types of stance expressions (1) were identified: Epistemic stance (1a) and attitudinal or modality stance (1b). The epistemic stance bundles comment on the knowledge status (i.e., certain, uncertain, probable or possible) of the information in the following proposition:

(1a) Being on the first class, 6 Jaya, *I am very sure* that all the teachers put their hope on us, the students of 6 Jaya. (W16a)
It was the happiest day of my life and *I hope we can* go holiday next time. (W01b)
In the flight, *I think It was* the happiest day of my life. (W26c)

One by one my friends being called to the front to get their result and my name will be called too and I was so surprised because *I didn't thought that* I will get 5A as my trial UPSR were bad. (S16b)
That day I woke up early in the morning because *I know I will* go to the most famous theme park in the world USS or Universal Studios of Singapore. (S20b)

I walk unto the stage with my mom and I heard applause from the students *and I think it* was the happiest moment in my life. (S08c)

Epistemic bundles found in both corpora were all personal (i.e., involving the writer or speaker). The attitudinal or modality stance bundles express the writer or speaker's attitude towards actions or events described in the following proposition. Personal and impersonal (i.e., involving other characters in the narratives) bundles with this function were identified. Attitudinal or modality stance bundles are used to express desire (1b1), obligation or direction (1b2) and intention or prediction (1b3):

(1b1) I am very excited to take my result *because I want to* know how much A, B, or C in my result. (W18a)
Soon, after the session of school end, I was decided to back home and *want to tell my* parents about the event. (W30b)
Last year, *my family want to* go picnic at Batu Layar. (W01c)

It is *because we want to* spend a lot of time because my father will have a lot of works to do after this. (S12a)
After my teachers gave my result *I don't want to* see that because I think that I got bad because when I answer the questions it's too hard for me. (S30b)
Last month my father *want to go to* Batu Layar with my family. It is because my father want to spend time with a family. (S01c)

(1b2) Then, we were opened anything at motor, *my brother asked me* to follow him bought a motor's thing. (W03c)
After finish played *my father asked me* to clean ourself. (W25c)
After we have a dinner we go back to home and have a sleep *because we have to* make sure our energy full to do the next activities. (W38c)

We stay at there at one week and after that we go to the beach and play go to the playground and my aunty *ask me to go* to the MRT station. (S23a)
So I also had to tell them that I have to move to a another school which is here SMK Sungai Tiram. (S10b)
That morning all of the standard 6 students *have to go to* the school hall. (S16c)

(1b3) I am very happy because I *can spend time with* my family members. (W15a)
And as usual, *I can't wait to* open all the gift that I've got and ask a help from my siblings to open it too. (W17b)
My *father said we will* go tomorow. (W26b)

It was the happiest day in my life because I *can spend time with* my family then I get a lot of present from my family members. (S11a)
So my uncle took us back to home and I think I *will go to the* USS by my own money with my family. (S20b)
I very hope *I can make my* parent happy for my SPM so I target I got 9A to my SPM. (S41c)

Second, bundles with discourse organizing function (2) serve to indicate transition (2a) of the events in the written or spoken narrative texts:

(2a) I help my aunty prepare the food. *After that, I go* to take my grandfather in the car. (W25a)
We catch the jelllyfish using my cloth and we put in the bucket. *After that, my family* and I taking the photos for our memories and put in frame. (W28b)
So, we always study together and always tried to help each other in every situation. *Other than that, we* also go to tuition class together. (W07c)

The movie is my favourites movie and after that we go to the zoo and my first experience I've no see the white tiger and I see at my eyes *and after that we* go to the underwater and I see on my eyes the big fish. (S23a)
After that we go to the strawberry farm. We pick our own strawberry. My sister took many strawberries because she love to eat strawberries very much. *After that we went* to the butterfly farm. (S42b)
After that I help my mom to wash all the dish. *And then I help* my father wash his car. (S35c)

Third, referential expressions (3) used for identification or focus (3a) function to highlight on a significant event (3a1), animate (3a2), inanimate (3a3) or abstract entity (3a4):

(3a1) It was *happiest day of my* life. (W02a)
I was very excited because *this is my first* time. (W38b)
We have to take a boat as there was no land road. By the way, *it was my first* time. (W13c)

The best day of my life when I got the PT3 result last year. (S27a)
It was a *happiest day in my* life when I got good result in PT3. (S30b)
The happiest moment that I still remember with extra fondness and make me feel more matured is when I follow my father to continue he's study in Iran. (S34c)

(3a2) *My family and I* went from Johor to the Melaka by a car. (W22a)
I ate the fried noodles and thanks *to my mom because* she cooked my favourite food. (W35b)
Before we go to the beach, *my sister and I* helped my mother to packing clothes, food and others things. (W15c)

Last weekend *my family and I* go to the trip at Batu Layar. (S01a)
Then I saw *my mother and my* young brother walk passing the hall and they smile at me and I just thought in my mind did I just get 5A too. (S16b)
My brother and I was so happy. (S26c)

(3a3) We play *the ball at the* beach. (W25a)
When we arrive, we packed all *the thing such as* food, clothes and camp at car. (W01a)
The Place that we went is Legoland. (W39b)

Then one of them was blowing *the things that we* call as a trumpet I don't know name but I just call it trumpet. (S17a)
After that I just join *the activity and my* brother and my father go to fishing. (S38b)

I have best experience in my life is my friend and I go to *other country such as* Brazil and Japan. (S21c)

(3a4) I am very happy because I can spend *time with my family members.* (W15a)
I'm a little upset on that time then I took a decision to get some *fresh air at the* window. (W17b)
It was *a sunny day, my* family and I were planned to do something that we never do the school holiday. (W40c)

It was my happiest day *in my life that* I can't forget ever. (S06a)
My happiest day *of my life was* going to Rio de Janeiro and Hong Kong with my friend. (S21b)
Even that holiday short I very happy. And this *was the happiest in* my life. (S31c)

Referential expressions that function as place reference (3b) serve to point to a place setting or location:

(3b) Lastly, my family and I *go back to hotel* and packed our things, take some rest. (W39a)
After 8 hours, just stay and sit *in the car, we* reached infront the grandparent's house that located at the Jalan Tok Bok, Kuala Krai. (W04b)
It was a sunny day, Zamri and I very bored at home and we suggest to go to *the Tanjung Balau beach.* (W21c)

When I *arrive at Melaka we* check in hotel and first we rest. (S31a)
Before that my father make a BBQ and we celebrate together *at the beach and* we do a lot of activities that I love like collecting sea shells, playing volleyball, strolling around the beach. (S11b)
After that we *go back home and* take a shower and rest because we must go to Japan tomorrow. (S21c)

Bundles with time reference (3c) serve to point to a time setting and bundles with quantity specification (3d) are used to specify quantity or amount of anything:

(3c) At that time I felt that I was so lucky to have met them when I moved to Johor. *At the same time* I wondered that I'll be missing all of this if I weren't friends with them. (W10a)
In the morning, we went by car at 8 a.m. with my family members to the beach. (W11b)
After we arrived, we go in the house to seeing our grandparents. (W04c)

So *the next day we* went to Art Festival at Pasir Gudang. (S24a)
Then *at 4.00 p.m. we* take lunch at Legoland. (S39b)
And on *the evening my father* took us to having a lunch in one restaurant at Melaka which the foods there was very nice. (S06c)

(3d) Ling said that her new school in KL is very productive and that there *are a lot of* artistic students there. (W10a)
I plan to buy the shoes to my father's birthday but I not *have a lot of* money. (W14b)
After we take some food, we go swimming and *we do some activities.* (W28c)

My mother prepare *a lot of food* to eat. (S11a)
Then when I came into school *I see many people* already go into the school hall. (S08b)
We *saw a lot of* animal like snakes and many more. (S37c)

Fourth, topic-oriented expressions (4) that depict an action or state (4a) are used to indicate an action (4a1) or state (4a2) of the writer, speaker or character(s) as well as the state of an inanimate entity (4a3):

(4a1) *We arrived at the* klia at 9.00 am and go to cafe have a breakfest. (W26a)
At the night, *we take a walk* around the village. (W13b)
We go to the beach by car. (W15c)

Then before we go home *we go to the* restaurant first to lunch. (S12a)
After 20 minute at the restaurant *we start our journey* to Star Hill. (S27b)
After that *we go to shopping* mall at Jusco. (S22c)

(4a2) We arrived at 12 o' clock *and I was very* excited because I can keep the experience as a memory in my life. (W03a)
My family and I were so excited to go to the Malacca. (W37b)
My *friends and I were* so excited and felt enjoyful. (W37c)

My *family and I was* very excited because can get new experience. (S36a)
The happiest day of my life was when I was around 3 years old *and my brother was* 5 years old. (S29b)
My *mother and I was* busy to cook my chocolate cake. (S42c)

(4a3) It *was a sunny day,* Zamri and I take a decision to go to the vacation at Hong Kong and Rio De Jeneiro. (W21b)
We went to Waterplex cinema 5D and watched Red Beard story and it *was so cool* and awesome cause the seats were moving and there was water splitting an over the place. (W29b)
We go to our room the room *was very beautiful and* the bed was very soft. (W39c)

The *food was so delicious* and we enjoyed it. (S42a)
And at *the place is very* happy and very big. (S23c)
In here *the place is very* beautiful scenery for holidays with family and have many activities we can do. (S36c)

Bundles that depict feelings or emotions (4b) are used to express the feelings or emotions of the writer, speaker or character(s) and elaboration or clarification bundles (4c) are used to elaborate or clarify the main content of a sentence or the prior sentence:

(4b) Two years ago my family and I went to holiday at Desaru. *I was very excited* to go to my holiday. (W38a)
When the day comes, I was so nervous I think that other people can hear my heart beat. (W10b)
My father said that he will took us for vacation next time if have a chance, so *we are so happy* to hear that. (W06c)

So when I came into her house it was like *I was so happy* because been so long since I've seen Uzmah. (S10a)
We were very excited but the situation is crowded because there a lot of people especially Chinese. (S37b)
I so happy because my father first time make delicious fish. (S01c)

(4c) Other than that, we went to the stall beside the road. *It is because my* father wanted to buy honey for him. (W42b)
Then, my friends asked me *to go to the* stage, after I walked along the hall, just in a sudden, my mom appeared and hug me tightly. (W08b)
But i was the sad *because it is the* last day in Japan. (W18c)

Then my best friends one more best friends Rashidah her name got called for 5A and I congrats her *because I was so* proud of her. (S16a)
So we decided in the next month there is Mother's Day we plan *to celebrate it and* make her happy. (S17b)
So even though Uzma decided to go to another secondary school we exchange phone number so that we can stay in contact *even though we are* far way. (S10c)

Fifth, bundles with special conversational function (5) serve to report or inform (5a) about something that has been done:

(5a) Our teacher *told us that we* had to send our painting and design to the organisor. (W24b)
The next day *my parent told me* that they are bringing me to Sunway Lagoon and I was shocked. (W29b)
We checked in a resort because *my father said we* looked very tired. (W12c)

Then *he said to me* he want to buy my design. (S24a)
I said to my mother I will fishing after I play a search treasure with my brother. (S01b)
He told me that when he was young they were friends and now only they got in contact. (S29c)

The information on the distribution of functional categories of four-word LBs in the written and spoken corpora is provided in the next chapter. The findings on the functions of LBs in the written corpus were compared to the findings on the functions of LBs in the spoken corpus to examine how similarly or differently they were used in both corpora.

3.6.2.4 The challenges in identifying the functions of lexical bundles

The identification of the functions of bundles was not straightforward. For instance, the analyst faces a challenge when she has to decide if she would interpret the bundle, *after that, my brother* as a discourse organizer or a referential bundle as this bundle is made up of fragments with both functions (i.e., *after that* is used to indicate transition whereas *my brother* is used to refer to a person). The analyst then decides to classify *after that, my brother* as a discourse organizing bundle as the main function of the bundle is to show transition from one idea to another. There is no right or wrong answer but the decision made is consistently applied to other instances of use as in, *after that, we went* may also function as a topic-oriented expression that is used to depict an action. But this bundle type was classified as a discourse organizer. However, bundles such as *that, we go to* is interpreted as functioning to depict an action although it contains the fragment, *that* which is part of the sequence *after that* because this bundle closely represents the function of depicting an action rather than functioning to ensure cohesion of the text.

Next, the analyst has another difficult decision to make in terms of interpreting bundles with learner language features such as *I went to go* and *I very happy because*. The functional characteristics of bundles were identified through a semantic perspective. In other words, the researcher went about identifying the functions of bundles by looking at the context in which they were used. For instance, on the surface,

I went to go would simply mean *went* as the past tense of going somewhere. But by adopting a semantic perspective, the bundle *I went to go* is in fact used to mean the desire to go:

> Finally, we go at to the car and come in to the house. We family are very tired but so happy because whole family can spend time together. I think, *I went to go* again to Malacca zoo and saw again. As a simple, it was the happiest on my life. (W03a)

Therefore, this bundle was classified as an attitudinal stance expression that is used to express desire. Moreover, although the bundle, *I very happy because* is interpreted structurally different from *I was very happy* as discussed above, functionally these two bundles are used to depict the feelings of the writer or speaker. Thus, these two bundles were grouped under the same functional category as these bundles are used to depict feelings or emotions (see Table 3.4: Category 4b). The researcher also decided to have a separate category to classify topic-oriented expressions that is not present in Biber et al.'s (2004) functional framework. Apart from that, another confusion arises when identifying the function of the bundle, *family and I went* whereby, the researcher has to decide if this bundle would be called a referential bundle since it makes reference to a group of people (i.e., *my family and I*) or a bundle that depicts an action since it is made up of the verb fragment, *went*. A decision is then made to classify this bundle as a topic-oriented expression that functions to depict an action. There is perhaps no right or wrong answer to the decision made but this is consistently applied to all the other instances of use found in the written and spoken narrative texts. These are some of the challenges faced during the functional analysis of the bundles.

3.6.3 Research question 3: The nature of language development

The nature of language development was observed based on the changes in the use of LBs over time as well as based on the structural and functional analysis of LBs in the written and spoken corpora. The researcher also performed a simple analysis by

adopting two different perspectives to analyse learner language. An EA was performed to identify the errors that occurred in the use of AdjP-based bundles in the structure, '1st person pronoun + copula *was/were* + AdjP' (e.g., *I was very happy*) in the written and spoken corpora over time. Overlapping AdjP-based bundles were not included in this analysis. The researcher identified the errors produced by the students in the use of AdjP-based bundles in their written and spoken narrative texts over time. The erroneous forms were also verified by a senior English teacher from SMK Sungai Tiram with 21 years of experience in the teaching field. Dulay, Burt and Karshen's (1982) surface structure taxonomy (as cited in Ellis & Barkhuzien, 2005) was used to describe the AdjP-based bundles with errors. This was followed by another analysis on the same AdjP-based bundles from the perspective of treating learner language in its own right where frequency of conventional forms and innovative forms of the same AdjP-based bundles used in the written and spoken corpora over time were analysed. As noted earlier in this chapter, in line with the orientation of this study which methodologically treats learner language in its own right, the term 'conventional forms' is used to refer bundles with structures that are acknowledged by traditional grammar whereas the term 'innovative forms' is used instead of 'errors' to refer to bundles with structures that do not fit into conventionally acknowledged grammatical items. A comparison of the findings based on these two different analysis was done to understand the difference between measuring learner language based on NS norms and treating learner language in its own right. Insights drawn from the two different analysis on AdjP-based bundles are discussed in Chapter 4.

3.7 Conclusion

The present study made use of two corpora of 126 narrative texts each to investigate the use of LBs in the written and spoken language of students over time. LBs with a

frequency of two occurrences or above in at least two different texts were extracted from each sub-corpus. The normalized frequency of bundle types per 1000 words and overall frequency of LBs normalized per 1000 words were obtained. The structures and functions of LBs found in the written and spoken corpora were analysed. The lists of LBs generated from both corpora were compared to find out any similarities or differences in the use of LBs in the written and spoken narrative texts of students. The narrative texts were collected from the same group of 42 students at three points in time within a six months period: April 2017, July 2017 and October 2017, to observe the nature of language development using longitudinal data. In the following chapter, the findings and discussion of the study is presented.

CHAPTER 4: FINDINGS AND DISCUSSION

4.1 Introduction

In this chapter, the findings of the study are reported where the findings on the use of LBs in the written and spoken corpora over time is presented. This is followed by the findings on the structures and functions of LBs and a discussion on how similarly or differently these bundles are used in written and spoken narrative texts of the students. The findings based on the two different approaches (i.e., EA vs. analysis of learner language in its own right) employed to analyse AdjP-based bundles is presented. A discussion on the nature of language development is then presented based on the observations of the use, structural and functional analysis of four-word LBs as well based on the insights drawn from the two different approaches that were used to analyse AdjP-based bundles in both corpora.

4.2 Research question 1:

1. What are the most frequent four-word LBs that occur in the written and spoken narrative texts of the students over time?

4.2.1 The use of lexical bundles in students' written and spoken narrative texts over time

As noted in Chapter 3, the raw counts of LB types (i.e., different LBs) used in the written and spoken corpora were not directly comparable due to the varied sizes of the written and spoken sub-corpora. The raw counts of LB types were normalized to frequency of occurrence per 1000 words in a sub-corpus. This measure was done to facilitate comparability of the findings between the sub-corpora of different sizes (Biber & Barbieri, 2007; Biber & Conrad, 2009). The complete lists of LBs extracted from the

written and spoken corpora with a frequency of two occurrences or above in at least two different texts are provided in appendices A and B due to space constraint. The raw counts of LBs and the normalized frequencies of LB types per 1000 words in the written and spoken narrative texts over time are presented in Figures 4.1 and 4.2 respectively.

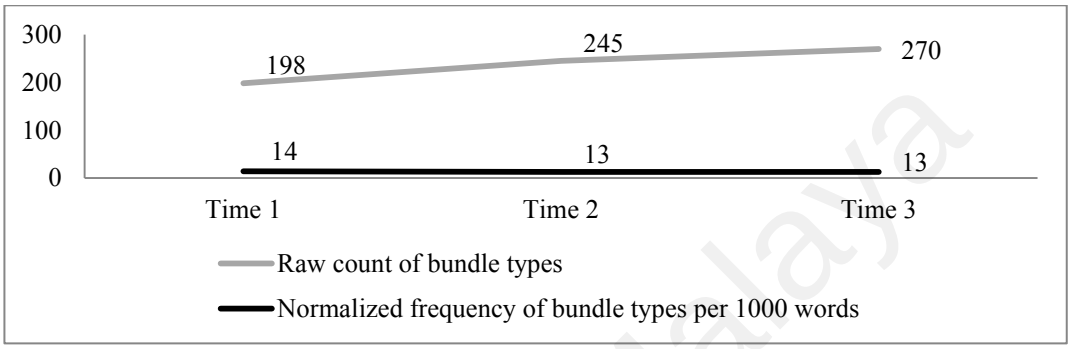

Figure 4.1: Raw count and normalized frequency of four-word lexical bundle types per 1000 words in the written narrative texts over time

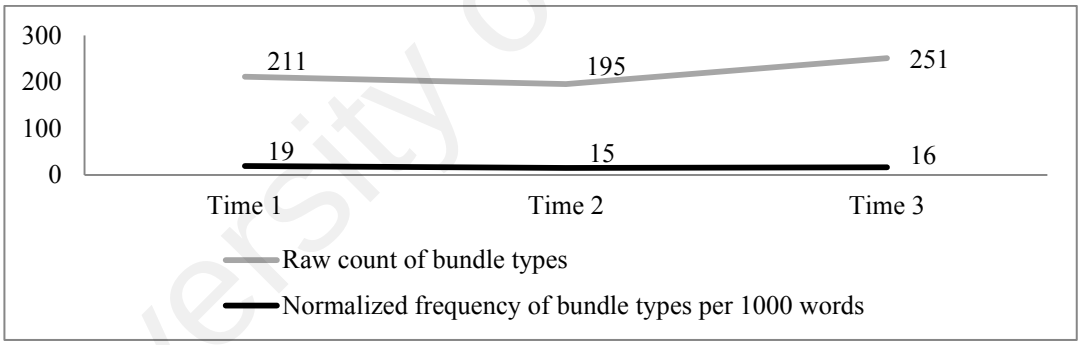

Figure 4.2: Raw count and normalized frequency of four-word lexical bundle types per 1000 words in the spoken narrative texts over time

The highest range of LBs is found in the spoken narrative texts at Time 1 accounting for 19 bundle types per 1000 words. Despite the fluctuation in the normalized frequency of LB types, it can be said that there are slightly more LB types found in the spoken narrative texts in comparison to the written narrative texts at all three points in time. In addition, the overall frequencies of occurrence of the bundles were analysed to investigate how frequently these bundle types occurred in the written and spoken

corpora. The overall frequencies of bundles in the written and spoken narrative texts normalized to per 1000 words are displayed in Figure 4.3. It was found that the LBs occurred slightly more frequently in the spoken narrative texts over time although they were almost as frequent in the written narrative texts.

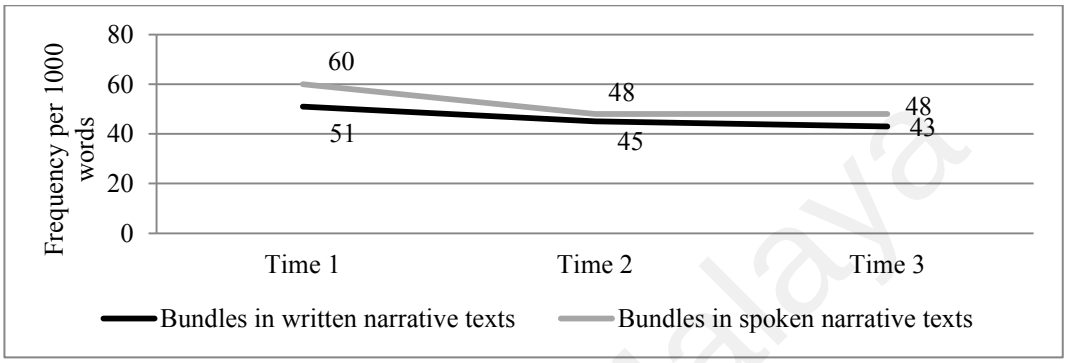

Figure 4.3: Overall frequency of four-word lexical bundles normalized per 1000 words in the written and spoken narrative texts over time

The overall frequencies of bundles normalized to per 1000 words show a gradually decreasing pattern from Time 1 to Time 3 in the written corpus whereas a drastic fall is noted from Time 1 to Time 2 which stagnates at Time 3 in the spoken corpus. With minimal differences in the use of overall frequencies of bundles (i.e., normalized to per 1000 words) in the written and spoken corpora over time, it can be said that the students made use of LBs slightly more frequently in the spoken narrative texts in comparison to the written narrative texts at all three points in time. However, they became less reliant on LBs in their written and spoken language over time.

Noteworthy is the fact that the findings here are in parallel with the findings of past studies. Biber et al.'s (1999, 2004) cross-sectional studies found that the spoken registers made use of larger stock of bundles than the written registers. Elturki and Salsbury's (2015) longitudinal study revealed that the spoken data of an individual learner contained more three-word and four-word FS than the written data. An inverse relationship was noticed by Elturki and Salsbury (2015) where the types of FS in the

speech of the learner increased as the types of FS in the writing of the learner decreased over a year of observation. However, in this study, the use of bundle types in the written and spoken narrative texts show a dissimilar pattern whereby bundle types decrease with slight variance from one point to another in the written corpus whereas bundle types fluctuate in the spoken corpus over time. The comparison of the findings of this study to the findings of past studies may not be direct due to the differences in the nature of the studies, but to a certain degree these studies share a common aspect of investigating written and spoken data.

Due to space limitation, the 50 most frequently used bundle types in the written and spoken narrative texts over time are presented in Tables 4.1 and 4.2 respectively in frequency order, with the recurring bundles indicated.

Table 4.1: The 50 most frequent four-word lexical bundles in the written narrative texts over time

Rk.	Time 1	Frq.	Time 2	Frq.	Time 3	Frq.
1	my family and i	44	my family and i	45	happiest day of my	38
2	happiest day of my	37	it was the happiest	37	was the happiest day	38
3	the happiest day of	36	was the happiest day	37	the happiest day of	37
4	was the happiest day	36	happiest day of my	36	it was the happiest	36
5	it was the happiest	34	the happiest day of	35	day of my life.	33
6	day of my life.	32	day of my life.	33	my family and i	20
7	and i go to	12	and i go to	14	**my friends and i**	17
8	and i went to	12	and i went to	11	my brother and i	9
9	family and i went	11	i go to the	10	*to go to the*	9
10	family and i go	8	my friend and i	9	we go to the	9
11	i go to the	8	*to go to the*	9	we go back to	8
12	i went to the	8	we go to the	9	**we arrived at the**	7
13	my brother and i	8	we went to the	9	after that, we go	6
14	last year, my family	7	*my sister and i*	8	*and go to the*	5
15	after that, my family	6	family and i go	7	and put in the	5
16	**family and i was**	6	at the beach. my	6	because we want to	5
17	we go to the	6	family and i went	6	**i was very excited**	5
18	**i was very excited**	5	have a lot of	6	in front of the	5
19	**my friends and i**	5	that, we went to	6	it was a sunny	5
20	of smk sungai tiram	5	after that, we went	5	my father and i	5
21	**we arrived at the**	5	friend and i were	5	*my mother and my*	5
22	year, my family and	5	happiest day in my	5	*my sister and i*	5
23	**and it was the**	4	i went to the	5	am very excited to	4
24	family and i were	4	last year, my family	5	and get ready to	4
25	**i am very excited**	4	my cousin and i	5	**and i was very**	4

Table 4.1, continued

Rk.	Time 1	Frq.	Time 2	Frq.	Time 3	Frq.
26	we go back to	4	year, my family and	5	**family and i was**	4
27	after that, i go	3	*and go to the*	4	**i am very excited**	4
28	after that, we go	3	are a lot of	4	**i think it was**	4
29	after that, we went	3	*day of my life*	4	*i very happy because*	4
30	am very happy because	3	*i very happy because*	4	i was so happy	4
31	am very sure that	3	in the morning, we	4	in the car. we	4
32	**and i was very**	3	it is because the	4	my brothers and i	4
33	and we go to	3	my brother and i	4	*on the next day,*	4
34	are a lot of	3	*my mother and my*	4	think it was the	4
35	because it was my	3	*on the next day,*	4	was a sunny day,	4
36	can spend time with	3	this is my first	4	we went back to	4
37	go back home. we	3	we decided to go	4	after that, my brother	3
38	i am so excited	3	we go back to	4	after that, we check	3
39	i am so happy	3	after that, my father	3	after that, we started	3
40	i am very happy	3	after that, we go	3	after that, we take	3
41	i am very sure	3	after we arrived at	3	after that, we went	3
42	**i think it was**	3	brother and i go	3	and i went to	3
43	i was going to	3	day in my life.	3	**and it was the**	3
44	it is because my	3	do a lot of	3	and went to the	3
45	last week, my family	3	family and i arrived	3	back to the hotel	3
46	my father and mother	3	family and i were	3	because we have to	3
47	on the way to	3	for a while and	3	*day of my life*	3
48	other than that, we	3	fresh air at the	3	day, my family and	3
49	packed all the thing	3	get ready to go	3	do some activities. the	3
50	smk sungai tiram will	3	go back to the	3	excited because this is	3

Table 4.2: The 50 most frequent four-word lexical bundles in the spoken narrative texts over time

Rk.	Time 1	Frq.	Time 2	Frq.	Time 3	Frq.
1	after that we go	23	day in my life.	17	happiest day in my	22
2	we go to the	23	we go to the	17	day in my life.	18
3	my family and i	19	happiest day in my	16	it was the happiest	18
4	that we go to	16	after that we go	14	after that we go	17
5	happiest day in my	14	my family and i	13	was the happiest day	16
6	day of my life.	12	that we go to	12	the happiest day in	12
7	day in my life.	11	it was the happiest	10	to go to the	12
8	family and i go	11	was the happiest day	10	my family and i	11
9	happiest moment in my	11	the happiest day in	9	day of my life.	10
10	**moment in my life.**	10	that i have to	8	that we go to	10
11	happiest day of my	9	told me that i	8	we go to the	10
12	i go to the	9	happiest day of my	7	my brother and i	9
13	and i go to	8	we arrive at the	7	at the night we	7
14	was the happiest day	8	i go to the	6	happiest day of my	7
15	and after that we	7	*after we arrive at*	5	the happiest day of	7
16	it was the happiest	7	and after that we	5	**the next day we**	7
17	have a lot of	6	day of my life.	5	*after we arrive at*	6
18	**my happiest day in**	6	day we go to	5	*and it was the*	6
19	**my mother and my**	6	it was a happiest	5	after that we went	5
20	the happiest moment in	6	last year my family	5	and i was so	5
21	**the next day we**	6	*next day we go*	5	day we go to	5
22	to go to the	6	the happiest day of	5	got a good result	5

Table 4.2, continued

Rk.	Time 1	Frq.	Time 2	Frq.	Time 3	Frq.
23	day in my life	5	to go to the	5	help my father to	5
24	go to the beach	5	*want to go to*	5	i went to the	5
25	the happiest day in	5	we went to the	5	my father said we	5
26	then we go to	5	year my family and	5	**my mother and my**	5
27	after that we take	4	*and go to the*	4	*want to go to*	5
28	after that we went	4	*and it was the*	4	we can see the	5
29	day of my life	4	in the morning we	4	*we go back to*	5
30	me to go to	4	my friend and i	4	*a happiest day in*	4
31	other than that we	4	want me to go	4	and i go to	4
32	see a lot of	4	*we go back to*	4	and we arrive at	4
33	the happiest day of	4	when i arrive at	4	because we want to	4
34	was my happiest day	4	when we arrive at	4	day in my life	4
35	we went to the	4	*a happiest day in*	3	even though we are	4
36	a happiest moment in	3	a happiest moment in	3	get ready to go	4
37	after that my father	3	after that we take	3	go back to the	4
38	and then we go	3	after that we went	3	i help my father	4
39	and this is my	3	and i go to	3	in my life is	4
40	**and we go to**	3	*and i want to*	3	**moment in my life.**	4
41	at the beach. it	3	and i went to	3	**my happiest day in**	4
42	at the same time	3	and take a rest.	3	my sister and i	4
43	because i can spend	3	at the beach and	3	*next day we go*	4
44	best day of my	3	can make my family	3	to spend time with	4
45	day we go to	3	day in my life	3	after that after we	3
46	family and i was	3	day of my life	3	and after that we	3
47	for me because i	3	go back to the	3	*and go to the*	3
48	he said to me	3	happiness day in my	3	*and i want to*	3
49	i and my brother	3	have a lot of	3	and it is the	3
50	i very happy because	3	i can go with	3	**and we go to**	3

The recurrences of LB types at all three points are indicated in grey, recurrences at Time 1 and Time 2 are indicated in light grey, recurrences at Time 1 and Time 3 in bold, and recurrences at Time 2 and Time 3 in italics. As can be seen, in the written corpus, the top six bundles yielded at Time 1 recurred at Time 2 and Time 3. The most frequently used bundle in the written narrative texts at Time 1 and Time 2 is *my family and I* which occurred 44 times and 45 times respectively. This bundle is presented below with examples from the written narrative texts at Time 1 (1) and Time 2 (2) respectively:

(1) After finish the school, *my family and I* went to the beach at 2.00 p.m. (W02a)
Last years, *my family and I* went to Singapore. (W23a)
My family and I were so happy and It was the happiest day of my life. (W37a)

(2) Last year, *my family and I* went to holiday at Malacca for three days and two nights during end-year holiday's. (W06b)
My family and I arrived at Singapore one day earlier. (W20b)
When *my family and I* arrived in Iran, we took a hotel to stay in three days. (W34b)

At Time 3, the most frequently used bundles are *happiest day of my* and *was the happiest day* (3) which occurred 38 times each in the written narrative texts. The bundles are presented below with examples from the written narrative texts at Time 3 respectively:

(3) It was the *happiest day of my* life. (W04c)
So, the conclusion is it was the *happiest day of my* life. (W05c)
So yes, I really think that it was the *happiest day of my* life. (W08c)

For me, he is a best brother ever and it *was the happiest day* of my life. (W12c)
I say "It *was the happiest day* of my life. (W22c)
I think It *was the happiest day* in my life. (W24c)

Only about quarter of the bundles which occurred in the written narrative texts at Time 1 recurred at Time 2 and Time 3, sharing only 12 of the top 50 most frequent bundles (in grey). About nine of the top 50 bundles were shared at Time 1 and Time 2 (in light grey); eight were shared at Time 1 and Time 3 (in bold); and seven bundles were shared at Time 2 and Time 3 (in italics).

On the other hand, the most frequently used bundles in the spoken corpus are *after that we go* and *we go to the* (4) which occurred 23 times each at Time 1, *day in my life* and *we go to the* (5) which occurred 17 times each at Time 2 and *happiest day in my* (6) which occurred 22 times at Time 3. These bundles are referred to below with examples from the spoken narrative texts:

(4) *After that we go* for lunch at the A'Famosa Resort. (S06a)
After that we go shopping and go to swimming pool. (S31a)
After that we go to Salang by ferry. (S36a)

Then before we go home *we go to the* restaurant first to lunch. (S12a)
After the pray *we go to the* Museum Abu Bakar. (S22a)
We go to the beach because we want celebrate my grandfather birthday. (S25a)

(5) So it was my happiest *day in my life.* (S05b)
Then we go to the stage together and that time suddenly all the students clapped their hands and I think it was the happiest *day in my life.* (S08b)
After we finish it we going to home and at home my little brother and sister was very tired and I help my father to put the bag at the room and clean the car and that day I was very tired but it was my happiness *day in my life.* (S25b)

We go to the market to buy food. (S20b)
After that we gather together and *we go to the* restaurant first to take a breakfast. (S27b)
After that *we go to the* strawberry farm. (S42b)

(6) In my mind it was the *happiest day in my* life. (S01c)
My *happiest day in my* life is when is during my UPSR result giving ceremony. (S16c)
It was a *happiest day in my* life. (S42c)

There were about 17 bundles of the 50 most frequent bundles that were shared at all three points in time in the spoken corpus (in grey); seven were shared at Time 1 and Time 2 (in light grey); five at Time 1 and 3 (in bold); and eight at Time 2 and Time 3 (in italics). The findings on the use of top 50 LBs in written and spoken narrative texts show that the students seemed to have made less use of the same set of bundle types in their written and spoken narrative texts over time. This may suggest that they have employed new set of bundle types in the written and spoken narrative texts over time.

As displayed in Table 4.1, five bundles among the top six most frequent bundles in the written corpus at all three points in time are overlaps of a longer bundle, *it was the happiest day of my life* (W01a). These overlapping bundles are *happiest day of my*, *the happiest day of*, *was the happiest day*, *it was the happiest* and *day of my life* (see Table 4.1: Time 1). This could be due to the fact that as part of the written task students were required to write a narrative that ends with the line, *'It was the happiest day of my life'*. Hence, the finding on the occurrence of five overlapping bundles among the top six most frequent bundles over time is not surprising. On the other hand, as presented in Table 4.2, there are overlapping four-word bundles that are part of a longer bundle, *it was the happiest day of my life* (S07a) in the spoken corpus as well. But it is worth mentioning that these overlapping bundles did not occur within the top six of the 50 most frequent bundles in the spoken corpus. They were rather spread out taking on the highest to lowest ranks within the 50 most frequent bundles. For instance, *it was the happiest* occurred at the 16th rank at Time 1 (7 times), 7th rank at Time 2 (10 times), and 3rd rank at Time 3 (18 times) and *the happiest day of* ranked at the 33rd place at Time 1 (4 times), 22nd place at Time 2 (5 times) and 15th place at Time 3 (7 times) in the spoken corpus. This finding is unforeseen because the same topic administered for the written task was given for the spoken task as well. It appears that students did not strictly adhere to the requirement of the spoken task as they did for the written task which resulted in a high use of bundles related to the topic given in their written narrative texts. It could be that they were slightly more explorative in the choice of bundles used in their speech.

Closer scrutiny into the bundle types used also revealed that the students were slightly more explorative in the choice of bundles used in their spoken narrative texts than in their written narrative texts. To illustrate this, the normalized frequency of

occurrence per 1000 words of three bundle types, *day of my life*, *day in my life* and *moment in my life* in the written and spoken corpora over time are provided in Table 4.3.

Table 4.3: The normalized frequency of occurrence per 1000 words of *day of my life*, *day in my life* & *moment in my life* in written and spoken corpora over time

	Rk.	Time 1	Rk.	Time 2	Rk.	Time 3
Written corpus	6th	day of my life (2.24)	6th	day of my life (1.72)	5th	day of my life (1.63)
	-	day in my life (0)	43rd	day in my life (0.16)	107th	day in my life (0.10)
	-	moment in my life (0)	-	moment in my life (0)	-	moment in my life (0)
Spoken corpus	6th	day of my life (1.06)	17th	day of my life (0.40)	9th	day of my life (0.88)
	7th	day in my life (0.97)	1st	day in my life (1.34)	34th	day in my life (1.42)
	10th	moment in my life (0.62)	54th	moment in my life (0.19)	40th	moment in my life (0.25)

As evidenced above, the normalized frequency of *day of my life* shows that it occurred most frequently in the written corpus at all three points in time. This bundle was not used at a high frequency in the spoken corpus over time. Instead, the students made use of alternate bundle types, *day in my life* and *moment in my life* as replacements for *day of my life* in their spoken narrative texts. These two alternate bundles are derived from the longer bundle, *it was the happiest day in my life* (S09a) and *it was the happiest moment in my life* (S08c) respectively. The alternate bundle, *day in my life* occurred at a very low frequency in the written corpus in comparison to the spoken corpus whereas *moment in my life* did not occur in the written corpus at all.

Two inferences can be made through the observations stated above. First, the students tend to be less flexible in the use of LBs in their written narrative texts which could be the reason why they heavily depended on one preferred bundle, *day of my life* and made very less use of the bundle, *day in my life* in the written narrative texts. They seem to be more flexible in the use of LBs in their spoken narrative texts as they made alternative use of three bundle types, *day of my life*, *day in my life* and *moment in my life* in their spoken narrative texts to convey the same meaning. Second, the findings here unveil the relative demands of the writing and speech. During a face-to face interaction

the students have minimal time to prepare, plan and produce their speech (Wray & Perkins, 2000). As a result, despite being able to produce lengthy written narrative texts in comparison to spoken narrative texts that were short, they depended more on these four-word sequences that are stored in their memory as substitutes to make meaning in their speech rather than writing. The inference made here is also supported by the initial findings of this study where not only did the students use slightly more variety of LBs in the spoken narrative texts but they also used them slightly more commonly in the spoken narrative texts in comparison to the written narrative texts.

The findings on the high use of the bundle, *day of my life* in the written corpus than in the spoken corpus as well as the bundle, *moment in my life* that occurred only in the spoken corpus tapped the curiosity of the researcher to investigate further on that. The researcher performed an analysis to examine how the students described their *day* and/or *moment* in their written and spoken language. Therefore, the adjectives that occurred before the nouns *day* and *moment* in both corpora were searched and generated. The types of adjectives occurring before *day* and *moment* in the written and spoken corpora over time together with the raw frequency and normalized frequency of occurrence per 1000 words of the most common adjective in brackets are presented in Tables 4.4 and 4.5 respectively.

Table 4.4: Types of adjective occurring before *day* in the written and spoken corpora over time

	Written corpus	Spoken corpus
Time 1	best (1), exciting (1), **happiest (44) (3.08)**, historical (1), important (2), nervous (1), sunny (1)	best (5), happening (1), **happiest (26) (2.29)**, happiness (3), happy (2), important (1)
Time 2	best (1), celebration (1), excited (1), **happiest (49) (2.56)**, happiets (1), hard (1), lucking (1), lucky (1), nervous (1), remembrale (1), scary (1), special (1), sunny (2), tired (1), tiring (1)	best (1), **happiest (25) (2)**, happiness (3), happy (2), hot (1), nervous (1), sunny (1)
Time 3	anxiety (1), best (1), big (2), challenging (1), **happiest (42) (2.08)**, happy (1), rainy (1), special (1), sunny (6), tiring (1), wonderful (1)	best (1), big (1), extreme (1), forgettable (1), **happiest (31) (1.93)**, happiness (4), happy (1), memorable (1), short (1), special (1), sunny (1)

Table 4.5: Types of adjective occurring before *moment* in the written and spoken corpora over time

	Written corpus	Spoken corpus
Time 1	**best (3) (0.21)**, exciting (1), funny (2), happiest (1), happy (1), precious (2)	enjoyment (1), fantastic (1), **happiest (15) (1.32)**, happy (3), memories (1), sad (1), special (1)
Time 2	exciting (1), **happiest (3) (0.16)**, happy (1), lovely (1), memorable (1)	best (2), excited (1), great (1), happier (1), **happiest (5) (0.4)**, happily (2), happy (3), precious (1)
Time 3	best (3), funny (3), **happiest (6) (0.3)**, happy (3), joyful (1), memorable (1), sad (1)	enjoyable (1), **happiest (7) (0.43)**, happiness (1), happy (3), precious (1)

It was found that *happiest* was the most commonly used adjective occurring with *day* in the written and spoken corpora over time (refer to Table 4.4 above). However, *happiest* was more commonly used in the written corpus than in the spoken corpus over time. The normalized frequency of adjective types occurring with *day* showed a fluctuating pattern in the written corpus over time. There were 0.49 adjective types, 0.78 types and 0.54 types at Time 1, Time 2 and Time 3 respectively. In the spoken corpus, the normalized frequency of adjective types used with *day* increased over time. There were 0.53 adjective types, 0.55 types, and 0.68 types at Time 1, Time 2 and Time 3 respectively. Surprisingly, the students made use of slightly more variety of adjectives with *day* in their spoken narrative texts in contrast to their written narrative texts at Time 1 and 3.

As for the adjective types occurring with the noun *moment*, it was found that *happiest* occurred with *moment* more commonly in the spoken corpus than in the written corpus over time (refer to Table 4.5 above). The students also made use of slightly more adjective types with *moment* in the spoken corpus than their written corpus at Time 1 and Time 2. For example, there were 0.42 instances of adjective types per 1000 words, 0.26 types and 0.35 types at Time 1, Time 2 and Time 3 respectively in the written corpus whereas 0.62 types, 0.63 types and 0.31 types at Time 1, Time 2 and Time 3 respectively in the spoken corpus.

It is also interesting to highlight that the phrase, *happiest* + *day* occurred more commonly in the written corpus in comparison to the spoken corpus. Conversely, the phrase, *happiest* + *moment* occurred more frequently in the spoken corpus instead. This could be an indication that these students tend to be more specific when they share their life experiences verbally than through their writing. As a result, they preferred using *moment*, a relatively specific choice of word more frequently as opposed to *day*, a quite general word in their spoken language.

In the following section, the findings on the structures of the bundles are presented.

4.3 Research question 2:

2. What are the structures and functions of the four-word LBs and to what extent do the LBs found in the written narrative texts differ from those found in the spoken narrative texts?

4.3.1 The structural analysis of lexical bundles

The structural categories of four-word LBs found in the written and spoken corpora were identified and presented in Chapter 3. The distribution of structural categories of four-word LBs in the written and spoken corpora over time is presented in Table 4.6 below (refer to Chapter 3, Table 3.3 for the modified structural framework). The findings on the structures of bundles used in the written and spoken corpora showed that most of the bundles in both corpora consisted of VP fragments all making up to more than 40% of the bundles in each corpus. The proportions of VP-based bundles were comparatively larger in the spoken corpus than in the written corpus at all three points in time.

Table 4.6: Distribution of structural categories of four-word lexical bundles in the written and spoken corpora over time

Structure	Examples	Written corpus			Spoken corpus		
		Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
		% of all structures			% of all structures		
VP-based							
(connector +) 1st/3rd person pronoun + VP fragment	I went to the, we arrived at, she went to the	23.9	22.1	20.39	30.09	31.41	30.21
(connector +) pronoun/ Noun phrase + VP fragment	because this is my, family and I went, a big smile plastered	6.39	7.27	4.81	3.98	3.33	4.17
Copula be + Noun phrase/Adjective phrase	was the happiest day, are a lot of, am very happy because	6.67	6.47	6.07	4.28	3.16	4.17
(connector +) VP fragment	and go to the, packed all the thing, cooked my favourite food	4.22	7.74	6.07	5.60	7.82	6.12
Preposition phrase + VP fragment	in my life was, by car and arrive, of my life was	-	0.23	-	0.74	0.33	0.78
Adjective phrase (with VP fragment)	I was so nervous, we are so happy, excited because this is	3.95	1.5	3.67	1.92	1	1.04
		45.13	**45.31**	**41.01**	**46.61**	**47.05**	**46.49**
DC-based							
(connector +) 1st/3rd person pronoun + dependent clause fragment	and I go to, he want to go, I was going to	3.95	3.46	3.89	3.69	5.16	6.64
WH-clause fragments	audience were wonder who, when I arrive at, don't know how to	1.78	0.9	0.46	1.47	2	1.56
(connector + (noun/verb/ adjective +)) *to*-clause fragments	a chance to see, and get ready to, very excited to go	3.54	6.47	8.82	3.69	8.15	7.68

Table 4.6, continued

Structure	Examples	Written corpus			Spoken corpus		
		Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
		% of all structures			% of all structures		
(VP +) *That*-clause fragments	told us that we, think that it was, that we go to	0.54	0.35	0.69	0.29	3.16	0.91
		9.9	**11.18**	**13.86**	**9.14**	**18.47**	**16.79**
NP/PP-based							
(connector +) Noun phrase with *of*-phrase fragment	a lot of food, day of my life, the end of the	14.83	12.70	13.4	6.64	3.33	3.91
Noun phrase with other post-modifier fragment	living room with my, happiest day in my, the first day at	1.9	3.81	1.83	12.73	11.15	11.2
Other noun phrase expressions	my family and I, days and one night, last year my family	17.10	15.94	11.91	8.8	9.82	10.29
(connector +) Prepositional phrase expressions	and on the evening, at the same time, in front of the	4.22	5.08	5.04	4.28	4.49	5.73
		38.05	**37.53**	**32.18**	**32.45**	**28.79**	**31.13**
AdjP-based							
(connector +) Adjective phrase expressions	and I so happy, so nervous because I, very fast and I	**1.21**	**1.85**	**3.44**	**2.21**	**1**	**1.3**
AdvP-based							
Adverb phrase expressions	other than that we, back to the hotel, and after that we	**0.41**	**0.69**	**2.06**	**2.95**	**1.53**	**1.3**
Others							
Intersentential bundles	(with my family. We), (are very happy. After), (me. It was a)	**5.3**	**3.44**	**7.45**	**6.64**	**3.16**	**2.99**
Total		**100**	**100**	**100**	**100**	**100**	**100**

The bundles in both the corpora were dominated by the structural type, '1st/3rd person pronoun + VP fragment'. Bundles that begin with '1st/3rd person pronoun + VP fragment' are provided below with examples from the written and spoken narrative texts respectively:

(7) After bathing, *I went to the* kitchen for breakfast. (W35a)
After that, we went to the Desaru Resort. (W22b)
Suddenly, *he gave me a* big box to me. (W12c)

It was the happiest day in my life because *I can spend time* with my family… (S11a)
We arrive at the destination at 9.00 a.m. (S40b)
It was the happiest moment for me and my family because the activities there was so enjoyable. (S06c)

Bundles with 1st person pronouns, *I* and *we* were more common than bundles 3rd person pronouns, *he*, *she*, *it* and *they* in this structural type. The students made substantial use of bundles with 1st person pronouns, *I* and *we* most probably to create a personal account of their written and spoken narratives.

More than half of the bundles in the written and spoken corpora were made up of VP-based and DC-based bundles. VP-based and DC-based bundles are said to be clausal elements as they consist of a verb component (Biber et al., 2004). The proportions of clausal bundles (i.e., VP-based and DC-based bundles) ranged at 55% and 65% of the bundles in the written and spoken corpora respectively. This suggests that both the written and spoken narrative texts are mostly clausal in nature. Second, the NP/PP-based bundles (i.e., phrasal bundles) too were found to be common in the written and spoken corpora but not more than VP-based bundles which are presented below with examples from the written and spoken narrative texts respectively:

(8) On the evening, we take *a walk at the* beach together and enjoy the fresh air. (W13a)
The PT3 results coming out, It was *the happiest day of* my life because I got to make parent happy and I happy because I got 7A's in my PT3 results. (W41b)
That morning at 6. a.m we get ready and gathered together *in front of the* hotel. (W17c)

Last year *my family and I* go to Pulau Tioman. (S36a)

I thank to all the teachers and my friends that have teach me so I get 5A in UPSR and it was the *happiest day in my* life. (S16b)

At the night we had a dinner in a famous restaurant at Cameron. (S12c)

NP/PP-based bundles ranged at 38% and 32% of the bundles in the written and spoken corpora respectively. Phrasal bundles were more commonly used in the written narrative texts than in the spoken narrative texts. It can be said that the written corpus is slightly more phrasal than the spoken corpus. Third, DC-based bundles ranged about 13% and 18% of the bundles in the written and spoken corpora respectively. The use of DC-based bundles in the written corpus increased over time. An increase in the use of DC-based bundles was also noted in the spoken corpus from Time 1 to Time 2 which slightly decreased at Time 3. The use of '*to*-clause fragments', a subcategory of DC-based bundles increased over time in the written corpus. Notably, the bundle, *to go to the* in this structure was found to be more common than other '*to*-clause fragments' in both corpora. Interestingly, in the written corpus only two instances of the bundle, *to go to the* were noted at Time 1 (9) which drastically increased to nine instances at Time 2 (10) and Time 3 (11):

(9) Next, our plan was *to go to the* green tea farm. (W12a)

First Day, we want *to go to the* beach. (W19a)

(10) Then, my friends asked me *to go to the* stage, after I walked along the hall, just in asudden, my mom appeared and hug me tightly. (W08b)

It was a sunny day, Zamri and I take a decision *to go to the* vacation at Hong Kong and Rio De Jeneiro. (W21b)

It wrote that my friend want to invite me *to go to the* vacation at Melaka Waterpark. (W27b)

(11) After we spent about an hour there, we decided *to go to the* green tea farm too. (W12c)
Every step I take *to go to the* stage, I feel so light. (W16c)
It was a sunny day, my friend names Hazim called me, he want to invite me join him *to go to the* Tanjung Balau Beach. (W27c)

In the spoken corpus, there were six instances of *to go to the* at Time 1 (12) which decreased by one to five instances at Time 2 (13) and drastically increased to 12 instances at Time 3 (14):

(12) Their make surprise *to go to the* my mom's hometown at the Kelantan. (S04a)
We are very excited because that was our first time *to go to the* very interesting place. (S06a)
After that my family bring me *to go to the* beach to spend time. (S11a)

(13) When all finish my father said to me *to go to the* river and play what you want he said. (S01b)
The teacher called the 5A students *to go to the* stage one by one and I waited for my name to be called. (S08b)
We standby *to go to the* airport to buy a flight ticket to Rio de Janeiro. (S21b)

(14) In the night we need to prepare and get ready *to go to the* SMK Temin Baru school in Pahang. (S17c)
They said they want bring my family, my cousin and my siblings *to go to the* most excited, most interesting, most fun at the world theme park USS. (S20c)
Friday my sister come home and she invite me *to go to the*
AEON Jusco. (S27c)

The '*to*-clause fragments' found in both corpora were mainly used to elaborate the main content in a sentence wherein the writer or speaker answered the questions 'where', 'why' or 'what' using bundles with this structure. Based on the findings on the use of the '*to*-clause fragment', *to go to the* illustrated above, there are no linear patterns of development noted in the use of *to go to the* in written and spoken language of

students over the span of six month. In the written corpus, initially, the students made use of *to go to the* in two simple sentence which expanded to complex sentences at Time 2 and Time 3. However, in the spoken corpus, the use of complex sentences was noted even from the initial production that expanded to more instances of complex sentences at the final production. Apart from that, AdjP-based bundles, AdvP-based bundles and intersentential bundles were all used at low proportions in both corpora.

Overall, the findings on the distribution of structural categories of four-word LBs in the written and spoken corpora over time show an inconsistent pattern. The findings on the VP-based and NP/PP-based bundles are in line with Biber et al. (2004) to a certain extent. Biber et al. (2004) found that the spoken registers (i.e., conversation and classroom teaching) consisted of a greater number of VP-based bundles whereas the written registers (i.e., academic prose and textbooks) consisted of a greater number of NP/PP-based bundles. Similarly, in this study, although VP-based bundles are most widely used in both the corpora, the spoken corpus still comprises a slightly larger proportion of VP-based bundles than the written corpus. NP/PP-based bundles, on the other hand, are slightly larger in the written corpus than the spoken corpus although not more than VP-based bundles. The findings on DC-based bundles too correspond to the findings of Biber et al. (2004) where they found that DC-based bundles were used in a larger quantity in the spoken registers than the written registers. Similar to that, this study reveals that DC-based bundles are more common in the spoken corpus in comparison to the written corpus. This finding is in contradiction to the past claim of O'Donnell (1974) which states that DC are significantly greater in writing rather than in speech (as cited in Akinnaso, 1982, p. 107). Ruan (2016) found the use of large proportions of VP-based bundles in the academic writing of Chinese students collected at four points during their four years of studies. He pointed out that these students tend to acquire VP-based and NP-based bundles before acquiring PP-based bundles. This

pattern is said to be a developmental order of LBs in the academic writing of the students (Ruan, 2016). In line to that, the findings on the higher proportions of VP-based bundles and NP-based bundles in this study too may indicate a developmental order of LBs. However, the validity of the developmental order of bundles can only be verified through an extended period of study on LBs in the written and spoken language of these students.

Although the bundles used in the written and spoken corpora are dominated by VP fragments (e.g., 45% and 47% respectively), NP/PP-based bundles too are used in considerably high proportions (e.g., 38% and 32% respectively). In spite of the differences in the modes of production, the written and spoken narrative texts contain a combination of bundles that are typically used in writing and speech. This could be due to the task type which required students to narrate their life experiences which resulted in a heavy reliance on bundles consisting of personal expressions, '1st person pronoun + VP fragments' in their written and spoken narrative texts as well as bundles that are more phrasal in nature even in their spoken narrative texts.

In the next section, the findings on the functions of the LBs in the written and spoken corpora are discussed.

4.3.2 The functional analysis of lexical bundles

The functions of four-word LBs found in the written and spoken corpora were identified and presented in the earlier chapter. The distributions of functional categories of four-word LBs in the written and spoken corpora are presented in Table 4.7 below (refer to Chapter 3, Table 3.4 for the modified functional framework). The complete lists of LBs according to the functional categories in the written and spoken corpora over time are provided in appendices C and D respectively due to space limitation. As can be seen in Table 4.7 below, more than half of the bundles in both corpora

functioned as referential expressions making up to over 50% of the bundles in the written and spoken corpora. The referential expressions that were used for 'identification or focus' in order to highlight on a significant event, animate, inanimate or abstract entity in the writer, speaker or character(s)'s life accounted for the highest proportions in both corpora, yet, comparatively larger in the written corpus than in the spoken corpus (e.g., 44% and 36% respectively). Examples of referential bundles with 'identification or focus' function in the written and spoken corpora are provided below respectively:

(15) I swear I will remember till my last breath. It was the *happiest day of my* life. (W07a)
My cousin and I very excited to go there. (W31b)
The Place that we went is Legoland. (W39b)

This is *happiest moment in my* life. (S26a)
Actually *my mother and my* brother and I was so happy because we want to get a surprise. (S26c)
I have best experience in my life is my friend and I go to *other country such as* Brazil and Japan. (S21c)

Referential expressions that functioned as 'time reference' were used in a greater proportion in the spoken corpus at all three points in time in contrast to the written corpus (e.g., 11% and 6% respectively). This indicates that students were more focused and precise in stating the time setting in the narration through speech during a face-to-face session in comparison to the narration through writing.

Table 4.7: Distribution of functional categories of four-word lexical bundles in the written and spoken corpora over time

Function	Example	Written corpus			Spoken corpus		
		Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
		% of all functions			% of all functions		
Stance expressions							
Epistemic stance	I think it was, I am very sure, I hope we can	1.87	0.24	1.49	-	2.41	1.07
Attitudinal/ modality stance							
Desire	because I want to, want to go to, and he want to	1.58	1.08	1.98	1.74	4.81	2.68
Obligation/ directive	have to go to, brother asked me to, because we have to	-	0.12	1.86	0.47	2.40	2.82
Intention/ Prediction	can spend time with, I can't wait to, SMK Sungai Tiram will	2.44	2.88	1.73	2.07	3.43	2.15
		5.89	**4.32**	**7.06**	**4.28**	**13.05**	**8.72**
Discourse organizers							
Transition	after that my family, other than that we, first and foremost we	**3.16**	**2.76**	**3.96**	**10.11**	**6.01**	**5.91**
Referential expressions							
Identification/ Focus	happiest day in my, my family and I, fresh air at the	44.54	41.23	37.75	36.49	28.35	33.56
Place reference	go back to hotel, go to the beach, in the living room	3.16	3.49	4.7	2.37	5.33	3.09
Time reference	at the same time, in the morning I, at 8.00 a.m. I	4.89	6.37	5.32	9	11	10.74
Quantity specification	buy some things for, a lot of food, some food to eat	3.3	2.89	1.24	4.58	1.55	3.35
		55.89	**53.98**	**49.01**	**52.44**	**46.23**	**50.74**

Table 4.7, continued

Function	Example	Written corpus			Spoken corpus		
		Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
		% of all functions			% of all functions		
Topic-oriented expressions							
Depiction of action/state	we go to the, my friends and I were, I saw my mother	25.72	28	26.98	24.64	26.80	24.96
Depiction of feelings/emotions	I was very excited, I am so happy, so nervous because I	6.32	4.08	8.29	4.42	2.06	2.82
Elaboration/ Clarification	it is because my, to go to my, even though we were	3.02	6.02	4.08	3.16	3.44	4.03
		35.06	**38.10**	**39.35**	**32.22**	**32.30**	**31.81**
Special conversational functions	he told me that, I said to my, my father said to	**-**	**0.84**	**0.62**	**0.95**	**2.41**	**2.82**
Total		**100**	**100**	**100**	**100**	**100**	**100**

Topic-oriented expressions ranged about 39% and 32% in both corpora respectively. Examples of topic-oriented bundles found in the written and spoken corpora are presented below respectively:

(16) Last year my family and I went to Langkawi. *I was very excited* to go for Langkawi because I never go Langkawi before. (W39a)
At the night, *we take a walk* around the village. (W13b)
After that, we continue our journey *to go to the* Deerland Lanchang, Pahang. (W37c)

Then before we go home *we go to the* restaurant first to lunch. (S12a)
We were very excited but the situation is crowded because there a lot of people especially Chinese. (S37b)
And after arrive I *help my father to* prepare the food. (S25c)

Topic-oriented expressions that functioned to 'depict an action or state' accounted for the highest percentages in the written and spoken corpora with minimal differences (e.g., 28% and 27% respectively). It was found that the most commonly used 'depiction of action/state' bundles in the written narrative texts are *and I go to* and *and I went to* at Time 1, *and I go to* at Time 2 and *we go to the* at Time 3. Although the bundles, *and I go to* and *and I went to* appear to have one meaning at the surface level, deeper analysis into concordance lines showed that they functioned to denote two meanings. To illustrate this, concordance lines are divided into two sets, each set indicating one meaning. The examples given here are of the most frequently occurring 'depiction of action or state' bundles, *and I go to* and *and I went to* in the written narrative texts at Time 1. These bundles are used to denote two distinct meanings, shown here in two sets of concordance lines.

Set 1

Last year, my family **and I go to** beach. It is because my aunty
game again and my group win. My cousin **and I go to** the nap for eat some food because
Last year, my family **and I go to** the beach at Tanjung Balau.
the road. At the night of the day. My family **and I go to** the Jonker-Walk. Jonker-Walk is

After finish the school, my family **and I went to** the beach at 2.00 p.m. After arrive
Last years, my family **and I went to** Singapore. Because my mother
Last year, my teacher **and I went to** the art festival. My teacher told me
water based activities such as slides. My brother **and I went to** the tallest water slide in the water
that happen in my life like when My family **and I went to** the Desaru resort in Melaca to
not back yet. After that, my cousin, my sister **and I went to** the night market. After back from
Last year, my family **and I went to** Tioman Island. We was very
Last year my family **and I went to** Langkawi. I was very excited to go

Here the bundles *and I go to* and *and I went to* are used to mean 'the characters' act of going to a place' whereby, 'to' functions as a preposition followed by nouns. The nouns that appear in these lines are: *beach, nap, Jonker-Walk, Singapore, art festival, water slide, Desaru resort, night market, Tioman Island* and *Langkawi.* The nouns all indicate a place reference.

Set 2

Go vocation Last year, my family	**and I go to** vacation at Batu Layar. We go at
In The Public holiday, My family	**and I go to** visited Tanjung Leman Beach. I
food. After I finish lunch, my father	**and I go to** do next activity like catching fish.
hotel to sleep. At the morning, My dad	**and I go to** jogging and have a breakfest. My
brother and I playing kite. My litle brother	**and I go to** find the some beautiful shell and I
and I go to find the some beautiful shell	**and I go to** ate some food because hungry.
that, my family and I cleaning the place	**and I go to** the last swimming. My family and
we check in the challet. Firstly, my family	**and I go to** snorkelling by boat. When we

friends at there. When afternoon my friends	**and I went to** play football. After play football
and another things. Before swimming, my family	**and I went to** ate some food and we swimming.
Last week, my family	**and I went to** go holiday at Malacca, Bandar
Two years ago my family	**and I went to** the holiday at Desaru. I was very

Here the similar bundles are used to mean 'the characters' act of going in order to carry out an action'. In these examples, 'to' occurs with verbs that signify activities (i.e., *vacation, visited, do, jogging, find, ate, swimming snorkelling, play, go* and *holiday*). In other words, *and I go to* and *and I went to* are not only used to depict the act of going to a place as evident in set 1 but they are also used to portray the act of going in order to perform an activity as seen in set 2.

Conversely, the most frequently occurring bundle *we go to the* in the written narrative texts at Time 3 in this function has one meaning as it is used to depict 'the characters act of going to a place' as illustrated in set 3 of concordance lines provided below:

Set 3

do holiday at the Tioman beach.	**We go to the** beach by car. Before we go
beach. We go to the beach by car. Before	**we go to the** beach, my sister and I helped
swim in the blue ocean. Then, an 3 p.m.	**we go to the** restaurant in the hotel to eat.
We finish visit at 12.30 p.m. After that,	**we go to the** Masjid Sultan Abu Bakar
fathers just sit on the mat. After finish ate,	**we go to the** water and played the ball. We
to drawing my face as free. After that,	**we go to the** vase making factory. My friends
have a energy we start our activities. First,	**we go to the** fun-fair. We took 15 minutes to
activities. The second day for our travel,	**we go to the** Angkor Wat. There have a many
too in the Melaka. Furthermore, the last day	**we go to the** town at the Melaka and buy

It was also found that students made use of the bundle, *we go to the* in the spoken narrative texts at all three points in time for the same function. At time 3, however, there were two bundle types that were most commonly used to 'depict an action or state': *that we go to* and *we go to the*. On the surface, these two bundles appear to be overlaps. But in fact they are two distinct bundles as presented in the examples below:

(17) After that we go back to the jetty and we have some prayers before we going back to the hotel. After *that we go to* the SMK Temin and ask them to prepare us good dinner for us. (S37c)

(18) *We go to the* hometown and get to see a beautiful scenery but cannot get to take some picture because my phone broken that time. (S04c)

The former (17) incorporates a part of a connecter and has two meanings, similar to the written bundles *and I go to* and *and I went to*. The latter (18) has only one meaning as it only refers to the characters' act of going to a place similar to the written bundle, *we go to the* shown in set 3.

One important observation is made through the concordance analysis of 'depiction of action or state' bundles with dual meanings (e.g., *and I go to*, *and I went to, that we go to*). Although these bundles seem to be structurally and functionally similar on the surface, they are not. The two different meanings of these bundles are represented by two different structural patterns as it is said that meanings of words can be distinguished by observing the patterns or phrases they typically occur with (Hunston, 2002). For instance, *and I go to* which is used to mean 'the characters' act of going to a place' (e.g., *and I go to beach*) represents the structural pattern '(connector +) 1st person pronoun + VP fragment' whereas the same bundle which means 'the characters' act of going in order to perform an activity' (e.g., *and I go to visited Tanjung Leman Beach*) is of the pattern '(connecter +) 1st person pronoun + DC fragment'. Moreover, intriguingly, these

dual meanings are typically found in bundles with the following pattern '*I/we* followed by *Verb* followed by *to'*. The verbs that frequently occur in this pattern in the written and spoken narrative texts are: *go* and *went*.

Another notable finding is based on the functions of the most commonly occurring bundles in the written and spoken corpora. As highlighted earlier, the most commonly occurring bundles in the written corpus are *my family and I* at Time 1 (44 times) and Time 2 (45 times), *happiest day of my* and *was the happiest day* at Time 3 (38 times). The bundle, *my family and I* functions to make reference to a group of people involved in an activity (19) whereas the bundles, *happiest day of my* and *was the happiest day* function to make reference to a significant event (20) in the writer's life:

(19) Last year, *my family and I* go to vacation at Batu Layar. (W01a)
After finish the school, *my family and I* went to the beach at 2.00 p.m. (W02a)
Last year, *my family and I* went to holiday at Malacca for three days and two nights during end-year holiday's. (W06b)

(20) Even we not got straight A's but we got about 7 A's. We do thanked to all teacher who teach us and it was the *happiest day of my* life. (W07c)
After a few minutes talking, teacher got to say one by one name of the excellent students who get straight A's. It was really unbelieveable when the teacher call my name after done the others. […] So yes, I really think that it *was the happiest day* of my life. (W08c)
Last school holiday, my eldest brother, Muhammas Zahrin had decided to take a vacation for us. This is because, a big company from Kota Kinabalu, Sabah had given a letter to him and told him that they will took him for working with them. Eventhough the place is not as far as we think, but I'm sure we will miss him as soon he leaves us. […] For me, he is a best brother ever and it *was the happiest day* of my life. (W12c)

On the other hand, as noted above, the most commonly used bundles in the spoken corpus are *after that we go* and *we go to the* at Time 1 (23 times), *day in my life* and *we*

go to the at Time 2 (17 times), and *happiest day in my* (22 times) at Time 3. In the spoken corpus, at Time 1, *after that we go* functions as a discourse organizer that is used to show transition of ideas (21) whereas *we go to the* is a topic-oriented expression that functions to depict the act of going somewhere (22):

(21) *After that we go* to Safari Park we have to use a track to look around the area. We see a lot of animal that near us like lion and tiger. We take some picture for memories. *After that we go* for lunch at the A'Famosa Resort. (S06a)
We first we don't take any ride we just walk around the USS. So *after that we go* to first game we ride is Transformers 4D. (S20a)
At the place we buy ticket for one big family. *After that we go* to Salang by ferry. (S36a)

(22) *We go to the* beach because we want celebrate my grandfather birthday. (S25a)
Then *we go to the* water cruise which is there a lot of ferry and go around the Melaka. (S37a)
Then *we go to the* butterfly farm. (S42a)

But in the spoken narrative texts, a shift is noted from a high reliance on a discourse organizing bundle (i.e., *after that we go*) to a high reliance on referential bundles, *day in my life* at Time 2 (17 times) and *happiest day in my* at Time 3 (22 times) which function to make reference to a significant event in the speaker's life (23):

(23) Last week is my birthday. So I'm very happy because I already official 16. So I wish for making my day so happy. […] I very excited then I cry because I not expectedly that they make me happy then one of my best friend give me a gift and my parents too. After that I said thank you to all my friends and family and my parents. So it was my happiest *day in my life.* (S05b)
Everyone have happy day in their life. I also have happy *day in my life.* It was Legoland trip that prepared by my school to Legoland. (S39b)
My *happiest day in my* life is when is during my UPSR result giving ceremony. (S16c)

The discourse organizer, *after that we go* is still used at a high frequency at Time 2 (14 times) and Time 3 (17 times) in the spoken narrative texts but not as common as the referential bundles, *day in my life* and *happiest day in my*. However, *after that we go* occurred at a low frequency in the written corpus. The heavy use of *after that we go* in the spoken corpus points to the demand of speech where the students are left with a relatively short period of time to organize and process the content of the narrative. As a result, they have made high use of the discourse organizer, *after that we go* to probably ensure the flow and transition of ideas in their spoken language. This is not the case for the production of the written narratives as they have time to plan and organize their narrative which could be why the students did not make heavy use of *after that we go* in the written corpus as they did in the spoken corpus. Moreover, the shift from the high use of discourse organizing bundle, *after that we go* at Time 1 to referential expressions, *day in my life* and *happiest day in my* at Time 2 and Time 3 respectively in their spoken narrative texts suggest that a change is taking place in the spoken language of these students over time which is becoming more like the written language of the students, ultimately to accommodate to the needs of the narrative genre.

Stance expressions and discourse organizers were found to be less common than the other functional categories in both corpora. Nevertheless, discourse organizing bundles were slightly more common in the spoken corpus than in the written corpus at all three points in time. For instance, the proportions of stance expressions were about 7% in the written corpus and 13% in the spoken corpus. Discourse organizers accounted for about 3% and 10% in the written and spoken corpora respectively. Noteworthy is the fact that epistemic stance bundles (e.g., *I am very sure*) were common in the written narrative texts than in the spoken narrative texts at Time 1 and Time 3. More interestingly, no epistemic stance bundles were found in the spoken corpus at Time 1. Bundles with

special conversational functions (e.g., *my father said we*) were not only found in the spoken corpus but were also present in the written corpus.

Overall, the findings on the distribution of functional categories of four-word LBs in the written and spoken corpora over time show an inconsistent pattern. The substantial use of referential expressions and minimal use of stance and discourse bundles in the written and spoken narrative texts, despite the difference in the modes of production may indicate the possible requirement of the narrative genre that is descriptive in nature. This is made evident through the fact that the functions of bundles in the corpora are rather distinctive in comparison to Biber et al.'s (2004) functions. The narration of life experiences through writing and speech is clearly a different genre from the written and spoken registers such as textbook, academic prose, conversation and classroom teaching. Hence, this requires for the use of a different repertoire of linguistic features. This is possibly why the written and spoken narrative texts are heavily reliant on 'literate' bundles (i.e., referential expressions) and less reliant on 'oral' bundles (i.e., stance expressions and discourse organizers).

The considerable use of topic-oriented expressions in both corpora can be seen as representing the characteristics of the narrative genre as well. This is made evident in the extensive use of topic-oriented bundles which are mainly clausal components consisting of VP-based bundles and DC-based bundles. Moreover, most of the topic-oriented bundles begin with '1st person pronoun + VP fragments'. The stance expressions and discourse organizers used in both the corpora constitute VP-based bundles beginning with '1st person pronoun + VP fragments' but are lesser in proportion in comparison to topic-oriented expressions. This is in contradiction to Biber et al.'s (2004) study which found that stance expressions and discourse organizers were dominated by VP-based bundles and DC-based bundles. From the observation, it can be said that VP-based bundles may appear to be used for similar functions when they

display the same structural patterns, but in fact they have different functional use in the corpora. This is obvious in the case of '1st person pronoun + VP fragments' that largely functioned as topic-oriented expressions (e.g., *we go to the*), followed by, stance expressions (e.g., *and I think it*) and discourse organizers (e.g., *after that we go*) that were used in lesser proportions in the written and spoken narrative texts.

Despite the differences in the corpora and the functional use of bundles, the findings of this study on the functions of LBs correspond to Biber et al.'s (2004) findings on the functions of bundles in the written registers (i.e., textbooks and academic prose) to a certain degree displaying a high use of referential expressions and a low use of stance expressions and discourse organizers. Biber et al. (1999, 2004) also found that academic prose and textbooks (i.e., written registers) consisted of large use of referential bundles and conversation (i.e., spoken register) consisted of a large use of stance bundles and discourse organizers. On the other hand, classroom teaching (i.e., spoken register) constituted a high use of both referential expressions and stance bundles. Likewise, in this study, despite the fact that referential bundles are largely used in both corpora, the written narrative texts consisted of a larger use of referential bundles than the spoken narrative texts. On the contrary, stance expressions and discourse organizers were less common in both corpora. But stance expressions were slightly more in the spoken corpus than in the written corpus at Time 2 and Time 3. Discourse organizers were slightly more common in the spoken corpus than in the written corpus at all three points in time.

Another striking observation on the functions of bundles used in written and spoken narrative texts is that they do not demonstrate a clear written or spoken production specific set of bundles such as in the case of Biber et al. (2004). To illustrate, unexpectedly the students made low use of stance expressions and high use of referential expressions in the spoken narrative texts. At the same time, the written

narrative texts consisted of stance expressions and bundles with special conversational functions displaying a mixture of 'literate' and 'oral' bundles to some extent in both the written and spoken corpora. Probably, this phenomenon is a clear indication of the characteristics of narrative genre. Therefore, it can be safely concluded that LBs are genre specific in which they primarily function to fulfill the needs of the genre that goes beyond the needs of the modes of production and this is realized in the functional use of bundles in the written and spoken corpora.

4.4 Research question 3:

3. How might the changes in the use of LBs observed over time explain about the nature of language development?

4.4.1 Findings on two different analysis on adjective phrase-based bundles: Error analysis vs. analysis of learner language in its own right

Two different analysis (i.e., EA and analysis of learner language in its own right) were conducted to investigate the use of AdjP-based bundles in the written and spoken corpora over time. These analysis were performed to understand the difference between measuring learner language based on the NS norms and treating learner language in its own right (see Chapter 3: section 3.7.3 for a detailed account of the procedure of analysis). The findings on the frequency of error types produced in the use of AdjP-based bundles in the written and spoken corpora over time are provided in Table 4.8.

Table 4.8: Frequency of error types in the use of adjective phrase-based bundles in the written and spoken corpora over time

	Error category	Written corpus		Spoken corpus	
		Frequency	% of total errors	Frequency	% of total errors
Time 1	1. Omission	0	0	5	100
	2. Addition	0	0	0	0
	3. Misinformation	13	100	0	0
	4. Misordering	0	0	0	0
	5. Blends	0	0	0	0
Time 2	1. Omission	9	100	0	0
	2. Addition	0	0	0	0
	3. Misinformation	0	0	2	100
	4. Misordering	0	0	0	0
	5. Blends	0	0	0	0
Time 3	1. Omission	14	66.7	2	100
	2. Addition	0	0	0	0
	3. Misinformation	7	33.3	0	0
	4. Misordering	0	0	0	0
	5. Blends	0	0	0	0

As can be seen in Table 4.8, most of errors produced by the students involved the omission of copula *was/were* (e.g., *I very happy because*) and the misuse or misinformation of copula *am/are* (e.g., *I am very happy*) in past tense narratives. It is noted that the frequency of errors in the use of AdjP-based bundles increased in the written narrative texts over time. However, the frequency of these errors reduced in the spoken narrative texts over time. The students produced more errors in the use of AdjP-based bundles in their written narrative texts than in their spoken narrative times over time. This suggests that the students may have not completely acquired the correct usage of copula *be* especially in their written language which resulted in the increase of the production of error types in the written corpus. At the same time, the reduction in the frequency of error types in the spoken narrative texts may suggests that these students are improving in the use of AdjP-based bundles in the spoken language over time. The omission of copula *be* can be said to be an interlingual error which is a result of the mother tongue influence of the students (Richards & Sampson, 1973; Ellis & Barkhuzien, 2005). The Malay and Tamil languages are claimed to be the mother

tongues of the participants of the study. These two languages do not operate according to the rule of copula *be* and hence, the rules of the mother tongue could have been transferred to the TL. This is perhaps why the students made use of these incorrect forms in the written and spoken corpora. The possible cause for the misinformation errors to occur is not clear. However, this could also indicate that the students have not completely acquired the correct usage of copula *was/were* to indicate past tense in the narratives.

Following that, the description of errors produced in the use of AdjP-based bundles in the written and spoken corpora are presented in Table 4.9. As highlighted in the previous chapter, Dulay et al.'s (1982) surface structure taxonomy (as cited in Ellis & Barkhuzien, 2005) was used to describe the AdjP-based bundles with errors. The corrected forms of the incorrect forms produced by the students in the written and spoken corpora are provided as reconstructions.

Table 4.9: Error description of adjective phrase-based bundles in the written and spoken corpora over time

		Error	Reconstruction	Surface structure description
Written corpus				
Time 1	1.	…I am very excited… (4)	…I was very excited…	Misinformation
	2.	I am so excited… (3)	I was so excited…	Misinformation
	3.	I am so happy… (3)	I was so happy…	Misinformation
	4.	I am very happy… (3)	I was very happy…	Misinformation
Time 2	1.	…I very happy because… (4)	…I was very happy because…	Omission
	2a.	…I very excited to… (2)	…I was very excited to…	Omission
	2b.	…I very excited to… (1)	…I were very excited to…	Omission
	3a.	…and I very happy… (1)	…and I was very happy…	Omission
	3b.	…and I very happy… (1)	…and I were very happy…	Omission
Time 3	1.	…I am very excited… (4)	…I was very excited…	Misinformation
	2a.	…I very happy because… (3)	…I was very happy…	Omission
	2b.	…I very happy because… (1)	…I were very happy…	Omission
	3.	…we are so happy… (3)	…we were so happy…	Misinformation
	4a.	…and I so happy… (1)	…and I was so happy…	Omission
	4b.	…and I so happy… (1)	…and I were so happy…	Omission
	5a.	…and I very happy… (1)	…and I was very happy…	Omission
	5b.	…and I very happy… (1)	…and I were very happy…	Omission

Table 4.9, continued

		Error	Reconstruction	Surface structure description
	6.	…because I very happy… (2)	…because I was very happy…	Omission
	7a.	…I so happy because… (1)	…I was so happy because…	Omission
	7b.	…I so happy because… (1)	…I were so happy because…	Omission
	8a.	I very excited because… (1)	I was very excited because…	Omission
	8b.	…I very excited because... (1)	…I were very excited because...	Omission
Spoken corpus				
Time 1	1.	…I very happy because… (3)	…I was very happy because…	Omission
	2.	…I very tired because… (2)	…I was very tired because…	Omission
Time 2	1.	…I'm very happy because…(2)	…I was very happy because…	Misinformation
Time 3	1a.	I so happy because… (1)	I was so happy because…	Omission
	1b.	…I so happy because… (1)	…I were so happy because…	Omission

As highlighted above, the students made two types of errors that are omission and misinformation. The first type of error involved the omission of copula *was/were* in AdjP-based bundles (e.g., *I very excited to*). The reconstructions of bundles without copula *be* (e.g., *I very excited to*) were based on the subject-verb agreement rule. To illustrate, Pattern 2a in the written corpus at Time 1 is part of the sentence, *I very excited to see them* (W33b) whereas Pattern 2b is part of the sentence, *My cousin and I very excited to go there* (W31b). Although on the surface these two bundles seemed to represent the bundle type, *I very excited to* close-text observation showed that the former contains a singular subject '*I*' and the latter a plural subject '*we*'. Therefore, the use of copula *be* in the sequence, *I very excited to* would differ based on the subject-very agreement rule. The reconstructions of these two bundle types are: 2a. *I **was** very excited to see them* and 2b. *My cousin and I **were** very excited to go there* as presented in Table 4.9 above. Second type of error involved the misinformation of copula *be* whereby the students made incorrect use of present tense, *am/are* to describe past accounts of events as in, *Last week was my birthday, I am very excited to know that now*

I was finally turns 16 (W05c). The reconstruction of this form is presented as *I was very excited* as indicated in Table 4.9 above.

In contrast to the EA, when learner language is treated as an independent system, the production of students is not measured to NS system but to his or her initial production which is taken as a baseline. As noted in the previous chapter, the term 'conventional forms' is used to refer bundles with structures that are acknowledged by traditional grammar whereas the term 'innovative forms' is used instead of 'errors' to refer to bundles with structures that do not fit into conventionally acknowledged grammatical items. The conventional forms and innovative forms of AdjP-based bundles identified in the written and spoken corpora over time are provided in Table 4.10.

Table 4.10: Conventional and innovative forms of adjective phrase-based bundles in the written and spoken corpora over time

		Conventional forms	Innovative forms
Written corpus	Time 1	I was very excited (5) we were so excited (2)	I am very excited (4) I am so excited (3) I am so happy (3) I am very happy (3)
	Time 2	I was so nervous (2) I was very excited (2) I was very happy (2)	I very happy because (4) I very excited to (3) and I very happy (2)
	Time 3	I was very excited (5) I was so happy (4) I was so excited (2) I was very happy (2)	I am very excited (4) I very happy because (4) we are so happy (3) and I so happy (2) and I very happy (2) because I very happy (2) I so happy because (2) I very excited because (2)
Spoken corpus	Time 1	I was so happy (3) I was so shocked (2) I was very nervous. (2)	I very happy because (3) I very tired because (2)
	Time 2	we were very excited (2)	I'm very happy because (2)
	Time 3	I was so excited (3) I was so happy (3) I was so happy. (2)	I so happy because (2)

As evidenced in the table above, the frequency of innovative forms of AdjP-based bundles was slightly more than the frequency of conventional forms in the written corpus at all three points in time; seven conventional forms and 13 innovative forms at Time 1, six conventional forms and nine innovative forms at Time 2 and 13 conventional forms and 21 innovative forms at Time 3. This shows that the students are becoming increasingly innovative in the use of AdjP-based bundles in their written narrative texts over time. A contradiction is noted in the spoken corpus where the frequency of conventional forms and innovative forms produced fluctuate over time. However, the frequency of conventional forms was slightly more than the frequency of innovative forms at Time 1 and Time 3 in the spoken corpus. For instance, seven conventional forms and five innovative forms were identified at Time 1 and eight conventional forms and only two innovative forms at Time 3. Thus, it can be said that the use of AdjP-based bundles in the spoken narrative texts is observed to be getting conventionalized over time. Based on the observation, it can be said that that these students have adopted a more conventionalized way of expressing their emotions in their spoken language. However, they have attempted to express their emotions in a more innovative way in their writing instead. This is also an indication that the students have become more cultured into using native-like forms in their spoken language than in their written language.

Some of the important insights drawn from the two different analysis conducted on the use of AdjP-based bundles are discussed. From the EA perspective, the AdjP-based bundle, *I was very happy* is identified as the correct form in the written and spoken narrative texts. The forms, *I am very happy* and *I very happy because* are identified as incorrect forms due to the misuse of copula *be* and the omission of copula *be* respectively. These two incorrect forms produced by the students are considered as errors because they deviate from the concept of traditional grammar following the NS

norm. On the other hand, based on the analysis of learner language in its own right as shown above, the AdjP-based bundles that do not fit into the concepts of traditional grammar are not rejected as incorrect forms or 'errors' but are treated as innovative forms. Placing side by side the findings on AdjP-based bundles from the EA perspective vs. the findings obtained from treating learner language in its own right, the former is one-sided as it only focuses on what the students did wrong by accepting the grammatical constructs based on the NS norms as the end-state in the language acquisition process. The EA approach also puts a value judgement on learner language where the students are expected to produce native-like language. This perspective also portrays these students as inadequate beings in acquiring the language. The latter, however, presents a much balanced evaluation on the forms that the students have acquired and how they have acquired them differently from the conventional grammatical structures instead of rejecting them as 'errors'. Therefore, there is no value judgement placed on the students as well as on their language production. The students are also not portrayed as deficient beings who have failed to acquire correct forms in their language production but as individuals who are capable of producing forms to meet and suit their language needs.

4.4.2 The nature of language development

The quantitative analysis revealed that a decreasing pattern was noted in the use of LB types in the written corpus at Time 1 to Time 2 which stagnated at Time 3 whereas a fluctuating pattern was noted in the use of LB types in the spoken corpus over time. The overall frequencies of LBs used in both corpora suggested that these students seemed to become less reliant on LBs over the six months of observation. Moreover, minimal differences were noted in the use of LB types as well as in the use of overall frequency of LBs from one point to another in the written and spoken corpora as well as between

the written and spoken corpora. This could be an indication that a six months period of observation of language development may not be enough to draw conclusions about the nature of language development. Furthermore, the findings on the top 50 most frequent LBs in both corpora showed that the students did not rely much on the same set of bundles over time as less than half of the top 50 most frequent bundles were shared by these students in their written and spoken narrative texts at all three points in time. For instance, some bundles occurred at Time 1 and Time 2 but disappeared at Time 3, others emerged at Times 2 and 3 without being evident in the initial productions and some occurred at Time 1, disappeared at Time 2 but recurred at Time 3 in the written and spoken narrative texts. This shows the change that is taking place in the use of LB types in their written and spoken narrative texts throughout the six months of observation.

Although quantitatively, there were not many changes in the use of LBs in the written and spoken corpora over time, qualitatively, there are quite a lot of changes noted in the use of LBs in both corpora which is illustrated using examples of referential bundles used in the written and spoken corpora over time. Table 4.11 presents the referential bundles with 'identification or focus' function that were used to refer to a group of people in the written and spoken corpora over time.

Table 4.11: Referential bundles functioning 'to refer to a group of people' in the written and spoken corpora over time

	Time 1	Time 2	Time 3
Written corpus	1. my family and I (44) 2. my brother and I (8) 3. my friends and I (5) 4. my father and mother (3) 5. aunty and my cousin (2) 6. **I and my sister** (2) 7. my aunty and my (2) 8. my brothers and I (2) 9. my family and they (2) 10. my father and my (2) 11. my mother and my (2)	1. my family and I (45) 2. my friend and I (9) 3. my sister and I (8) 4. my cousin and I (5) 5. my brother and I (4) 6. mother and my (4) 7. mother and my sister (3) 8. my brothers and I (3) 9. my friends and I (3) 10. my friends and my (3) 11. me and my family (2)	1. my family and I (20) 2. my friends and I (17) 3. my brother and I (9) 4. my father and I (5) 5. my mother and my (5) 6. my sister and I (5) 7. my brothers and I (4) 8. my father and mother (3) 9. my friend and I (3) 10. mother and my father (2) 11. my brother and my (2) 12. my cousin and I (2)
Spoken corpus	1. my family and I (19) 2. my mother and my (6) 3. **I and my brother** (3) 4. **I and my cousin** (2) 5. **I and my friend** (2)	1. my family and I (13) 2. my friend and I (4) 3. my mother and my (3) 4. and my young brother (2) 5. **I and my family** (2) 6. my brother and I (2) 7. my family and my (2) 8. my father and mother (2)	1. my family and I (11) 2. my brother and I (9) 3. my mother and my (5) 4. my sister and I (4) 5. my friend and I (3) 6. **I and my family** (2) 7. **me and my family** (2) 8. my aunty and my (2) 9. my cousins and I (2) 10. my father and mother (2) 11. my mom and I (2) 12. my mother and I (2) 13. so my father and (2)

As can be seen, most of the referential bundle types in this function recurred in the written corpus over time. The students also made substantial use of these referential expressions that functioned to refer to a group of people in the written corpus from their initial production to the final production. But this is not witnessed in the spoken corpus. In the spoken corpus, the students made use of few referential bundles in this function at Time 1 which expanded at Time 2 and Time 3. Initially, at Time 1, they made heavy use of the bundle, *my family and I* for 'identification or focus' which could generally refer to any family member. But they became more specific over time which is evidenced in the use of referential bundle types that referred to particular members of the family: *my young brother*, *my brother and I* and *my father and mother* at Time 2 and *my sister and I*, *my aunty and my*, *my cousins and I*, *my mom and I*, *my mother and I* and *so my father*

and at Time 3. Based on this finding, it can be inferred that the students display specificity in their written language from their initial production till their final production. However, they show a developing pattern towards specificity in the use of referential bundle types functioning to refer to a group of people in their spoken language over time. Therefore, it can be said that increasing in specificity in the language can be an indication of developing in the language.

Moreover, as shown in Table 4.11, a few innovative forms in this function are noted in both corpora which are *I and my sister* at Time 1 in the written corpus which disappeared over time, *I and my brother*, *I and my cousin* and *I and my friend* at Time 1 in the spoken corpus which disappeared over time, *I and my family* at Time 2 as well as *I and my family* and *me and my family* at Time 3 in the spoken corpus. Interestingly, the bundle, *I and my family* found in the spoken corpus at Time 2 and Time 3 co-occurred with the conventional form *my family and I*. More interestingly, it was found that two students made use of both innovative forms and conventional forms in their spoken narrative texts. For instance, student 19 made use of the conventional form *my family and I* (3 times) together with the innovative form *I and my family* (1 time) in his spoken narrative text at Time 3:

(24) On this journey *my family and I* want to go Melaka. At Melaka have a many place for example mall and Wonderland. On this year *my family and I* decide to go Melaka because we want to have new experience. First day *I and my family* move at 8.00 p.m. We arrive at Melaka at 12.00 p.m. [...] After we go to Wonderland *my family and I* go to eat. (S19c)

Similarly, Student 38 made use of the innovative form *me and my family* and the conventional form *my family and I* in his spoken narrative text at Time 3:

(25) And then after we arrive at Cambodia my aunty was pick *me and my family* to my grandmother's house. [...] After that *my family and I* was eat together with

my aunty and the next activity is we go back to my grandmother house and we talk about the next activity. (S38c)

The evidence here suggests that even though these students have been cultured into making use of conventional forms (e.g., *my family and I, my brother and I* and *my cousin and I*) which is realized in the substantial use of conventional forms in this functions in both corpora, they still preferred using the innovative forms (e.g., *I and my family, me and my family, I and my brother* and *I and my cousin*). This shows that in the course of language development, the use of learner language features (e.g., *I and my family, me and my family*) may not necessarily mean the inadequacy of the students in acquiring the correct forms but rather a matter of choice of the students in deciding which form they intend to use in order to convey the intended message.

In addition to that, inconsistency and dynamicity are noted in the written and spoken language of these students over time which suggest that the nature of language development is inconsistent and dynamic instead of fixed and linear. For instance, as noted in the earlier section, the students made use of more adjective types occurring before the nouns *day* and *moment* in the spoken narrative texts than in the written narrative texts at two points in time. They also made use of more AdjP-based bundles in the spoken narrative texts than in the written narrative texts at Time 1. The inconsistencies in the use of increased adjective types and in the use of increased AdjP-based bundles in the spoken corpus at one point in time and in the written corpus at another point in time suggest that students may take on diverging developmental paths in their written and spoken language although the written and spoken tasks are based on the same topic. This again shows the complexity involved in the development of the written and spoken language of the students which contradicts the conventional idea of 'developmental ladder' metaphor that views language development as a consistent and linear process. The patterns of inconsistency and dynamicity highlighted in the written

and spoken language over time could also indicate that the students are continuously transforming their linguistic world according to their goals and needs (Larsen-Freeman, 2006) as they develop in their language. In this endeavour, the absence of consistency and linearity during the course of development does not mean that they are regressing in their language production.

Apart from that, the findings on the use of conventional and innovative forms of AdjP-based bundles in the written and spoken corpora showed evidence of students becoming more innovative in their written language while becoming more conventionalized in their spoken language over time. This finding is in fact quite unforeseen because a great amount of instructions on the written language is given to the students in the classroom. They only receive a minimal amount of instruction on the spoken language. Despite being taught the native-like forms in the classroom the students were more innovative in expressing their emotions in their written language while they adopted a more conventional way to express their emotions in their spoken language. Therefore, it can be inferred that despite instructional experience, students' language use is continually changing where dynamicity is at work.

4.5 Conclusion

In this chapter, the findings of the present study have been presented in which the use, structures and functions of four-word LBs in the written and spoken corpora over time was investigated, the extent to which they are different structurally and functionally in the written and spoken corpora were examined and the nature of language development was observed. Discussions on insights drawn from the findings were presented in this chapter as well. In the next chapter, the summary of the findings is provided followed by the implications of the study, the limitations of the study and suggestions for future research.

CHAPTER 5: CONCLUSION

5.1 Introduction

An introduction to the present study was provided in Chapter 1. Relevant literature was reviewed in Chapter 2, the method and procedure of the study were discussed in Chapter 3 and the findings and discussion of the study were presented in Chapter 4. In this final chapter, the summary of the findings is presented. The implications of the study, the limitations of this study and suggestions for future research are provided in this chapter as well.

5.2 Summary of the findings of the study

The summary of the findings observed in this study of the longitudinal written and spoken corpora of 42 students are presented here. First, four-word LBs were found to be slightly more common in the spoken corpus than in the written corpus over time. Quantitatively, there were not many changes noted in the use of LB types and overall frequency of LBs in the written and spoken corpora over time. The students seemed to become less reliant on the use of LBs in the written and spoken narrative texts within the six months of observation. The students tended to be less flexible in the use of LB types in the written narrative texts but were relatively more flexible in the use of LBs types in the spoken narrative texts. The findings here unveil the relative demands of the written and spoken language. Due to minimal preparation time during the speech, they were more likely to rely on LBs that were stored and retrieved from their memory (Wray & Perkins, 2000).

Second, structurally, more than half of the bundles in both corpora were made up of clausal bundles (i.e., VP-based and DC-based bundles). This suggests that both the

written and spoken narrative texts are mostly clausal in nature. NP/PP-based bundles (i.e., phrasal bundles) were found to be common in the written and spoken corpora as well but not more than VP-based bundles which indicates that the written narrative texts are slightly more phrasal than the spoken narrative texts. The findings on the VP-based and NP/PP-based bundles are in line with Biber et al. (2004) to a certain extent. Biber et al. (2004) found that the spoken registers consisted of a greater number of VP-based bundles whereas the written registers consisted of a greater number of NP/PP-based bundles. Similarly, in this study, although VP-based bundles were most widely used in both the corpora, the spoken corpus still comprised a slightly larger proportion of VP-based bundles than the written corpus. The proportions of NP/PP-based bundles were slightly larger in the written corpus than in the spoken corpus although not more than VP-based bundles.

Third, functionally, most of the bundles in both corpora functioned as referential expressions. This was followed by bundles that functioned as topic-oriented expressions. The substantial use of referential expressions and minimal use of stance and discourse bundles in the written and spoken narrative texts despite the difference in the modes of production (i.e., written and spoken) may indicate the possible requirement of the narrative genre that is descriptive in nature. This is made evident through the fact that the functions of bundles in the corpora are rather distinctive in comparison to Biber et al.'s (2004) functions. The narration of life experiences through writing and speech is clearly a different genre from Biber et al.'s (1999, 2004) written and spoken registers (i.e., textbook, academic prose, conversation and classroom teaching). Hence, this requires for the use of a different repertoire of linguistic features. This is possibly why the written and spoken narrative texts are heavily reliant on 'literate' bundles (i.e., referential expressions) and less reliant on 'oral' bundles (i.e., stance expressions and discourse organizing bundles).The functional analysis of the

bundles also revealed that the bundles used did not demonstrate a clear written or spoken production specific set of bundles such as in the case of Biber et al. (2004). The students made low use of stance expressions and high use of referential expressions in their spoken narrative texts. At the same time, the written narrative texts consisted of stance expressions and bundles with special conversational functions displaying a mixture of 'literate' and 'oral' bundles to some extent in both written and spoken narrative texts. This phenomenon could be a clear indication of the characteristic of the narrative genre.

Fourth, two different analysis on AdjP-based bundles were conducted (i.e., EA and analysis of learner language in its own right) to understand the difference between measuring learner language based on the NS norms and treating learner language in its own right. Two error types were identified through the EA in the use of AdjP-based bundles: the omission error and misinformation error of copula *be*. The EA revealed that the frequency of errors in the use of AdjP-based bundles increased in the written corpus over time whereas the frequency of errors in the use of AdjP-based bundles decreased in the spoken corpus over time. This suggests that the students may have not completely acquired the correct usage of copula *be* especially in their written language. However, the use of these bundles in the spoken language is observed to get better over time. The omission of copula *be* can be said to be interlingual errors which is a result of the mother tongue influence of the students (Richards & Sampson, 1973; Ellis & Barkhuzien, 2005). The possible cause for the misinformation error in the use of copula *am/are* instead of *was/were* in order to indicate past tense could indicate the students' inability to make use of the correct form of copula *was/were*. On the other hand, the analysis of learner language in its own right on the use of AdjP-based bundles showed that these students have adopted a more conventionalized way (e.g., *I was very happy*) of expressing their emotions in their spoken language. This is not the case in their

written narrative texts as they have attempted to express their emotions in a more innovative way (e.g., *I very happy because*) in their written language. From here, it can be said that the students have become more cultured into using native-like forms their speech than their writing.

Taken together the findings based on these two different analysis conducted, it can be inferred that the EA approach is one-sided and it places a value judgement upon the language production of the students where it only accounts for what the students did wrong. On the other hand, the treatment of learner language in its own right does not impose a value judgement on the students' production. The latter displays a much balanced view on the production of students and how they have produced them differently from the traditional grammatical perspective.

Fifth, based on the findings on the use, structural and functional analysis of four-word LBs as well based on the insights drawn from the two different analysis on the use of AdjP-based bundles in the written and spoken corpora, the nature of language development is observed to involve complexity, inconsistency and dynamicity. Moreover, the findings based on the use of referential expressions showed the complexities involved in the students' language development where the use of bundles may be a matter of choice of the students to convey the intended message rather than the inadequacy of the students in producing correct forms. Besides, increasing specificity in the use of referential expressions was observed in the spoken narrative texts over time. This suggests that the increasing specificity in the spoken language may be an indication of development that is taking place in the spoken language. Apart from that, the inconsistencies in the use of increased adjective types and in the use of increased AdjP-based bundles in the spoken corpus at one point in time and in the written corpus at another point in time suggest that students may take on diverging developmental paths in their written and spoken language although the written and spoken tasks are

based on the same topic. The increasingly innovative style employed in the written language and the increasingly conventionalized style in the spoken language over time despite instructional experience can be said to be displaying dynamic patterns in the written and spoken language production.

5.3 Implications of the study

This study, on a practical level, contributes to the national secondary school English language classroom to a certain extent. The less reliance on LBs as building blocks in their written and spoken language over time does not mean that LBs are to be treated as unimportant sequences. Researchers argue that LBs can act as frames whereby the possible variations that occur within the frames can be identified (Hunston, 2002; Bennett, 2010). For instance, the pattern, *I was [very] happy* can be considered as a frame to which the variations of adverb (e.g., *very*) used for degree modification can be identified and taught to students. Below are first 50 lines from the BNC written texts to illustrate the adverbs occurring with *happy*. The adverbs that are used to modify degree that precede the adjective *happy* in these lines are: *very*, *most*, *really*, *so*, *as*, *totally*, *quite*, *more than* and *equally*. It is noted that some of the adverbs occur several times, others only once or twice.

1	the test to be positive. I was a very	**happy**	man. There was clearly a need to
2	than just physical attraction to make a lasting,	**happy**	marriage. Strong friendship takes time
3	Friendship takes time. In most	**happy**	marriages, husband and wife
4	any moment. I remember feeling really	**happy**	when I was told I had at least six
5	by some secret action, and now I am so	**happy**	I found your letter concerning my
6	concerning my person. I am thankful and	**happy**	that there was a strange and
7	being the number of items to be catalogued.	**Happy**	indeed is the curator of a select
8	features of the working middle class,	**happy**	in its tranquillity, its labour and its
9	collectors and travellers were	**happy**	to have copies of pictures in
10	The common characteristic is the	**happy**	acceptance of whatever
11	many ugly people would appear to be just as	**happy**	, and just as emotionally fulfilled,
12	that challenge, just as I think he would have been	**happy**	to agree that it is possible to speak

13	could not find a hotel they were	**happy**	with.' Mr Kidd is also chairman
14	customers leave a generous gratuity, obviously	**happy**	with the meal and service.
15	being, with a young family at home, he is very	**happy**	to have switched to contract
16	and Pierre for £1.50 a month. I wasn't at all	**happy**	but I didn't have much choice.
17	glasses in a spontaneous salute to the	**happy**	couple, to the perfect English
18	coated with a kind of raw ketchup. He was a totally	**happy**	— if less than salubrious — man.
19	the milkman's horse every morning she's	**happy**	. Little sisters are spasmo.
20	I was beginning to feel quite	**happy**	again. I screwed my
21	Henry Tyler would not have described her as a	**happy**	woman, but afterward he
22	Festival's come of age of late and I am more than	**happy**	that my sax playing
23	and film. His three plays about Scotland (THERE IS A	**HAPPY**	LAND, BORDER WARFARE,
24	the verge of retirement and Paul would be	**happy**	to take my place. 'I hear you're
25	that you want to share with us, I'm more than	**happy**	to arrange another lecture for you
26	I still carry it around with me. We both look so	**happy**	and relaxed in it. Then, as now
27	the belief that they were the leaders of tomorrow, the '	**happy**	few' who would one day be
28	and at the end of it all, everyone is quite	**happy**	to settle for a draw. It's rather
29	The manager there was quite	**happy**	to take people on social security
30	's against you at the moment.' 'We'd be	**happy**	with just a council flat but there
31	fairly quickly became pregnant. Despite this	**happy**	event, the marriage seemingly
32	as well. John and I got on — and I was	**happy**	to join in with the general
33	go to Rome; Rome, where I had spent so many	**happy**	days in the past, would be my final
34	Incongruously, this was also a very	**happy**	period for me in many ways.
35	the body buried before you do anything else? I'm quite	**happy**	to wait.' 'You're probably right.
36	'How about one for the	**happy**	couple?' she suggested.
37	like it — and the rail staff at Colchester have been quite	**happy**	to let me have a break in
38	nature's own living pest controllers	**happy**	in the organic gardens
39	and add a touch of humour too. Until next month,	**happy**	reading and good gardening.
40	r (and daisy flowers always make me feel	**happy**	the same arrangement and
41	days with colour, bold foliage and, let's not forget,	**happy**	memories. Leucanthemella
42	also grows and looks well by water, but is equally	**happy**	in the deep rich soil of a
43	I try to maintain a regime that keeps wildlife	**happy**	. And that means regular work.
44	flowers in winter Many spiraeas will be	**happy**	in partial shade, brightening
45	the final turn, you are	**happy**	that you have got the field
46	the modern tendency is to be	**happy**	with a few hours of stall avoidance
47	as you that you are healthy and	**happy**	during pregnancy, and that you should
48	if they're not satisfied. If you aren't	**happy**	with the service you're getting from
49	Gradually she confided in Jay who was	**happy**	to listen and soothe. Flattered, even,
50	read her a bedtime story.	**Happy**	days! But it had all soured later.

Here the student does not acquire *happy* as a single word unit but as containing several phraseologies as the student is exposed to a few different ways of expressing the extent of being *happy* with the use of degree modifying adverbs. The student faces a difficult task in acquiring several adverbs occurring with *happy*. At the same time, the

student gains clarity as he or she gets more information about its use (Hunston, 2002). Even the bundles with the pattern, '1st person pronoun + copula *was* + AdjP can be considered as a frame. The occurrence of different adjectives in this frame can be identified and taught to students as well. This might be a promising way to teach students the preferred ways of expressing meanings in their written and spoken narratives.

DDL approach has begun to gain the interest of the teachers in recent years (Hyland, 2013). The use of corpora in DDL enables students to search and analyse facts about language use. In this study, it is suggested that DDL can be a complementary pedagogical tool in the language classroom in Malaysia (O'Keefee et al., 2007). In other words, DDL approach can be considered as an added instructional tool in the teaching and learning process (see Kamariah & Su'ad, 2011, 2014 for the efficacy of using this method in Malaysia). For instance, the use of frames to acquire preferred phrases as emphasized above can be made possible through DDL method. Although this practice seems alien to the national secondary level English classroom, it is rather not impossible (Hajar, 2014). Schools today are equipped with computer facilities and teachers do carry out the teaching and learning process in computer labs. DDL approach as a learning activity is inductive as it involves students to get hands-on experience in learning how to use preferred phrases to make meaning (Chambers, 2010). This approach is said to be far more informative than dictionaries, grammar books and so on (Chambers, 2010). The implementation of DDL approach can be carried out by either the teacher who makes use of a corpus to identify and teach phraseology to the students or students themselves who can directly make use of the corpus. During DDL session, the teacher does not become the sole source of knowledge but a facilitator of the learning session. At the same time, it is also important to highlight that the implementation of DDL in the English classroom is not without complications. This is because this approach as a

pedagogical method will require expertise in handling corpus tools (Hyland, 2013) (see also Hunston, 2002 for a detailed explanation on challenges of DDL in language teaching).

In addition to that, on an empirical level, most importantly, this study calls to reshape the conventional notions attached to learner language as a flawed system of TL to an independent system (Larsen-Freeman, 2006; Chau 2012, 2015; Cook, 2012; Garcia, 2014). The empirical evidence of the present study shows the disadvantage of evaluating learner language from an EA perspective (see Chapter 4, section 4.4.1 on two different analysis on the use of AdjP-based bundles) as well as the advantage of methodologically treating learner language in its own right. By arguing for the need to treat learner language in its own right, the researcher does not intend to simply disqualify the profound works of grammarians as well as traditional grammar. What is more worrying is the underlying assumption that anything that is of learner language that is unfitting to the socially accepted and idealized NS language norm is basically an 'error'. The concept of error is challenged here when the need for reconsideration does not lie at the part of the student but at the socially and traditionally accepted conceptualization of structures of the TL. For example, the innovative forms such as *I very happy because* should be viewed as a product of learner language and not as an 'error' due to lack of mastery of a certain linguistic item. Students will always appear to be failures if they are measured in terms of what they cannot produce. Learner language will always be a flawed language system if it is measured in terms of NS norms. The findings of this study also pose a question on the traditional assumptions placed upon the nature of language development as a fixed and linear process. This is because complexity, inconsistency and dynamicity are observed in the written and spoken language of 42 students over time.

5.4 Limitations of the study and suggestions for future research

There are a number of limitations in this study. Firstly, this study is aimed to identify the four-word LBs found in the written and spoken corpora of students over time. Therefore, the evaluation of the important bundles for language teaching is beyond the scope of this study. Nevertheless, this can be a starting point for future research to consider the evaluation of important LBs that can contribute to language teaching. Moreover, as far as the application of the present study is concerned, this study has little to contribute to language teaching. As noted in the first few chapters, LBs are uninterrupted sequences retrieved through an automated process. As a consequence, a large proportion of LBs that are made up of fragmented phrases are extracted from the corpora (see appendices A & B) and this further complicates the analysis of structures and functions of the LBs extracted. This is why quite a number of main structural categories and subcategories of LBs were identified in the present study (see Chapter 3 for the complete modified structural framework). Many of these fragmented bundles seem to be meaningless sequences to be contributed for language teaching. It is also realized through this study that LBs are not sequences that students use to create meaning but they are rather research tools used to analyse the students' language production. This implicates that future researchers who intend to investigate LBs in the language use of secondary school students should exercise caution.

Secondly, the empirical evidence suggests that research spanning longer than six months is required to track language development. This is because not many differences were noted in the quantitative analysis of LB types and overall frequency of LBs used in the written and spoken narrative texts over time. Thus, it is suggested that future studies should consider a longer period of observation to track language development. Thirdly, the sizes of the written and spoken corpora are rather small for generalizability of the findings. The written and spoken data samples of the corpora are limited to 42 students.

In the future, researchers can consider including a wider range of written and spoken samples of students of English. Apart from that, this study takes on a longitudinal approach to studying the use of LBs as well as observing the nature of language development, hence, it does not account for individual variability of students in the use of LBs in the written and spoken narrative texts which can also be considered as a limitation. Lastly, this learner corpus study does not take into account the influence of instructional experiences in the use of LBs in the written and spoken corpora over time during the six months of observation. This can be another direction for future phraseological studies that deal with LBs in the language use of students to investigate the influence of input gained during instructional experiences.

5.5 Conclusion

The aims of this study have been to investigate the use, structures and functions of four-word LBs in the written and spoken corpora over time as well as to observe the nature of language development. The findings of the present study support a range of phraseological studies and SLA studies, extend the studies in the past and contribute to the research on LBs (e.g., Biber et al., 1999, 2004; Hyland, 2008a; Chen & Baker, 2010, 2016) by examining the use, structures and functions of LBs as well as examining the nature of language development. The students seemed to become less reliant on LBs in the written and spoken narrative texts over time. LBs tend to be genre specific, ultimately functioning to meet the needs of the narrative genre in the present study which goes beyond the needs of the modes of production. The nature of language development is observed to involve complexity, inconsistency and dynamicity within the written and spoken language as well as between the written and spoken language of students. The nature of language development is also observed to include developing

towards specificity and a matter of choice of the students in making use of preferred bundles to convey the intended message in their written and spoken narrative texts.

REFERENCES

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.

Adolphs, S., & Knight, D. (2010). Building a spoken corpus: What are the basics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 38-52). Abingdon: Routledge.

Akinnaso, F. N. (1982). On the differences between spoken and written language. *Language and Speech, 25*(2), 97-125.

Allan, R. (2016). Lexical bundles in graded readers: To what extent does language restriction affect lexical patterning? *System, 59*, 61-72.

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

Anthony, L. (2014). AntConc (Version 3.5.6) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software.html

Arshad Abdul Samad. (2004). Beyond concordance lines: Using concordances to investigating language development. *Internet Journal of e-Language Learning & Teaching, 1*(1), 43-51.

Barber, C., Beal, J. C., & Shaw, P. A. (2009). *The English language: A historical introduction* (2nd ed.). New York: Cambridge University Press.

Bennett, G. R. (2010). *Using corpora in the language learning classroom.* Michigan: University of Michigan Press.

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. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics, 14*(3), 275-311.

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

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at…*: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL QUARTERLY, 45*(1), 5-35.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. England: Pearson Education Limited.

Bonelli, E. T. (2010). Theoretical overview of the evolution of corpus linguistics. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 14-27). Abingdon: Routledge.

Botley, S., & Dillah, D. (2007). Investigating spelling errors in a Malaysian learner corpus. *Malaysian Journal of ELT Research, 3*, 74-93.

Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 35-55). Cambridge: Cambridge University Press.

Chambers, A. (2010). What is data-driven learning? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 152-166). Abingdon: Routledge.

Chan, S. H., Hadi Kashiha, & Tan, H. (2014). Lexical bundles: Facilitating university "talk" in group discussions. *English Language Teaching, 7*(4), 1-10.

Chau, M. H. (2008). Developing phraseological competence: Insights from structural and functional analyses of a learner corpus. Unpublished master's dissertation, University of Nottingham.

Chau, M. H. (2012). Learner corpora and second language acquisition. In K. Hyland, M. H. Chau & M. Handford (Eds.), *Corpus application in applied linguistics* (pp. 191-207). London: Continuum.

Chau, M. H. (2015). *From language learners to dynamic meaning makers: A longitudinal investigation of Malaysian secondary school students' development of English from text and corpus perspectives*. Retrieved from University of Birmingham eTheses Repository (ID Code 6087).

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30-49.

Chen, Y.-H., & Baker, P. (2016). Investigating critical discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics, 37*(6), 849-880.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and non-native speakers? *Applied Linguistics, 29*(1), 72-89.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics, 32*, 45-61.

Conrad, S. (2010). What can a corpus tell us about grammar? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 227-240). Abingdon: Routledge.

Conrad, S., & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. *Applied Linguistics, 20*, 56-71.

Cook, V. (1992). Evidence for multicompetence. *Language learning, 42*(4), 557-591.

Cook, V. (2012). Multicompetence. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-6). Oxford: Wiley-Blackwell.

Corder, S. P. (1967). The significance of learner's errors. *International Review of Applied Linguistics in Language Teaching, 5*(4), 161-170.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*(4), 397-423.

Cowie, A. P. (1998a). Introduction. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 1-20). Oxford: Clarendon Press.

Cowie, A. P. (1998b). Phraseological dictionaries: Some east-west comparisons. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 209-228). Oxford: Clarendon press.

Creese, A., & Blackledge, A. (2010). Translanguaging in the bilingual classroom: A pedagogy for learning and teaching? *The Modern Language Journal, 94*(1), 103-115.

Crossley, S., & Salsbury, T. L. (2011). The development of lexical bundle accuracy and production in English second language speakers. *International Review of Applied Linguistics in Teaching, 49*(1), 1-26.

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics, 3*(1), 59-80.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literature New Series, 2*, 225-246.

Ebeling, S. O., & Hasselgård, H. (2015a). Learner corpora and phraseology. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 207-229). Cambridge: Cambridge University Press.

Ebeling, S. O., & Hasselgard, H. (2015b). Learners' and native speakers' use of recurrent word-combinations across disciplines. *Proceedings of the LCR2013 Conference, 6*, 87-106. doi: http://dx.doi.org/10.15845/bells.v6i0.810

Ellis, N. C. (2008). Phraseology: The periphery and the heart of language. In F. Meunier & G. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 1-13). Amsterdam: John Benjamins.

Ellis, N. C., O'Donnell, M. B., & Romer, U. (2013). Usage-based language: Investigating the latent structures that underpin acquisition. *Language Learning, 63*(1), 25-51.

Ellis, R. (1985). *Understanding second language acquisition.* Oxford: Oxford University Press.

Ellis, R. (1997). SLA and language pedagogy. *Studies in Second Language Acquisition*, *19*(1), 69-92.

Ellis, R., & Barkhuizen, G. (2005). *Analyzing learner language.* Oxford: Oxford University Press.

Ellis, N. C., 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.

Elturki, E., & Salsbury, T. (2015). Little voice mine: Semantic prosody and formulaic sequences in a longitudinal learner corpus. *English Teaching & Learning, 39*(4), 93-132.

Firth, J. R. (1957). Modes of meaning. In J. R. Firth, *Papers in linguistics 1934-1951* (pp. 190-215). London: Oxford University Press.

Garcia, O. (2014). TESOL Translanguaged in NYS: Alternative perspectives. *NYS TESOL JOURNAL, 1*(1), 2-10.

Granger, S. (1998a). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 145-160). Oxford: Oxford University Press.

Granger, S. (1998b). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3-18). London: Longman.

Granger, S. (2003). The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly, 37*(3), 538-546.

Granger, S. (2014). A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast, 14*(1), 58-72.

Granger, S., Gilquin, G., & Meunier, F. (2015). Introduction: learner corpus research – past, present and future. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 1-5). Cambridge: Cambridge University Press.

Greaves, C., & Warren, M. (2010). What can a corpus tell us about multi-word units? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 212-226). Abingdon: Routledge.

Hadi Kashiha, & Chan, S. H. (2014). Structural analysis of lexical bundles in university lectures of politics and chemistry. *International Journal of Applied Linguistics & English Literature, 3*(1), 224-230.

Hajar Abdul Rahim. (2014). Corpora in language research in Malaysia. *Kajian Malaysia, 32*(1), 1-16.

Hakuta, K, (1974). Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning, 24*(2), 287-298.

Halliday, M. A. K. (1994). *Functions of Language.* (2nd ed.). London: Arnold.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics, 19*(1), 24-44.

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.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, S. (2010). How can a corpus be used to explore patterns? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 152-166). Abingdon: Routledge.

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. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics, 32*, 150-169.

Hyland, K. (2013). Corpora, innovation and English language education. In K. Hyland & Lillian L. C. Wong (Eds.), *Innovation and change in English language education* (pp. 218-232). London: Routledge.

Jespersen, O. (1924). *The philosophy of grammar.* Abingdon: Routledge.

Johnson, K. (2008). *An introduction to foreign language learning and teaching (learning about language)* (2nd ed.). London: Longman.

Kamariah Yunus, & Su'ad Awab. (2011). Collocational competence among Malaysian undergraduate law students. *Malaysian Journal of ELT Research, 7*(1), 151-202.

Kamariah Yunus, & Su'ad Awab. (2014). The impact of data-driven learning instruction on Malaysian law undergraduates' colligational competence. *Kajian Malaysia, 32*(1), 79-109.

Kennedy, G. (1998). *An introduction to corpus linguistics.* London: Longman.

Kennedy, G. (2003). *Structure and meaning in English: A guide for teachers.* London: Longman.

Kramsch, C., & Whiteside, A. (2007). The fundamental concepts in second language acquisition and their relevance in multilingual context. *The Modern Language Journal, 91*, 907-922.

Larsen-Freeman, D. (1997). Chaos/Complexity science and second language acquisition. *Applied Linguistics, 18*(2), 141-165.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*(4), 590-619.

Larsen-Freeman, D. (2011). The emancipation of the language learner. *Second Language Learning and Teaching, 2*(3), 297-309.

Larsen-Freeman, D., & Cameron, L. (2008). Research methodology on language development from a complex systems perspective. *The Modern Language Journal, 92*(2), 200-213.

Larsen-Freeman, D., & Long, M. H. (1991). *An introduction to second language acquisition research.* Essex, England: Longman Group.

Leńko-Szymańska, A. (2014). The acquisition of formulaic language by EFL learners: A cross-sectional and cross-linguistic perspective. *International Journal of Corpus Linguistics, 19*(2), 225-251.

Li, W. (2017). Translanguaging as a practical theory of language. *Applied Linguistics, 00*(0), 1-23.

Lu, X. (2010). What can corpus software reveal about language development? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 184-193). Abingdon: Routledge.

Mitchell, R., Myles, F., & Marsden, E. (2013). *Second language learning theories* (3rd ed.). London: Routledge.

Moon, R. (1998). Frequencies and forms of phrasal lexemes in English. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 79-100). Oxford: Clarendon Press.

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

Nemser, W. (1971). Approximative systems of foreign language learners. *IRAL, 9*(2), 115-123.

Nieto, S. (2009). *Language, culture, and teaching: Critical perspectives*. New York: Routledge.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching.* Cambridge: Cambridge University Press.

Ong, C. S. B., & Yuen, C. K. (2014). A corpus study of structural types of lexical bundles in MUET reading texts. *3L: The Southeast Asian Journal of English Language Studies, 20*(2), 127-140.

Ong, C. S. B., & Yuen, C. K. (2015). Functional types of lexical bundles in reading texts of Malaysian university English test: A corpus study. *GEMA Online Journal of Language Studies, 15*(1), 77-90.

Ortega, L. (2013). SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language Learning, 63*(1), 1-24.

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics, 25*, 26-45.

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.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32,* 130-149.

Paramasiwam Muthusamy, & Atieh Farashaiyan. (2017). A corpus-based comparative study of Malaysian ESL learners and native English speakers in compliment patterns. *International Journal of Linguistics, 9*(5), 232-246.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistics theory: Nativelike selection and native like fluency. In J. C. Richards & R.W. Schmidt (Eds.), *Language and communication* (pp. 191-226). London: Longman.

Richards, J. C., & Sampson, G. P. (1973). The study of learner English. In J. C. Richards (Ed.), *Error analysis: Perspectives on second language acquisition* (pp. 3-18). London: Longman.

Ruan, Z. (2016). Lexical bundles in Chinese undergraduate academic writing at an English medium university. *RELC Journal*, 1-14.

Sampson, G. (1980). *Schools of linguistics: Competition and evolution.* London: Hutchinson.

Scott, M., & Tribble, C. (2006). *Textual patterns.* Amsterdam: Benjamins.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching, 10*(3), 209-231.

Shirato, J., & Stapleton, P. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research, 11*(4), 393-412.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Siti Aeisha Joharry, & Hajar Abdul Rahim. (2014). Corpus research in Malaysia: A bibliographic analysis. *Kajian Malaysia, 32*(1), 17-43.

Smith, M. S. (1994). *Second language learning: Theoretical foundations.* London and New York: Longman.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences in EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes, 12*(3), 214-225.

Stubbs, M. (2008). The search for units of meaning: A tribute to John McHardy Sinclair (14 June 1933 – 13 March 2007). Germany: University of Erlangen.

Stubbs, M. (2009). The search for units of meaning: Sinclair on empirical semantics. *Applied Linguistics*, *30*(1), 115-137.

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Wang, Y. (2017). Lexical bundles in spoken academic ELF: Genre and disciplinary variation. *International Journal of Corpus Linguistics, 22*(2), 187-211.

Wei, Y., & Lei, L. (2011). Lexical bundles in the academic writing of advanced Chinese EFL learners. *RELC Journal, 42*(2), 155-166.

Wray, A., & Perkins, M. R. (2000). The function of FL: An integrated model. *Language & Communication 20*(1), 1-28.