Lexical Bundles in University EAP Exam Writing Samples:
CAEL Test Essays

by

Randy Appel

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

Master of Arts

School of Linguistics and Language Studies
Carleton University
Ottawa, Ontario

May 31, 2011

© 2011, Randy Appel
NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.
Abstract

This study investigates the use of 4-7 word lexical bundles in the writing of NNS writers taking the Canadian Academic English Language (CAEL) assessment. Written essays sourced from CAEL test archives were used to compile two corpora of non-native English writers’ discourse. These written essays were divided into two corpora of low-level and high-level writers so that comparisons could be made between the two groups. The identified lexical bundles in each corpus were analyzed along three separate lines: functional analysis was used to demonstrate the differences in how lexical bundles were used by each group of writers to structure their discourse; an analysis of the length of lexical bundles used by each group showed how each group differed in terms of how they made use of longer strings of recurrent language; comparing the lexical bundles to the two reading articles given to test takers during their examinations showed how the two groups of writers differed in terms of their reliance on this source in their writing.

The investigation of lexical bundles in non-native English writers discourse contributes to a growing understanding of the differences in how non-native English writers make use of lexical bundles in their writing.
Acknowledgements

First and foremost I would like to thank my family for all the kindness, caring and support I continuously receive from them. My parents have always made themselves available to me no matter how busy they are in their own lives, and without their support I would never have been able to accomplish this task. My brother Darrin has also helped a great deal in this process by giving me the time necessary to return to school and pursue my studies.

I would like to thank Dr. David Wood who has proven to be an incredibly knowledgeable and supportive supervisor. Working under his guidance has been a huge benefit to me and I appreciate the many hours he has devoted to helping me through this process.

My external examiner, Dr. Hedy McGarrell, has also provided insightful comments and suggestions that have proven helpful.

I would also like to thank Wendy Fraser, Dr. Janna Fox and Dr. Devon Woods for their contributions to this process.
Table of Contents

ABSTRACT ............................................................................................................................. I
ACKNOWLEDGEMENTS ......................................................................................................... II
LIST OF TABLES ..................................................................................................................... V
LIST OF APPENDICES .......................................................................................................... VI
1 INTRODUCTION .................................................................................................................. 1
   1.1 RESEARCH QUESTIONS ............................................................................................ 15
   1.2 THE CANADIAN ACADEMIC ENGLISH LANGUAGE (CAEL) ASSESSMENT .......... 16
   1.3 EVALUATION OF WRITTEN RESPONSE PORTION OF THE CAEL TESTS .......... 18
2 CORPUS LINGUISTICS AND SECOND LANGUAGE ACQUISITION .................................... 22
   2.1 RATIONALISM VS. EMPIRICISM ............................................................................ 22
   2.2 CORPUS LINGUISTICS ............................................................................................ 25
   2.3 CORPORAS IN SECOND LANGUAGE ACQUISITION (SLA) ..................................... 32
   2.4 GENERAL VS. SPECIALIZED CORPORAS .............................................................. 34
   2.5 REPRESENTATIVENESS OF CORPORAS ................................................................. 36
3 FORMULAIC LANGUAGE .................................................................................................... 39
   3.1 FORMS OF FORMULAIC LANGUAGE ....................................................................... 42
4 LEXICAL BUNDLES ............................................................................................................. 45
   4.1 LEXICAL BUNDLES AND PSYCHOLINGUISTIC VALIDITY ...................................... 45
   4.2 IDENTIFICATION OF LEXICAL BUNDLES ............................................................... 47
   4.3 PRAGMATIC FUNCTIONS OF LEXICAL BUNDLES ................................................. 49
      4.3.1 Stance Expressions ............................................................................................ 49
      4.3.2 Discourse Organizers ....................................................................................... 50
      4.3.3 Referential Bundles ......................................................................................... 51
5 METHODOLOGY ................................................................................................................ 55
   5.1 IDENTIFICATION CRITERIA OF LEXICAL BUNDLES ............................................. 61
      5.1.1 Identification Criteria for 4-word Lexical Bundles in the LLC and HLC .......... 62
      5.1.2 Identifying lexical bundles of five or more words ........................................... 69
   5.2 WORDSMITH TOOLS 5.0 - LEXICAL BUNDLE IDENTIFICATION ......................... 73
   5.3 ANALYTIC FRAMEWORK ......................................................................................... 74
      5.3.1 Functional Taxonomy - Biber et al. (2004) ....................................................... 74
      5.3.2 Frequency of occurrence and distribution ....................................................... 75
      5.3.2.1 Lexical Bundles and the Required Readings ................................................ 75
6 RESULTS AND DISCUSSION ............................................................................................... 77
   6.1 INITIAL OVERVIEW OF LEXICAL BUNDLES IN THE LLC AND HLC ................. 77
   6.2 FUNCTIONAL ANALYSIS ......................................................................................... 78
      6.2.1 Lexical Bundles in the LLC - Functional Analysis ........................................... 80
         6.2.1.1 Stance Bundles in the Lower Level Corpus (LLC) .................................... 80
         6.2.1.2 Discourse Organizing Bundles in the LLC .............................................. 82
         6.2.1.3 Referential Bundles in the LLC ............................................................... 84
      6.2.2 Lexical Bundles in the HLC - Functional Analysis ........................................... 86
         6.2.2.1 Stance Bundles in the Higher Level Corpus (HLC) .................................. 86
         6.2.2.2 Discourse-Organizing Bundles in the HLC .............................................. 87
         6.2.2.3 Referential Bundles in the HLC ............................................................... 88
      6.2.3 Functional Analysis - Comparing the LLC and HLC ....................................... 90
   6.3 FREQUENCY ANALYSIS OF LEXICAL BUNDLE LENGTHS .................................... 94
      6.3.1 LLC - Frequency analysis of 4-7 word lexical bundles .................................... 95
List of Tables

Table 5.1 Corpus Statistics ........................................................................................................ 56
Table 5.1.1 Summary of the criteria used for the identification of lexical bundles .......... 68
Table 6.2.1.1 Stance bundles in the LLC ............................................................................... 80
Table 6.2.1.2 Discourse organizing bundles in the LLC ......................................................... 82
Table 6.2.1.3 Referential Bundles in the LLC .......................................................................... 84
Table 6.2.2.1 Stance Bundles in the HLC ............................................................................. 87
Table 6.2.2.2 Discourse-organizing bundles in the HLC ......................................................... 88
Table 6.2.2.3 Referential bundles in the HLC ......................................................................... 88
Table 6.2.3 Summary of lexical bundles in the LLC & HLC ................................................. 91
Table 6.2.5 Summary of lexical bundles in the LLC, HLC and Biber et al 2004 (academic prose) ................................................................................................................. 93
Table 6.3.1 4-7 word lexical bundles in the LLC ................................................................. 95
Table 6.3.2 4-6 word bundles in the HLC ............................................................................. 97
Table 6.3.3 Summary of 4-7 word lexical bundles in the LLC and HLC ......................... 100
Table 6.4.1 Originally produced lexical bundles in the LLC .............................................. 103
Table 6.4.2 Originally produced lexical bundles in the HLC .............................................. 104
Table 6.4.3 Originally produced lexical bundles in the LLC & HLC ................................. 108
List of Appendices

Appendix A: Description of CAEL assessment band scores.................................122
Appendix B: 4 word lexical bundles in the Lower Level Corpus.............................123
Appendix C: 4 word lexical bundles in the Higher Level Corpus............................126
Appendix D: 4-7 word lexical bundles in the Lower Level Corpus..........................128
Appendix E: 4-7 word lexical bundles in the Higher Level Corpus..........................131
1 Introduction

The traditional view of language production has often centered on generative grammar rules that are in large part the result of work done by researchers such as Noam Chomsky in the late 1950’s and early 1960’s. Under the generative grammar approach, sentences are classified as either grammatical, or ill-formed based on their adherence to a predefined set of grammatical rules, and language composition is seen as a process of stringing together individual words, one-by-one, in order to express coherent thought.

Although the generative grammar model still holds weight in many linguistic circles and presents a powerful model for our understanding of language composition/processing, it is not without its flaws. One of the major issues associated with the generative grammar approach to language production is that its theories were developed in the absence of computer-aided corpora and therefore “limited to what a single individual could experience and remember” (Sinclair, 1991, p.1). Additionally, the generative grammar system is unable to adequately account for the tendency of native-users of a language to produce pre-formed chunks such as idioms and other commonly re-occurring word-strings. As we will see in the coming sections, these re-current word-strings, or formulaic sequences as they are often called, form a crucial part of language production in both speaking and writing.

As an alternative to the Chomskyan tradition of looking at language production as a rule-governed procedure whereby sentences are created word-by-word using the language’s underlying grammatical system, a relatively new concept has emerged whereby, it is argued, these rules are often by-passed in favour of pre-constructed chunks of language, otherwise known as formulaic sequences. This growing recognition of the
importance of pre-formed chunks of language has led the way to new theories and models to describe how we use language. Usage-based models are one example of a broad category that incorporates formulaic sequences, as a way of explaining language production and processing. In this sense “usage-based” should be looked at as a general term encompassing numerous different types of models, however as Barlow & Kemmer (2000) show, there are many characteristics that are generally shared by any language model falling under this heading. As the name suggests, these models of language rely on the analysis of actual language use in context as the basis for many of their linguistic claims, often making use of corpora. The use of corpora in the development of language models eliminates one of the criticisms of generative grammar models of language i.e. “avoidance of any engagement with what people actually say” (Wray, 2002, p. 13). Also, being a ‘bottom-up’ approach, usage based-models “focus on the role of learning and experience in language acquisition” (Barlow & Kemmer, 2000, p. xi) and recognize frequency of use as an important aspect of linguistic theory. This recognition of the role frequency plays in language use and processing is important in relation to formulaic sequences, and more specifically to lexical bundles. Lexical bundles, a sub-category of formulaic language are defined based on their frequency of occurrence and have been shown to play an important role in the way writers structure their discourse (Biber 2004, Biber & Barbieri 2007, Cortes 2004/2006). While lexical bundles will be defined in further detail below, it is useful to begin with an overview of formulaic language in general before launching into a discussion of lexical bundles. The following paragraphs provide a definition of formulaic language, along with examples which attempt to show
the important role this aspect of language plays within language composition and processing.

The term *formulaic sequence* can be seen as a somewhat broad term that encompasses a wide range of pre-fabricated sentence stems and longer clauses, even full length sentences, which are seemingly stored whole within the minds of speakers/writers. These sequences range from the very short “in conclusion” to much longer formulations, such as the proverb “Don’t count your chickens before they hatch”. Through the examination of formulaic sequences in an increasingly wide range of corpora, researchers have begun to provide greater descriptions of the formulaic language native speakers produce within a variety of registers, genres, and disciplines. While this wealth of research has added a great deal to the field, it has also resulted in a somewhat fractured subject of study with large numbers of researchers adopting a variety of terms to label this aspect of language (Wray lists dozens of such labels in her 2002 book *Formulaic Sequences and the Lexicon*). Despite the variety of terms and definitions used to describe this concept, Wray’s own 2002 definition of *formulaic sequence* is often cited by scholars interested in this subject, and may be helping to provide a higher level of definition to this growing field.

*a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar* (Wray, 2002, p.9)

This definition is still somewhat broad in terms of classifying what actually constitutes a formulaic sequence, however it does mention one of the key points in the
description of formulaic language, i.e. that they are stored and produced whole, as single
units, despite the fact that they may be composed of any number of words. This notion of
being stored as a single unit is an important point since it removes formulaic sequences
from the traditional generative grammar rules of language production.

Despite the fact that formulaic sequences are produced in the absence of
generative grammar rules, their use should not be seen as an indication that these
sequences are unable to be analyzed using said rules, in fact, they most often can. Even
though these sequences are often grammatically sound and could theoretically be
produced using the underlying grammatical rules of the language, the user incorporates
these sequences into their language production as prefabricated units, allowing the user to
skip word-by-word composition and instead produce fully formed sequences as single
chunks of language. As Sinclair (1991) has stated in his work on collocations (one of the
many labels used to refer to formulaic language), “the language user has available to him
or her a large number of semi-preconstructed phrases that constitute single choices, even
though they might appear to be analyzable into segments” (Sinclair, 1991, p. 110).

Wray’s (2002) definition of formulaic sequence given above provides a good fit for these
preconstructed phrases, and ties in well with the “idiom-principle” Sinclair has put
forward. In his 1991 book, Sinclair presents the idea, that language production functions
under two main dichotomous principles; the rule governed open-choice principle (where
grammatical rules dictate which word choices are available to fill each slot in a phrase)
and the idiom-principle (where multi-word units are inserted into discourse, allowing the
user to skip word-by-word text generation). While a unified model would be preferred to
the dichotomous model presented by Sinclair (1991), his proposal of an idiom principle
into his model of language composition represents an important addition. This ability to move from the open-choice principle to the idiom-principle is important for numerous reasons. First of all, the use of formulaic language aids in the production of quick/accurate discourse, since the speakers/writers can more quickly create longer passages of language by using larger chunks of language than would be available to them if they were creating their discourse word by word. This is supported by the fact that "researchers feel that the storage capacity of memory is vast, but that the speed for processing those memories is not, so that we must learn short cuts for making efficient use of this processing time" (Crick: 1979: 219 as cited in Nattinger & DeCarrico, 1992, p. 31). In relation to this point, making use of the idiom-principle and therefore multi-word units (formulaic sequences) can be seen as a way of making more efficient use of our large memories as opposed to relying on our relatively limited ‘on-the-fly’ processing abilities. Additionally, the use of formulaic sequences can be seen as a benefit to the hearer/reader since as Wray argued in her 2002 book on formulaic language, "the more novel our output is for the hearer, the more likely it is to be misunderstood" (Wray, 2002, p. 94). Therefore, by using formulaic sequences a speakers/writers increases the chances that what they are trying to communicate will be understood more quickly and with less processing effort by those with whom they are communicating.

While in the past, idiomatic structures and other forms of formulaic language have often been neglected by researchers addressing language issues, the ever increasing research into formulaic language has led to a growing recognition of the importance of formulaic language and the idiom-principle:
the overwhelming nature of this evidence leads us to elevate the principle of idiom from being a rather minor feature, compared to the grammar, to being at least as important as grammar in the explanation of how meaning arises in text (Sinclair, 1991, p. 112).

Following the publication of this statement in 1991, numerous corpus linguistic researchers (Chen 2010; Cortes 2006; Levy 2003) and psycholinguists (Schmitt, Grandage, Adolphs 2004; Underwood, Schmitt, Galpin 2004) have chosen to place formulaic language at the centre of their studies, contributing to a growing recognition of the importance of this aspect of language.

Although it has been shown that a relatively high degree of formulaic language is present in situations where it is necessary to convey “large amounts of information under severe time constraints” (Schmitt & Carter 2004; Kuiper & Haggo 1984; Kuiper 1996; Wray 2002), i.e. sports announcers, auctioneers, etc, this is by no means the only type of situation where formulaic sequences are seen to be at play in the language processing and production of native speakers. Not only is the idiom principle common in speech, where time constraints are more apparent due to the immediate processing/feedback and response needed, it is easy to identify numerous applications of this principle in writing as well, where native speakers consistently re-produce strings of words as though they were single units (i.e. by and large, on the other hand, as a result, first of all, etc). The use of formulaic language in writing serves many beneficial roles for writers, as Levy (2003) has identified in her work on lexical bundles. Included among these benefits is the fact that “it helps with automaticity or fluency in writing as it believed that there is
neither enough time nor processing space for recalling discrete grammar rules and then applying them to produce an appropriate vocabulary item in its correct grammatical form” (Levy, 2003, p. 4). As we will see later on, this may be particularly true for timed writing tasks since timed writing allows for fewer revisions of any text that is produced.

As previously mentioned, the use of formulaic sequences is also beneficial to anyone reading text, since formulaic language is more likely to be understood by the reader than more novel constructions (Wray, 2002, p. 94). This also helps to explain at least one way in which these structures may become formulaic, if we hear/read a sequence we find easy to understand and well formed, we are more likely to repeat it in our own language production. It seems completely plausible that the more often a sequence is repeated the more likely it is to become stored fully formed in our mental lexicon.

The plausibility of the idiom principle, and therefore the existence of formulaic sequences themselves, has been echoed by numerous scholars, including Altenberg (1998), who in an examination of spoken discourse in the London-Lund Corpus of Spoken English states that this method “is in fact quite plausible, and perhaps frequently utilized, alternative to a more rule governed mode of production” (Altenberg, 1998, p. 116). Nattinger and DeCarrico (1992) in their work on lexical phrases and language teaching also seem to share this view:

- Ones knowledge of language and ones ability to use it, in other words, includes knowledge of how to create sentences ‘from scratch’
- and knowledge of prefabricated patterns (grammatical competence),
- and knowledge [of] how to select and retrieve ready-made
form/function composites (pragmalinguistics competence) for appropriate situations or contexts (socio-pragmatic competence) (Nattinger & DeCarrico, 1992, p.13).

This dichotomous ability of native-speakers to produce language not only on-line while operating under the open-choice principle, but also through the use of the idiom-principle to pull from memory pre-constructed sequences may be one of the keys to achieving native-like fluency in a language. According to Boers, Eyckmans, Kappel, Stenger and Demecheleer (2006) who tested the directed instruction of formulaic sequences to non-native speakers, the increased use of formulaic sequences in non-native speakers’ oral discourse was shown to increase perceived fluency by outside judges. The directed instruction of these sequences to the experimental group in Boers et al’s study was shown to result in greater fluency, improved range of expression, and higher accuracy when compared to the control group. Greater use of formulaic sequences was associated with more native-like speech since correctly memorized sequences could be used without fear of error in their internal structures. Using formulaic sequences can help reduce the possibility of errors since, when correctly used, errors would only be found between formulaic sequences, and not within them. Although their work presents a potentially valuable approach to second language acquisition, language testing and assessment, and native-like fluency studies, as of yet, to my knowledge no similar work has been done using formulaic sequences in written discourse.

As Pawley and Syder have pointed out in their seminal work on native-like fluency and selection, it is not uncommon for non-native speakers to "produce an utterance that is grammatical but unnatural or highly marked"
(Pawley & Syder, 1983, p. 12). Although a common experience among many non-native speakers, most native-speakers of a language will have little to say as to why these supposedly grammatically correct utterances are so non-native like in nature. Since formulaic sequences are often simply embedded in the minds of speakers without being subject to analysis, they present a potentially difficult area of language for non-native speakers to gain a comfortable grasp of. Even shorter 2-word sequences can present difficulty since many word pairings do not follow any preset rules, and often go against the way an English as a Second Language (ESL) speaker would construct the same thought/expression in their first language (L1). Fontenelle, in his work on lexical functions in dictionaries explains this point by comparing the French and English usage of *attention*:

*no rule whatsoever can account for the acceptability of ‘pay attention’ and the illformedness of ‘do attention’. A cross-linguistic comparison with French testifies to this, since ‘payer attention’ is inadmissible in French (‘faire attention’ being the only acceptable combination) (Fontenelle, 1998, p. 192).*

Fontenelle is not the only researcher who sees formulaic sequences as a potential stumbling-block for non-native speakers, according to Liu and his work on idiomatic language, the fixed form of idioms (a type of formulaic sequence) is extremely difficult for non-native speakers to understand and make use of:
for L2 learners even fixedness in form alone should perhaps make an expression an idiom because the compositional invariability itself presents learners with the difficulty of having to grasp the expression as a whole if they want to use it correctly and meaningfully (Liu, 2008, p.15).

Unfortunately generative grammar rules cannot be used to alleviate any of the issues these researchers have identified. Instead, the use of grammatical rules simply lets the speaker know which combinations are possible given the parameters of the language. While the knowledge of which utterances are grammatically acceptable is no doubt beneficial, it also implies a greater amount of variability and flexibility within the language than is actually used by most native-speakers. Many areas of our language production fall into very specific patterns, even though the grammar system allows greater variability. For example, Pawley and Syder (1983) use time keeping as one example of the very specific patterns our language falls into:

1. It’s twenty to six.
2. It’s six less twenty.
3. It’s forty past five.
4. It exceeds five by forty.
5. It’s a third to six.
6. It’s ten minutes after half past five.

Even though each of the examples above is grammatically correct, a native speaker would generally only consider the first example acceptable. Not only are
examples 2) through 6) non-native-like and would likely identify the speaker/writer as non-native if used, they also prove more difficult to interpret because they break from the pre-established language patterns we use when talking about the time. These pre-selected language formulas are seen in many aspects of language, not confined solely to time keeping, and are an important part of being accepted into a language community since straying from these structures may help to identify someone as an outsider. While grammar rules make available to the speaker an endless array of choices, it is the selection of only a relatively small number of these that will be interpreted as being native like. Without a knowledge of, and an ability to use these sequences correctly, ESL learners can be easily marked as non-native like in their composition, despite having a high-level of proficiency with the grammar system.

Although the influence of formulaic sequences on native-like language would no doubt be minute if they were to compose only a small portion of the language native speakers produce, research has shown that these sequences do in fact play a significant role in the way we communicate with others. Analysis on selected works has shown that formulaic sequences appear prominently in native speaker discourse and that over half of both spoken and written language is made up of these multi-word sequences (Erman & Warren, 1999, as cited in Bybee, 2001, p. 15). While idioms may be the most easily identifiable sequences that fit into this category (i.e. kick the bucket, don’t put all your eggs in one basket, like a fish out of water, etc), formulaic sequences exist in all facets of our speech/writing and are used on a daily basis. Types of formulaic sequences include, but are not limited to: social formulas (How’s it going?), discourse-focused formulas (on
the other hand), framing statements (in conclusion), and comparison formulas (in contrast to).

In theory, the ability to use formulaic sequences will not only aid the speaker/writer in composing quick/accurate discourse, but may also help those being communicated with to more easily interpret what is being said. Although it is difficult to test the internal processes a listener goes through in attempting to understand speech, research has investigated how formulaic sequences are read by both native and non-native speakers. It is well known that while reading, our eyes do not move in consistent sweeping lines across the page, carefully reading each word in turn, instead eye movement during reading is marked by jerky patterns, stopping on some words, skipping over others and occasionally going back to previous words for further study. In a study of time spent fixated on specific words while reading, researchers found that “terminal words in formulaic sequences are processed more quickly than the same words in non-formulaic contexts”. This result was seen as providing support “for the position that formulaic sequences are stored and processed holistically” (Underwood, Schmitt, Galpin, 2004, p. 167). More recently, Tremblay (2009) has also explored the processing of lexical bundles by native English speakers. Looking exclusively at 4 and 5-word lexical bundles serving referential functions, Tremblay (2009) used three separate self-paced reading tests in an attempt to determine whether lexical bundles were read more quickly than control non-lexical bundles of identical word length. From these tests, Tremblay (2009) was able to conclude that lexical bundles are in fact read with greater speed than other sentence fragments. Additionally, testing participants’ ability to correctly recall lexical bundles in two sentence-recall experiments, it was found that lexical bundles were
recalled more accurately than non-lexical bundles. These findings alone, however, did not provide sufficient support that lexical bundles are in fact stored fully formed in the minds of speakers/writers, i.e., the possibility still existed that lexical bundles are just processed word-by-word, albeit at a faster rate than non-lexical bundles. To address this issue more effectively, Tremblay (2009) decided to use a "regression-type immediate free recall experiment with event-related brain potential recordings" (Tremblay, 2009, p. 90). Analyzing the event-related potentials, Tremblay came to the following conclusion:

Given that the frequency of a whole sequence modulated event-related potentials in a manner that was similar to the effect of single-word frequency on the same brain waves, I conclude that at least some aspects of multi-word sequences are stored and retrieved holistically (as opposed to being computed on-line). (Tremblay, 2009, p. 91)

Tremblay's (2009) in-depth analysis of the processing of lexical bundles provides additional support for the notion that lexical bundles are indeed a distinct type of lexical unit that can be stored holistically in the mind. While these are only two studies and additional research is needed, it does suggest that, with at least some formulaic language, readers are able to pull the meaning of formulaic sequences from the text without needing to analyze all of their constituent parts, thereby speeding up the reading process and aiding in the quick interpretation of text. It would therefore seem important not only for ESL speakers to develop an ability to produce a wide variety of these sequences, but also to be able to interpret them quickly.
when they are conversing with native speakers or reading texts produced by native
speakers.

The term *formulaic sequence* covers a broad spectrum encompassing many
different types of holistically stored and produced multi-word sequences, however the
present study has chosen to deal with a specific sub-category of formulaic language:
lexical bundles. Lexical bundles are a much more specifically defined sub-set that is
confined to statistically derived sequences that are based on frequency counts. Being
statistically derived, the identification of lexical bundles is dependent upon computer
aided corpus analysis. The search for, and identification of lexical bundles has yielded
interesting results in native speaker writing, however research into L2 writing is lacking
when it comes to this type of research. As of yet, research done into the lexical bundles
L2 writers produce has most often been performed using general corpora which do not
take into account the varying abilities of the L2 writers whose writing has contributed to
the corpora used in these studies. Instead these writers are grouped together to form a
general view of ESL writing. While this technique can give a broad view of the
discrepancies between L1 and L2 writers, it does little to explain the differences within
L2 writer populations and how these writers may progress toward a native like level of
language production. In their own work on formulaic sequences, Schmitt and Carter
(2004) have expressed the need for further research into how ESL writers acquire and
make use of these sequences “There may well be an underlying systematicity to the
acquisition and use of L2 formulaic language, but there is simply not enough focused
research at present to say very much with conviction” (Schmitt & Carter, 2004, p. 13).
Additionally, more research into the differences among L2 writers may help to better explain how native speakers judge the proficiency of non-native writers.

1.1 Research Questions

While prior research into formulaic language and lexical bundles in L2 academic writing has most often focused on native speakers, or made the choice to group all L2 speakers together without differentiating between levels of proficiency, the research presented in this paper attempts to show differences among various levels of proficiency within a population of L2 writers, as evaluated by outside judges.

Given the previously described need for a greater level of knowledge in relation to the lexical bundles used by non-native speakers, the present research aims to answer the following two research questions:

1. **Is the frequency with which lexical bundles are used greater for L2 writers evaluated as being more advanced (greater perceived fluency)?**

2. **Do the functions of these lexical bundles differ according to perceived level of fluency, as judged by those evaluating the written response portion of the Canadian Academic English Language assessment tests? If so, in which ways?**

In order to answer these research questions, it was first necessary to identify a valid and useable source of L2 English writing. Given the nature of many of the previously assembled L2 English language corpora (i.e. speakers/writers included were not generally sorted based on their English language ability), none of the publicly available L2 English
corpora were suitable to address the aforementioned research questions. To adequately address these research questions, it was therefore necessary to gain access to a large sample of L2 English writing that would allow for the grouping of writers based on their level of fluency. The archives of the Canadian Academic English Language (CAEL) assessment provided an excellent source of writing that would allow for the achievement of this goal.

1.2 The Canadian Academic English Language (CAEL) assessment

The CAEL test is a standardized assessment of test takers’ proficiency in academic English. Using nine different grading bands, the test evaluates students attempting to gain admission to English speaking Canadian universities and colleges. The assessment evaluates a students’ ability to use English “as it is used in Canadian universities and colleges” (Canadian English Language Assessment, 2010) and uses four different sections to evaluate test takers’ language abilities: reading response, lecture response, oral language response, and written response. Although different universities and colleges use different minimum score requirements for admission into undergraduate and graduate level programs, the CAEL website lists a minimum score of 50 (out of a maximum possible score of 90) as necessary in order to be considered sufficiently proficient to be admitted to English speaking post-secondary institutions (Canadian English Language Assessment, 2010). Since the CAEL assessment uses a topic-based approach, it is unique among standardized tests, and can provide a highly specialized corpus for analysis since past tests are archived and available to researchers if the proper procedures have been followed in order to gain access to them.
Although both the oral and written response portions of this test provide potentially valuable information for study, given the goals of my research, I have chosen to focus on the handwritten “written response” portion of this test. Following the reading task, lecture response, and an oral language response, the final portion of the CAEL assessment is the “written response”. In this section, students are expected to write anywhere from one to three pages in response to a prompt. While test takers are given a total of 2 hours and 15 minutes to complete the entire test, the written response portion is allocated a total of 45 minutes and is the final task test takers are required to complete. The prompt used to elicit response asks the test taker to: “agree or disagree with a claim”, “argue for or against a position”, or “discuss advantages and disadvantages of a course of action” (Canadian English Language Assessment, 2010). Since the test version used for the research presented in this paper is still being used today, it was necessary to change some of the content words so that the ‘true’ topic of the test would not be revealed. Previous tests are kept archived under lock and key in order to ensure the confidentiality of test takers’ writing and keep the topic of the tests secure and confidential. It is important for the reliability of these tests and the CAEL test system itself that test topics are kept secure and private.

Many types of research have been done using CAEL test data in the past, however, no other research of this type has been performed using essays written by non-native English writers of differing levels of proficiency. One of the major advantages of studying formulaic sequences in this manner is the highly specialized nature of the discourse test takers have created. Since test takers write on only one topic per test, the writing can easily be compared and analyzed for the presence of lexical bundles. Lexical
bundles are often closely related to the type of writing being performed and the context of
the writing, being able to study their use in a very confined context is a rare opportunity
within corpus linguistics. Additionally, since the prompt used in this version of the
CAEL test asks writers to present their argument by either agreeing or disagreeing with
the given prompt, we are also presented with a specific type of argumentative writing
which holds its own framing statements and formulaic language.

1.3 Evaluation of written response portion of the CAEL tests

The criteria used to evaluate each component of the CAEL assessment was created
by a panel of seasoned ESL/EAP professionals and test developers coming from the
Language Assessment and Testing Research Unit (LATRU) of Carleton University’s
School of Linguistics and Applied Language Studies. Writing performance in the written
portion of the CAEL assessment is divided into nine grading bands ranging from 10-90.
These bands span the range from “Very Limited Writer (10-20) to Expert Writer (80-90)”
(Canadian English Language Assessment, 2010). Please refer to Appendix A for a
detailed description of each of the nine grading bands used in the CAEL assessment.

Evaluation of test taker writing follows the “collaborative read-aloud marking
protocol” whereby a team of three judges meets to evaluate the writing produced by test
takers. During these meetings, one judge will read aloud the written response portion of
the test while the other two judges listen attentively.\(^1\)

\(^1\) All judges are experienced ESL/EAP teachers who have been trained in the methods
used to evaluate the CAEL assessment.
The writing samples are read aloud in order to minimize the impact of surface features in the text such as spelling and handwriting on the score assigned. The impact of these surface features has been researched extensively and can be substantial. Raters are encouraged to place greater emphasis on other more substantial features (e.g., meaningfulness, cohesiveness, flow, etc.) of the essays and to discount these surface features when assigning a score (Canadian English Language Assessment, 2010).

Once the written response portion of the test has been read in full, each judge will record some brief comments along with their impression of the writing and a score for the essay that was just read (10-90). Once each judge has decided on a grade for the writing in question, the three judges will compare grades to see if a consensus has been reached. If all three raters share the same evaluation of the test, this mark is recorded, and the process is repeated for the next writing sample (Fox, Pychyl, Zumbo, 1993). If all three judges do not share the same evaluation of the writing in question, a discussion begins where judges explain their rational for the grade they have assigned to the writing. In the rare occurrence where consensus can still not be reached after discussions have been held, a second group of judges can be called on to re-evaluate the test in question.

With the goal of ensuring inter-rater reliability, every rating session includes the re-evaluation of previously marked writing samples. These re-marked tests are recorded, and correlations are calculated. The most recent correlation figures reported by CAEL (based on n=178) show that 87.1% of the remarked tests were given the same grade in
each of the grading sessions, 10.1% had a difference of 1 band score, and 2.2% had a change of 2 band scores (Canadian English Language Assessment, 2010).

The writing produced by CAEL test takers can be considered to come from a unique sub-section of the academic genre/register since it is a time constrained writing task and the ability to revise the discourse produced is therefore limited. While “the purpose of the CAEL Assessment is to test English in use for academic purposes (EAP)” (Fox, 2002, p. 9), it does not provide an exact replication of the writing environment that would generally be experienced under ‘normal’ academic settings (i.e. writing for take home assignments). Despite these potential limitations, research performed by Fox (2004) has shown that the CAEL test is in fact an effective evaluation tool in “identifying EAP support classes and projecting the learning trajectory” (Fox, 2004, p. 459). Additionally, the writing portion of the CAEL assessment is viewed as the most critical component of academic English language proficiency, and attempts to reflect the “attitudes, beliefs, values and understandings of what it means to use language for academic purposes” (Fox, 2002, p. 22).

In previous research involving lexical bundles and the academic register, Levy (2003) used two corpora of student and professional writing to compare the function and distribution of lexical bundles. While the professional corpus came from published academic materials, the student corpus was composed of test taker’s writing sourced from student placement tests written by freshman or undergraduate transfer students which were graded according to perceived proficiency. In this way, the student corpus used by Levy (2003) is comparable to the two CAEL corpora used in the present study since they are both topic-specific time-constrained tests written by test takers aiming to enter into
English speaking institutions. Since most of the individuals taking the CAEL Assessment do so as part of the process of applying for admission to college or university programs, it can be argued that the writing produced by CAEL test takers is an adequate measure of academic writing.
2 Corpus Linguistics and Second Language Acquisition

Over the years corpus linguistics has grown into an important way in which to investigate language use. In this chapter, a brief history of corpus linguistics is provided along with the ways in which this type of empirical approach contrasts with other methods of linguistic study. Since corpus studies come in various forms, different approaches to the creation of corpora, and the types of corpora commonly used will also be discussed.

2.1 Rationalism Vs. Empiricism

Linguistic study can, in general, be seen to fall under two broad categories: empirical investigations, and rationalist introspection (McEnery & Wilson, 2002). Under the rationalist approach, internalized knowledge (such as the familiarity with one's native language) generally serves as the basis for any claims that are made on the language under study. The rationalist approach to linguistic inquiry uses introspection and 'artificial behavioural data' as the main data on which linguistic claims are based. This artificial behavioural data can come from a variety of data collection techniques, with one of the most popular being elicitation activities where researchers attempt to elicit language from the speaker/writer through a variety of methods.

Rationalist theories are based on the development of a theory of mind, in the case of linguistics, and have as a fundamental goal cognitive plausibility. The aim is to develop a theory of language that not only emulates the external effects of human language processing, but actively
seeks to make the claim that it represents how the processing is actually undertaken (McEnery & Wilson, 2002, p. 5)

Conversely, the empirical approach contrasts rationalism by using actual evidence of language in use as the primary source from which conclusions and judgments about a language are drawn. This actual evidence of language in use, often comes in the form of corpora which can provide a large resource of ‘real’ language for researchers to study. Each of these two approaches has its champions who have provided important advancements to linguistic study by implementing their preferred method of investigation. On the rationalist side, we can refer to the work of, among others, Noam Chomsky whose concept of generative grammar has had a long lasting impact in this field and has helped advance our understanding of the underlying systems governing language production/processing. From the empirical camp, Michael Halliday is perhaps one the best known examples of an empirical researcher, and can be seen as relying more on actual examples of language use for the basis of many of his claims and subsequent development of the well-know lexicogrammatical approach: Systemic Functional Linguistics.

Despite the contributions many rationalist linguists have made, generative grammar and the I-Language Vs. E-Language divide for example, not all areas of linguistic study are well suited to this method of enquiry. For example, Hunston, in her 2002 book on corpus linguistics believes that native speaker intuition is poorly suited to at least four areas of study: judgments about collocations, judgments about frequency, semantic prosody and pragmatic meaning and details of phraseology (Hunston, 2002, p.
Although a native speaker has experience of very much more language than is contained in even the largest corpus, much of that experience remains hidden from introspection" (Hunston, 2002, p.20). Furthermore, Stubbs, in his own work on corpus linguistics claims that “for many areas of language use, introspection provides no interesting evidence” (Stubbs, 2007, p. 129). The empiricist approach, and by extension corpus linguistics\(^2\), can be seen as an empirical method that attempts to provide answers to the areas of enquiry which rationalism cannot easily reveal. As opposed to relying on introspection and ‘artificial behavioural data’, empirical investigations instead, rely heavily on naturally occurring discourse. This discourse is usually collected into some kind of corpus where it can be stored, and subsequently analyzed either qualitatively or quantitatively to provide a description of the language.

While the two aforementioned approaches offer distinct benefits, for the reasons mentioned above, the study of formulaic language and lexical bundles using a frequency-based approach is better suited to empiricism. This is especially true when dealing with the formulaic language produced by non-native speakers since the language abilities of L2 speakers can be seen as lying beyond the scope of native speaker introspection.

As previously mentioned, corpus linguistics is one of the most widely used empirical approaches to linguistic study. It is therefore important to discuss what corpus linguistics is, and how it has evolved into its present form.

\(^2\) Leech (1991) lists four key features of corpus linguistics including “focus on a more empiricist, rather than rationalist view of scientific inquiry. (Leech, 1991, p. 107)
2.2 Corpus linguistics

Although corpus linguistics has existed in one form or another long before the advent of computers and the ability to store large amounts of text within them, today, the use of computers is inextricably linked with corpus linguistics. The roots of corpus linguistics stretch back through decades of linguistic research, even though these methods may not have been labeled as such at the time: “While not identifying themselves with the term corpus linguistics, field linguists, e.g. Boas (1940), and later linguists of the structuralist tradition all used a basic methodology that we can undoubtedly call corpus-based” (McEnery & Wilson, 2002, p. 3). Under the structuralist approach, discourse was collected and stored (into what would today be called corpora) and subsequently analyzed in an attempt to classify linguistic elements present within this collected body of texts. Although the collection and investigation of actual language use has existed for an extended period of time, a major problem faced by early corpus linguists was the fact that the storage and analysis of this data was a necessarily messy ordeal. Lacking the use of computers and corpus analysis software, corpus studies were labour intensive projects that at times needed to employ huge workforces to sift through many of the larger sized corpora that had been collected. Kaeding (1897), a German linguist, is cited by both McEnery and Wilson (2002) and Kennedy (1991) as employing at least 5,000 people to sort through his corpus of nearly 11 million words. Utilizing a work force of this size no doubt created many problems and affected the quality of analysis and any conclusions that were made from this analysis. Due in part to the unreliable nature of this and other early corpus linguistic work, many early corpus studies were seen as extremely time consuming, error prone and somewhat unreliable forays into linguistic inquiry. These
criticisms are summed up in Abercrombie's 1966 definition of the 'pseudo-procedure'. Although Abercrombie does not explicitly state that corpus linguistics is the subject of his critique, both McEnery and Wilson (2002), along with Leech (1991) see corpus linguistics as the target of Abercrombie's criticisms in the chapter “Pseudo-procedures in Linguistics”. For example, Abercrombie states that:

pseudo-procedures are of two sorts: in some, the alleged procedure is quite literally impossible; in others, although it may not be literally impossible, it would be so arduous and time-consuming as a way of conducting an investigation that no one in their senses would ever set out to use it (Abercrombie, 1966, p. 114-115).

While this can be seen as a somewhat accurate assessment of corpus linguistics at the time Abercrombie's book was published, these early criticisms of corpus linguistics need no longer apply, as computers largely handle the 'arduous and time-consuming' tasks Abercrombie wrote of. With the development and advancement of the computer age, these problems were largely eliminated, and today the use of computers has allowed corpus linguistic work to move beyond the level of a 'pseudo-procedure' into a well-respected and valuable method of linguistic study. While much of the corpus work done before computers relied on the storage and analysis of large collections of slips of paper on which text was written (Svartvik, 1991, p. 7), researchers have moved away from this manner of text collection, with computerized storage taking the lead. “By organizing huge masses of data, technology can make visible patterns which lie outside unaided
human perception and which no amount of introspection or manual analysis could
discover” (Stubbs, 2007, p. 131). An example of this ability to identify “patterns which
lie outside unaided perception” is found in the corpus work of Tao (2001). In this study,
Tao used a single written corpus, along with three spoken corpora to investigate the use
of remember. Contrary to previous introspective work by Van Valin and Wilkins (1993)
which stated that remember appears with “third-person full NP” and is used by “a third
person reporting past events”, Tao comes to the conclusion that “remember utterances are
not about a third-person event; they are instead primarily concerned with the speaker and
the addressee present in the interactional setting” (Tao, 2001, p. 122). The use of corpus
data in Tao’s study shows the potential benefits an empirical approach may have in
comparison to introspection when it comes to linguistic study, since the conclusion come
to by Tao is not in line with the original introspective judgments provided by Van Valin
and Wilkins. An additional example that shows the value of an empirical approach
comes from a conference on problems in linguistic analysis in English where Noam
Chomsky provides an introspective judgment concerning the use of perform:

Chomsky: The verb perform cannot be used with mass word objects: one can
perform a task but one cannot perform labour.

Hatcher: How do you know, if you don’t use a corpus and have not studied the
verb perform?

Chomsky: How do I know? Because I am a native speaker of the English
language.

(Hill, 1962, p. 29 as cited in McEnery & Wilson, 2001, p. 11)
Despite the confidence with which this statement was made, this can also be used to show the potential importance of an empirically based approach. As McEnery and Wilson (2001) point out, a search of the BNC corpus reveals that one can indeed perform 'mass word objects', as 'perform magic' is a completely acceptable construction (McEnery & Wilson, 2001, p. 11).

As computer applications began to spread, linguists took note and made use of their abilities to further linguistic study. The first computerized corpus, the *Brown Corpus of English (BCE)*, was developed in the 1960s by Kucera and Francis at Brown University in the United States. This computerized corpus was an important step in corpus linguistics since it helped to reduce/eliminate many of the aforementioned problems associated with non-computerized corpus studies and allowed researchers to have a much more stable and reliable way of storing large amounts of data. Although the sophisticated computer software necessary to analyze the data linguists were collecting began to appear soon after the original BCE was created, the storing of large corpora on computers was a huge first step toward modern day corpus linguistics.

Francis, due to his involvement in the creation of the first computerized corpus, was one of the pioneers of a computer based approach to corpus studies, and it may be most appropriate to reference his own definition of corpus: “a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis (Francis, 1982, p.7, as cited in Francis 1991, p. 17). As Hunston points out in her own work on corpus linguistics however, “It might be more proper to say that corpora are a way of collecting and storing data, and that it is the corpus access
programs — presenting concordance lines and calculating frequencies — that are the tools” (Hunston, 2002, p. 20). In other words, a corpus is of little use without an ability to adequately analyze it. It is important to note how closely entwined the use of computers and corpus linguistics have become, indeed much of the work done today with the larger corpora we now have would be impossible without the use of computers. The link between computers and corpus linguistics is made explicit by Biber, Conrad and Reppen who list “extensive use of computers for analysis” as one of the four essential characteristics of a corpus-based research (Biber, Conrad & Reppen, 1998, p. 4).³

Today, concordancing programs are necessary to search through the increasingly large sized corpora modern researchers continue to collect. While numerous general-purpose programs exist which can be used to search through these corpora, other researchers have found the need for more specialized concordancing programs that allow them to search for specific features that are of interest to their studies (De Cock, Granger, Leech, McEnery 1998; Granger & Rayson, 1998; Cortes 2004; Levy 2003). With multiple programs available for corpus analysis it can be somewhat confusing for researchers to chose a reliable program to use in their research. As a potential aid to this issue, Ari (2006) reviewed and evaluated three of the most popular concordancing programs available, coming to the conclusion that WordSmith Tools (in version 3 at the

---

³ In 1991, Leech thought it appropriate to modify corpus linguistics to include computer in its title (“Computer Corpus Linguistics”), however today, the “computer” portion of this label seems to be implied whenever corpus studies are mentioned.
time of the review) is “among the software reviewed here... the most efficient in its search for lexical bundles” (Ari, 2006, p. 39).  

WordSmith Tools was originally released in 1996, and as of late 2010 is now in its 5th iteration. Over this period of time, the program has grown into a well-accepted and widely used program that has been put to use by numerous researchers looking at a variety of linguistic features in the corpora they use in their studies (Chen 2010, Rachel 2009; Gabrielatos, 2007; Hyland, 2008, etc).

Through the use of computer aided corpus analysis, many interesting and important discoveries have been made in a variety of areas within linguistics and applied language studies. In a 2010 study of the lexical bundles in Electrical Engineering introductory textbooks and English for Specific Purposes (ESP) textbooks, it was revealed that ESP textbooks misrepresent target language use (Chen 2010). In another corpus-based study of academic writing, it was shown that both proficient college writers and professional writers used discourse structuring lexical bundles with greater frequency than non-proficient college writers (Levy, 2003). Corpus based studies investigating the discourse produced by non-native English writers have also proven to be a valuable information resource. Virtanen (1998), in a study of the use of direct questions in argumentative student writing found that non-native writers tended to use more direct questions in their writing than native English writers. Additionally, using an automatic profiling technique, Granger and Rayson (1998) found that non-native English writers

---

4 Although his research was published in 2006, and new versions/programs have surfaced since his initial evaluation, WordSmith Tools is still one of the most popular and widely accepted concordance programs available for purchase.

5 For a more complete list of publications using WordSmith Tools, please refer to the WordSmith website at www.lexically.net.
tended to create written discourse that was much more speech-like in nature, and that French learners in their study “display practically none of the features typical of academic writing and most of those typical of speech” (Granger & Rayson, 1998, p. 129).

While corpora are most often of a finite size, including those mentioned in the preceding paragraphs, it is important to point out one notable exception, the monitor corpus. A monitor corpus is a continually growing (and therefore non-finite) body of texts that can be particularly useful for lexicographic researchers. Since new text is constantly being added to monitor corpora, researchers can scan the accumulated data for changing meanings of pre-existing words, while also looking for the occurrence of new lexical items which may eventually lead to inclusion in updated dictionaries (McEnery & Wilson, 2001, p. 30). An example of a monitor corpus is the Collins Birmingham University Language Database (COBUILD). As of late 2010, this corpus contained approximately 2.5 billion words, with new text being added on a monthly basis (MYCOBUILD, 2010).

While there are obvious benefits to using corpora for language studies, empirical approaches should not be seen as the only approach to linguistic study, or the sole method of achieving analytical success. Both the rationalist and empirical approaches offer distinct benefits and, when possible, should work in tandem to provide a more accurate assessment and description of the object of study. As Charles Fillmore stated during the 1991 Nobel Symposium on Trends in Corpus Linguistics:

I have two main observations to make. The first is that I don’t think there can be any corpora, however large, that contain information
about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way. My conclusion is that the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body (Fillmore, 1991, p. 35)

While these words were written almost 20 years ago, their meaning still holds true today. Corpus linguistics presents researchers with a valuable set of tools that can be used to accomplish many important tasks. Despite the advantages of computer aided corpus linguistics however, it is still necessary for individual researchers to bring their own knowledge, intuition and introspection into this analysis so that a more complete picture of the object of study can be found.

### 2.3 Corpora in Second Language Acquisition (SLA)

In addition to the four areas of linguistic research that Hunston (2002) identified as lying beyond native speaker introspection, “Chomsky himself (1964) cautioned that his rejection of performance data as a source of evidence was inappropriate for language acquisition studies” (McEnery & Wilson, 2001, p. 13). Keeping this in mind, the use of corpora to study the language used by non-native English speakers seems well suited. Indeed, as Mair (1991) has stated “corpus data are useful because they often are just better data than those derived from introspection or elicitation, and for the study of certain problems they are indeed the only available data” (Mair, 1991, p.98). This can be
particularly true for ESL/EFL speech/writing, since it is near impossible for native
speakers to offer adequate introspective observations concerning the language non-native
speakers/writers produce.

While many English language corpora have been compiled in order to analyze
linguistic differences between; spoken and written language, various university
disciplines, and regional dialects, it was not until the early 1990s that corpora of learner
English began to appear (Granger, 1998, p. 4). One of the best known of these learner
corpora is the International Corpus of Learner English (ICLE). The ICLE is composed
of essays written by advanced EFL students in their 3rd and 4th years of undergraduate
study. The writing samples included in the ICLE come from students with a wide variety
of L1s, and it is therefore possible to use the corpus as a way of investigating the impact a
writers’ L1 has had on their English language writing. An investigation into the impact
of writers’ L1 on their English language writing using the ICLE is seen in Altenberg and
Tapper’s 1998 study “The use of adverbial connectors in advanced Swedish learners’
written English”. In this study a sub-corpus form the ICLE is used to compare Swedish
students’ L2 English writing with native English writers. Through this investigation, the
authors conclude: “in contrast to many EFL learners of other language backgrounds, the
Swedish learners tend to underuse conjuncts” (Altenberg & Tapper, 1998, p. 92).
Without the use of corpus data to verify this claim, it would be very difficult for anyone
to come to this conclusion. Additionally, Virtanen (1998) used the ICLE to compare the
use of direct questions in argumentative writing by non-native writers to those used by
native writers in the Louvain Corpus of Native English Essays (LOCNESS). Based on
the analysis of these two corpora, Virtanen came to the conclusion that non-native
speakers tend to overuse direct questions in their writing when compared to their native English speaking counterparts (Virtanen, 1998, p. 105).

Despite the increasing number of corpus studies focusing on ESL/EFL writing, very few studies have provided detailed reports on the differences in use of formulaic sequences among non-native speakers of differing levels of proficiency. Instead, previous studies focusing on formulaic sequences within non-native speaker populations have chosen to treat non-native speakers as a uniform group, being identified with broad strokes as either ‘low/high-level’, ‘advanced’ or simply as 'non-native'. The analysis done is therefore most often presented in comparison to native-speaker discourse in an attempt to show differences between native and non-native writers (the two studies mentioned above also show this tendency). While this approach is beneficial in that it can shed light on differences between these two groups, and potentially show problem areas L2 speakers/writers face as it relates to their use of formulaic sequences, it does little to show the progression L2 speakers/writers make on their path toward native-like language abilities. Similarly, by not providing distinctions between L2 speakers/writers of varying English language abilities, we are given little information about how native-speakers judge this aspect L2 speakers/writers language.

2.4 General Vs. Specialized Corpora

Although there are many types of corpora that can be used to achieve very different goals, it is beneficial to make a broad distinction between general and specialized corpora. As Hunston has defined it, a general corpus “will include as wide of a spread of texts as possible” and additionally will contain “texts of many types” (Hunston, 2002, p. 14-15). These corpora attempt to give a general picture of the
language under study. Examples of general corpora abound, including the first computerized corpus mentioned above, the Brown corpus. The Brown corpus is classified as a general corpus since it contains texts of various sources, including novels, religious texts, newspaper articles, etc. For similar reasons, the COBUILD monitor corpus would also be classified as a general corpus.

While general corpora are obviously an important and valuable source of linguistic information for researchers, they do not always represent the most appropriate approach for all topics of study. For example, when researchers have specific goals they are trying to achieve, specific features they are looking to analyze, or a subsection/dialect of a language, a general corpus will most likely not provide an adequate resource for this research.

As stated, much of the prior corpus research done on L2 lexical bundle use has chosen to group all L2 speakers together instead of differentiating between levels of language proficiency. This approach has helped researchers identify numerous differences between native and non-native speakers, but has done little to explain the differences in formulaic language use among L2 writers of differing proficiency levels. Access to the CAEL test archives has allowed the current research to move away from this established model, and instead compare the lexical bundles used by L2 speakers of differing language abilities. Using the CAEL corpus data also has allowed for the creation of a highly specialized corpus that can be used to answer the specific research questions mentioned above.
2.5 Representativeness of Corpora

If we look back at Francis’s 1982 definition of corpus linguistics, we can see the importance placed on corpus representativeness “a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis (Francis, 1982, p.7, as cited in Francis 1991, p. 17). While corpora of random texts can easily be compiled by any researcher, this type of random sampling will not be able to answer all questions related to language use. As with the well-known “garbage in, garbage out” philosophy, any corpus analysis can only be as good as the corpus being analyzed. If the corpus is not an accurate representation of the language the researchers are attempting to study, any analysis of the corpus will no doubt become an exercise in futility. An important first step toward the goal of creating a representative sample from which to compile any corpus is to accurately define the object of study. While it is easy for researchers to very broadly define the area of language, or language subset of their study, more specific parameters are needed in order to ensure that sampled texts included in the corpus will provide a true representation of the object of study. For example, while researchers may say that their corpus is a representation of 'songwriter lyrics', this would include an incredibly wide range of languages, genres and time periods that would most likely not fulfill the true intentions of the author of said research. In order to provide a more valuable corpus and subsequent analysis, the ‘sampling frame’ for the corpora should be more strictly defined (McEnery & Wilson, 2002). In the previous example of songwriter lyrics, this may have the effect of defining the sampling frame as only those songs in one particular language, genre and year as admissible into the corpus. This will result in a more strictly defined object of study and a more
representative corpus that will be better suited to answering the researchers' specific interests.

Additionally, corpus size is also important to the representativeness of any corpus. While this is by no means the only determiner of the representativeness of a corpora, it should be noted that a corpus is only a sample of the language subset it is attempting to reflect, and a larger sample size will no doubt create a more accurate representation of that subset of the language under study (as long as the samples included in the corpus maintain adherence to the predefined sampling frame). While larger corpora are generally believed to provide a more accurate representation of the language subset under study, this is by no means an absolute judgment. Depending on the language feature the researcher is attempting to analyze, there may only be a limited amount of text available to be included in the corpus. Although corpus size is important to the representativeness of a corpus, anyone compiling a corpus must make sure to follow their pre-defined sampling frame to ensure only appropriate texts are included. While past analysis of lexical bundles has often involved the use of large-scale corpora consisting of hundreds of thousands, and even millions of words, valuable analysis has also been obtained using much smaller corpora Howarth (1996), Chen (2008). In addition to these two examples, Partington (1998) cites Mparutsa, Love & Morrison (1991) and Ma (1993) as additional examples of success using small-scale corpora (Partington 1998, p. 5). A detailed explanation of the techniques used, and approaches taken in a selection of these studies will be discussed later on.

Despite the effort and precautions needed prior to and during the creation of any corpus, as Hunston states, “the real question as regards representativeness is how the
balance of a corpus should be taken into account when interpreting data from that corpus” (Hunston, 2002, p. 30). When all is said and done, any corpus analysis is only a representation of what was contained within the corpus itself, and caution must be taken in generalizing anything that is found within the corpus to populations that are not represented by it.
3 Formulaic Language

Formulaic sequences are *multiword strings of language that are stored fully formed in the minds of speakers/writers*. It has been argued that these sequences hold a similar place in our mental lexicon to more ‘standard units’, i.e. individual words (Peters, 1983). Since these sequences may be stored in our minds in a similar fashion to other lexical entries, formulaic language may play a significant role in our language production/processing. The use of formulaic sequences allows the speaker/writer to insert fully formed chunks of language into their discourse, thereby allowing the speaker/writer to bypass generative grammar rules in favour of a quicker form of language production. This version of language composition can be viewed as allowing the user to create larger stretches of discourse while at the same time ‘pulling’ fewer blocks of language from the mental lexicon. One of the main advantages of this form of language composition/processing is that it can allow for quick and accurate context specific discourse to be created by the communicator (Wray, 2002). Over the past few decades, an increasing number of researchers have made formulaic language the focus of their studies, and a growing recognition of the importance of formulaic language in native-like composition/processing seems to be taking hold.

In *The Units of Language Acquisition* (Peters, 1983), an analysis of children’s first language acquisition leads the author to conclude that the first units of language acquisition are often not simply single words, as had been often assumed, but are instead represented by multi-word units. Furthermore, the authors’ research leads her to the belief that “In terms of storage and use...there is no difference between such long units and units that happen to be minimal: to the language learner they are all units, and are
stored in the lexicon and retrieved as such” (Peters, 1983, p. 89). The conclusions reached by Peters present a potentially significant benefit to speakers/writers. The ability to store and retrieve longer word-strings would theoretically allow users to produce larger strings of language more quickly than would be possible if they were composing language word-by-word. This may also be similar to a technique musicians use of chunking together musical scales and other strings of notes: “In discussing the processing of semantically complex items, Jackendoff draws an analogy with chunking in music (1988: 125): ‘Any musician can attest that one of the tricks to playing fast is to make larger and larger passages form simplex units from the point of view of awareness – to ‘chunk’ the input and output” (Cowie, 1998, p. 31). Just as a construction project can be facilitated through the use of a relatively small number of semi-preconstructed pieces (as opposed to using a high number of smaller individual parts) language composition may also be facilitated by using semi-preconstructed sequences. This ability to pull longer word strings from memory also seems to be supported by research into language processing, as “these researchers feel that the storage capacity of memory is vast, but that the speed for processing those memories is not (Crick 1979: 219), so that we must learn short cuts for making efficient use of this processing time (Nattinger & DeCarrico, 1992, p. 31). By storing multi-word units in our memory, it is plausible that we are able to make more efficient use of our relatively limited processing capabilities by using prefabricated chunks of language when we compose our discourse.

Not only do these sequences hold potential benefits for those using them to communicate with others, the use of formulaic language may also hold benefits for the hearer/reader as has been demonstrated by Underwood, Schmitt, and Galpin (2004) who
used an eye movement study to look into the processing of formulaic sequences by both native and non-native readers. In this study, the researchers used an SMI Eyelink system to track the eye movements of participants as they read through a text containing both formulaic and non-formulaic sequences. The expectation was that this would provide some insight into how formulaic sequences are processed, a subject that the authors believe has been ‘inadequately researched’. Under the assumption that eye fixations during reading provide an indication as to what processes are occurring in readers minds, the researchers provide mean data on eye fixations (both frequency and duration) for both groups of readers. Analysis of their findings leads the researchers to conclude that: although native speakers generally use shorter and fewer fixations than non-native speakers, both groups see a significant decrease in the mean number of fixations when looking at terminal words in formulaic sequences (compared to these same words when they appeared outside of formulaic sequences). In addition, the data also shows a significant drop in the mean duration of fixations on terminal words in formulaic sequences among native speakers (again, compared to terminal words in non-formulaic sequences). This study can be seen as a potential indication that the use of formulaic sequences can not only ease the processing load of the speaker/writer but also the hearer/reader. Coupled with the aforementioned judgment that novel sequences are more likely to be misinterpreted, the findings found in Underwood et al.’s 2004 study, and Tremblay (2009) show the important role formulaic language plays in language comprehension.

While the use of formulaic language is very likely a popular method native speakers use in order to produce quick/accurate discourse, it may also be one of the keys
to a speaker/writers' ability to produce native-like language. As Pawley and Syder (1983) have shown, “only a small proportion of the total set of grammatical sentences are native like in form” (Pawley & Syder, 1983, p. 215). By incorporating formulaic sequences into their language repertoires, non-native speakers may be able to better avoid the common problem of producing grammatically correct, yet non-native like utterances.

3.1 Forms of Formulaic Language

Formulaic language can take on many forms from very short sentence fragments to longer full-length phrases. Additionally, these sequences are not always continuous, and may hold variable slots within the formula that allow for substitution/variation (i.e. It is [very] nice to meet you). As previously mentioned, there exists a wide variety of terms which can be seen as falling under the general heading of formulaic language. These terms include, but are not limited to: collocations, idioms, metaphors, compounds, and proverbs (Allerton, Nesselhauf, & Skandera, 2004). In Allerton et al. 2004 book on the subject of phraseological units, each of the first five chapters is devoted to defining and explaining these terms. While distinctions are made between these labels, the chapters also show that a great deal of overlap exists among many of these terms and that the lines between terms are often blurred. Even within individual terms, different researchers have adopted these labels to refer to varying notions, something that can create potential problems and misunderstandings.

Even though Allerton et al.'s 2004 attempts to distinguish between the aforementioned terms are beneficial to the reader since they can provide some ground work for the subsequent chapters in their book, the classification scheme used by Nattinger and DeCarrico (1992) may be seen as a more appropriate model for
distinguishing between types of formulaic language. In their 1992 book on lexical 
phrases, Nattinger and DeCarrico (1992) use four structural criteria to categorize lexical 
phrases:

1. Length and grammatical status: Since formulaic language varies in 
terms of both length and grammatical correctness, the first criteria used 
in identifying these sequences is to determine both the length of the 
sequence in question, and the grammaticality of said sequence.

2. Canonical/Non-canonical: Since many formulaic sequences contain 
forms that are not of a standard variety (by and large, once and for all), 
it can be valuable to distinguish between these two possibilities.

3. Fixed/Variable: While many formulaic sequences are totally fixed, in 
other words no alteration is possible without losing the original 
meaning (in a nutshell cannot be changed to in a grape, and hold your 
horses cannot be changed to your horses must be held), other sequences 
do hold variable slots that allow for varying degrees of substitution (to 
make a [very] long story short).

4. Continuous/Discontinuous: This criteria is used to distinguish between 
sequences that appear in unbroken blocks, and those that may be 
separated by other lexical items (not only X, but also Y).

Seeing each of these criteria as existing on a continuum as opposed to possessing 
absolute values, Nattinger and DeCarrico (1992) use the above criteria to identify 4 types
of lexical phrases: polywords (for the most part, by the way), institutionalized expressions (nice meeting you, have a nice day), phrasal constraints (to [wrap] this up, a [long] time ago), and sentence builders.
4 Lexical Bundles

Although formulaic sequences can be viewed as a broad category covering repeated language chunks found in many different types of discourse, the term “lexical bundle” is most often confined to the description of formulaic sequences found within written discourse using a frequency-based approach. Douglas Biber, one of the leaders in the study of formulaic sequences defines lexical bundles “simply as the most frequently recurring sequences of words” (Biber & Barbieri, 2007, p. 264). While the discussion of formulaic sequences in the previous chapter looked at general characteristics of a wide variety of formulaic language, lexical bundles have been able to create a distinct category of their own within this area of research.

Since lexical bundles are statistically defined multi-word strings, their identification generally relies on computer aided corpus analysis using concordancing programs such as WordSmith Tools. One of the main advantages of using this form of identification is that it removes any personal bias a researcher may have in identifying a sequence as formulaic while also offering the added benefit of creating a repeatable/verifiable, and therefore more potentially more viable source of linguistic information. While not all corpora used in lexical bundle research are available to the public, those studies which do use publicly available corpora open up their research to others who can use their own concordancing programs to verify their results.

4.1 Lexical Bundles and Psycholinguistic Validity

Lexical bundles represent an interesting and unique area within the study of phraseology and formulaic language. While most formulaic language has a basis in
psycholinguistic validity, since lexical bundles are identified solely on the basis of frequency of occurrence, their use does not necessarily mean that they are stored fully formed in the minds of those producing these structures. Wray (2002) addresses the issue of frequency in relation to formulaicity with the following statement: “if a string is required regularly, it is likely to be stored whole for easier access […] but it does not have to be” (Wray, 2002, p. 31). This statement puts lexical bundles into somewhat of a middle ground position since they may, or may not, be stored fully formed in the mental lexicon. Schmitt et al. (2004) attempted to test whether lexical bundles are actually stored in a prefabricated manner within the minds of speakers by using an oral-response test. In this test, participants first listened to, and then attempted to recite a short narrative that had several corpus derived from recurrent word-clusters embedded into the narrative. Once the participants’ attempts to re-create the narrative were recorded, these recordings were subsequently analyzed with the aim of identifying which of the recurrent word-strings embedded into the story were also produced by the participants during their attempts to recreate said narrative. Lexical bundles embedded into the narrative, which were also produced by participants during their re-telling of the story, were seen as an indication of formulaic status. Testing both native and non-native English speakers, the researchers found that, although on average native speakers were found to reproduce the narrative more accurately that their non-native speaking counterparts, frequency based identification of recurrent word clusters is not sufficient to indicate formulaic status.

While the results of Schmitt et al.’s 2004 test showed no link between prefabricated status and frequency of occurrence, these results have been disputed by Chen (2008), who believes that “the researchers mixed the modes in the research design,
which has negatively affected the effectiveness of this study" (Chen, 2008, p. 49). The results of Schmitt et al.'s study are obviously debatable, however seeing as their study represents one of the few attempts thus far to test the psycholinguistic validity of lexical bundles, it should not be ignored. Even though lexical bundles are identified on the basis that they are frequently occurring, caution should be taken in labeling these word strings as holistically stored units. Lexical bundles do however provide a valuable starting point in the analysis of any corpus. Since lexical bundles allow researchers to look at how words are used in conjunction with other words, they can still be seen as a profitable area of linguistic study that may or may not be holistically stored and produced by speakers/writers.

4.2 Identification of Lexical Bundles

The identification of formulaic language has taken on various forms, with researchers deciding to focus on different aspects for the identification of these sequences (structure, function, transparency, etc). These methods include, but are not limited to: phonologic coherence (Bybee, 2001), psycholinguistic validity (Schmitt, 2004), and corpus-based frequency (Biber, 2004). While each of these foci attempts to identify formulaic language using different tools, they all contribute to many of the same goals, primarily a better understanding of the use, acquisition and processing of formulaic sequences. Of the aforementioned research approaches, the only one used for the identification of lexical bundles is that of corpus studies employing a frequency-based approach (Altenberg 1998, Biber 2004, Cortes 2004, Hyland 2008). Under this method, corpora of various sizes have been analyzed using sophisticated computer software to identify the most commonly occurring recurrent word clusters. Used by Biber, Conrad,
and Cortes (2004), Chen (2008), Cortes (2008) and many others, corpus studies implementing frequency based counts have proven invaluable in identifying formulaic sequences and lexical bundles in a wide variety of corpora both big and small.

Since lexical bundles are simply defined as ‘the most frequently recurring sequences of words’ their length generally varies from 2-6 words in length. While any string of at least two words can be accepted as a lexical bundle, longer word strings are generally less frequently occurring than shorter strings. Upon a review of the literature dealing with lexical bundles, we can see that the most common lexical bundle length being searched for has generally been set at 4-words with a frequency cut-off of 20 occurrences per million words (Hyland (2008), Cortes (2008), Chen (2008)). While these figures can vary depending upon the size and specialization of the corpora being analyzed, the frequency cut-off point in these studies does represent a good starting point for any lexical bundle identification project.

In addition to a frequency cut-off point, the notion of range must also be addressed in any study of this type. Range deals with the spread of formulaic sequences within a corpus, i.e. how many texts within the corpus must contain a formulaic sequence before it qualifies as a lexical bundle. As with the aforementioned research that suggests using 4-word lexical bundles with a cut-off point of 20 occurrences per million words, a range of 5 texts also seems to be a well-accepted criterion for lexical bundle identification (Cortes 2004, 2006; Biber 2004; Chen 2008). Again, while these figures represent a good starting point for any research into this aspect of language, these figures may need to be altered depending on the size and level of specialization of the corpus being studied.
4.3 **Pragmatic Functions of Lexical Bundles**

Lexical bundles are not only important to the creation of discourse due to the fact that they are frequently occurring, they also serve specific functions that aid in the creation of meaning within a text. Following Biber et al. (2004), we can identify 3 major functions that lexical bundles play in structuring discourse:

1) Stance expressions

2) Discourse organizers

3) Referential expressions

Each of these types of lexical bundles serve a specific function within discourse, and will therefore be discussed in the following sections.

**4.3.1 Stance Expressions**

Stance expressions are generally used to “express attitudes or assessments of certainty that frame some other proposition” (Biber et al, 2004, p. 384). These structures also give the reader a way of interpreting the information that follows. The category of stance expressions branches off into two subcategories which each offer their own unique functions; *epistemic*, and *attitudinal/modality*.

The first subcategory of stance bundles, epistemic bundles, can be further divided into personal and impersonal epistemic stance bundles. Each of these two categories can be used by the speaker/writer to express opinions, including degrees of certainty. While personal epistemic stance bundles are most often used to express degrees of uncertainty (*I*
think it is), impersonal stance bundles are more often used to express certainty (are more likely to).

Attitudinal/Modality stance bundles serve several functions, including the expression of desire (I don’t want to), obligation/directive (you have to do), and ability (I want you to be able to).

Of the four registers (classroom teaching, textbooks, conversation, academic prose) Biber et al. (2004) analyzed for lexical bundle usage, stance bundles were most common in the oral registers: classroom teaching, and conversation (Biber et al., 2004).

4.3.2 Discourse Organizers

As their name suggests, discourse-organizing bundles serve the purpose of providing a structure and organization to the discourse they are used in conjunction with. These bundles can be furthered divided into: topic introduction/focus bundles and topic elaboration/clarification.

Topic introduction/focus bundles often “occur with both first and second person pronouns” (Biber et al., 2004, p. 392). Used in this manner, Biber et al. identify a number of discourse organizing bundles in the classroom teaching register:

If you look at the answers that are given, there’s only two answers that have one big M… (classroom teaching, Biber et al., 2004, p. 393)

What do you think the text is trying to tell us when they call attention that often conflict doesn’t appear suddenly? (classroom teaching, Biber et al., 2004, p. 393).
Topic elaboration/clarification bundles are used to expand on a topic. These bundles are common among both textbook writers and classroom teachers as seen in the following examples:

When you come to class next time—and I'm gonna look at grammar you know I mean I expect you to have things spelled relatively correctly… (classroom teaching, Biber et al., 2004, p. 393).

We know that if the project is in the same line of business as the firm’s other projects [...] then stand alone risk translates into high corporate risk [...] On the other hand, if the project is not in the same line of business, then it is possible that the correlation may be low (textbook, Biber et al., 2004, p. 393).

While these bundles have been observed in both classroom teaching and conversation, discourse-organizing bundles appear most commonly in classroom teaching (Biber et al., 2004)

4.3.3 Referential Bundles

“Referential bundles generally identify an entity or single out some particular attribute of an entity as especially important” (Biber et al., 2004, p. 393). As with the previous two categories of lexical bundles, referential bundles can be further divided into four sub-categories: identification/focus, imprecision indicators, specification of attributes, and time/place/text reference.
Identification/focus bundles are often used to draw attention to what immediately follows the bundle and may hold "a discourse organizing function" (Biber et al., 2004, p. 394), such as for introduction or summarization, as shown in the following examples:

*One of the things* they stress in parenting is to be consistent and particularly with parents um some parents are inconsistent between siblings. Uh fathers are notorious for letting their little darling girls get away with what they swat the boys about… (classroom teaching, Biber et al., 2004, p. 394)

Schizophrenia typically uh will mean that uh separation from reality uh it can mean uh uh you know extreme periods of euphoria and extreme periods of depression it can mean a lot of things—and that's *one of the* problems of schizophrenia. (classroom teaching, Biber et al., 2004, p. 394)

Imprecision bundles, on the other hand, are used for reference functions. These bundles can "either indicate that a specified reference is not necessarily exact, or to indicate that there are additional references of the same type that could be provided" (Biber et al., 2004, p. 394)

I think really we now have what about, six weeks left in class or something *like that.* (classroom teaching, Biber et al., 2004, p. 394)
Specification of attributes bundles are used to give attributes of the subject following the bundle. These bundles can provide various attributes to the head noun in question, including both concrete (size, weight etc) and abstract concepts.

These figures give an idea of the size of the ethnological community in Russia (textbook, Biber et al., 2004, p. 395)

Rather than reading textbooks and solving textbook problems, students must define and constantly refine the nature of the problem... (textbook, Biber et al., 2004, p. 395)

The last referential subcategory is the time/place/text-diexis bundle. These bundles are often “multi-functional, referring to a place, time, and/or text diexis, depending on the particular context” (Biber et al., 2004, p. 396).

So you have to record that, since the asset was sold at the end of the year (classroom teaching, Biber et al., 2004, p. 396)

She’s in that...uh..office down there...at the end of the hall (classroom, Biber et al., 2004, p. 396)

Uh I’m going to start actually with the end of the chapter (classroom teaching, Biber et al., 2004, p. 396)
Referential bundles appear most often in classroom teaching. These bundles are also apparent within textbooks and academic prose (Biber et al., 2004).
5 Methodology

Before discussing the methodology used to investigate the research questions, a restatement of these questions is appropriate:

1) *Is the frequency with which lexical bundles are used greater for L2 writers evaluated as being more advanced (greater perceived fluency)?*

2) *Do the functions of these lexical bundles differ according to perceived level of fluency, as judged by those evaluating the written response portion of the Canadian Academic English Language assessment tests? If so, in which ways?*

In order to address the above research questions, an empirical corpus based approach was used to investigate the differences in lexical bundle use between high-level and low-level non-native English test takers who wrote the CAEL test. The decision to use the CAEL test was based in part on the fact that the written portion of the test can be considered to come from a specialized academic register. While not all test takers were proficient in this register, it provides an interesting point of analysis since other studies looking at native speaker use of lexical bundles have also often chosen to focus on the academic written register. Additionally, since the academic written register presents problems for both native and non-native writers alike, this research may prove beneficial to our understanding of how non-native writers progress in their academic writing.

Recognizing that a highly specialized corpus was needed to answer the previously stated research questions, a strict sampling frame was also necessary. This sampling frame allowed for the inclusion of only one test version (i.e. only tests that dealt with the same subject were included in the corpora). Since formulaic sequences and lexical
bundles are often topic-specific, using only one test version allowed for the analysis of a confined set of lexical bundles used by students of differing levels of proficiency.

After deciding what writing would be included in the corpora, the next step was to decide how the corpora should be divided. While it would have been plausible to create nine separate corpora for each of the grading levels used by CAEL, for the present research, it was deemed more valuable to divide the CAEL tests into two broad bands, those scoring 40 and below (Low Level Corpus – LLC), and those scoring 50 and above (High Level Corpus – HLC). The LLC contains a total of 339 tests with 75,553 tokens appearing throughout the accumulated texts. In general, non-native English writers with a lower level of proficiency tended to write shorter written responses and therefore more tests of this type were needed to achieve the desired corpus size. A check of the differences in average written response length reveals a mean length (in number of words) of 223 for the LLC and 298 for the HLC. Because of this difference in mean length for the written responses of the LLC and HLC, the HLC contains a smaller number of tests than the LLC. The HLC therefore contains a total of 254 tests with 75,755 tokens appearing throughout the accumulated texts. A summary of the statistics for the two corpora is provided in the table below.

Table 5.1 Corpus Statistics

<table>
<thead>
<tr>
<th>Corpus</th>
<th># of files included</th>
<th>Tokens (running words)</th>
<th>Distinct Words</th>
<th>Mean # of Words/essay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower Level Corpus</td>
<td>339</td>
<td>75,553</td>
<td>3,382</td>
<td>223</td>
</tr>
<tr>
<td>Higher Level Corpus</td>
<td>254</td>
<td>75,755</td>
<td>3,964</td>
<td>298</td>
</tr>
</tbody>
</table>
As previously mentioned, the CAEL web site sets a minimum cut-off point of 50 for admission into (at least some) English speaking colleges and universities. Given this admission criterion, it was decided that dividing writers along these lines would provide a more suitable analysis of the differences in L2 student writing as it relates to lexical bundle use. Additionally, since the present study focuses on only one test version, it was possible to gain access to a more substantial amount of texts to be included in each of the two corpora than would have been possible if the corpora were divided so that one corpus was used for each of the nine grading bands.

While past analysis of collocations and lexical bundles has often involved the use of large-scale corpora consisting of hundreds of thousands, and even millions of words, valuable analysis has also been obtained by using much smaller corpora (Howarth 1996; Levy 2003; Chen 2010; Ringbom 1998; Hasselgren 2002; Lorenz 1998). Howarth’s 1996 study used two separate corpora to analyze the differences in collocation use between native and non-native English writers. The first corpus of native English writers’ discourse was actually created by combining two separate corpora of academic texts and textbooks to form a single “Native English Writing Corpus” (180,000 words in total). The second corpus used by Howarth was much smaller, amounting to only 25,000 words. This second corpus of non-native English writing was made up of ten essays written by non-native English writers from a variety of different language backgrounds. Once the two corpora involved in his study had been created, Howarth sorted through the data, attempting to identify all of the collocations appearing in each of his corpora. Choosing to forgo the use of minimum frequency of occurrence as a cut-off in the identification of collocations, Howarth identified 6,500 collocations in the native and
non-native corpora, with the vast majority coming from the native-writer corpus. Operating under the assumption that the collocations identified in the native English writer corpus were 'standard' forms, and therefore sequences non-native writers should be aiming to use in their own writing, Howarth was able to discover that the writing produced by non-native English writers included 'a much greater incidence of non-standard phraseology' (Howarth, 1996, p. 186).

Ringbom’s 1998 study investigated vocabulary frequencies in the writing produced by advanced learners of English coming from seven different western European learner corpora (French, Spanish, Finnish, Finland-Swedish, Swedish, Dutch, German), and compared their writing with a corpus of native English writing. Using a cross-linguistic approach, Ringbom (1998) took 7 western European learner corpora, ranging from 60,487 words to 129,549 words from the International Corpus of learner English (ICLE). In order to analyze the differences in writing produced by each of these groups of non-native writers with that produced by native English writers, Ringbom chose to compare the most frequently occurring words in each of the corpora. Adjusting all frequency figures to reflect occurrences per 10,000 words, Ringbom was able to discern that all learner groups in the study tended to overuse words that denote a non-finite and therefore vague interpretation. Examples of words that were more commonly occurring in each of the learner corpora, and also fit this category, included “people”, “things”, “more”, “all”, “some”, and “very”. Each of these words was found to be much more frequent in the learner corpora than the native English writer corpus. This finding was seen as providing evidence for the pre-existing view that advanced learner language is more vague than the language used by native English writers.
Lorenz (1998) presents an additional example of success using small-sized corpora. In Lorenz's 1998 study, he made use of four relatively small corpora (around 100,000 words each) to investigate how L1 German speaking learners of English made use of intensification adjectives in their writing. Comparing both German teenagers and German university students to British teenagers and British university students, Lorenz used computer aided corpus analysis to discover that the writing found in the German native-speaker corpora contained far more intensification adjectives (important, good, hard, difficult, easy, etc) than the native-English speaking British corpora. Furthermore, investigation of the corpora used in his study led to the statement that "the learners not only use more intensification, they also use it in places where it is semantically incompatible, communicatively unnecessary or syntactically undesirable" (Lorenz, 1998, p.64).

In line with some of these smaller scale corpus based studies, the present research also employs the use of two relatively small corpora, each composed of around 75,000 words. These two corpora consist of a total of 593 separate tests written between 2006 and 2010. Since the written response portion of the CAEL test is hand-written on the provided test booklet, all of CAEL tests used in this study needed to be transcribed before any analysis could be performed. All of the tests used in the present study were typed out and saved as .txt files for later analysis using WordSmith Tools 5.0.

During the transcription of these tests, several issues arose that needed to be addressed. First of all, while overall spelling was generally accurate, numerous spelling errors were identified in the CAEL test writings. These errors covered a wide range, from minor mistakes, to serious errors that obscured what the test takers true meaning
might have been. Additionally, since all of the tests were hand written, and each writer had differing levels of writing legibility, not all of their writing could easily be transcribed. Since the computer program being used to search for lexical bundles, WordSmith Tools 5.0, would only pull out recurrent sequences which were identical in form (spelling), it was necessary to have, as much as possible, standardized spelling throughout the corpora. Knowing the limitations of the software used in the analysis of the collected corpora, and that spelling was not the focus of the present study, the decision was made to correct small spelling errors in test takers writing. This decision to correct minor spelling errors is also supported by the method in which the CAEL assessment is evaluated. As mentioned earlier, CAEL tests are read aloud when being judged, and this process tends to lessen the value of spelling in the grading process. While corrections did occur during the transcription process, these were done only when it was clear what word(s) the test taker had been attempting to put down on the page. This resulted in a number of alterations being made to test taker writing, but was deemed necessary in order to achieve the type of analysis required to answer the aforementioned research questions. In addition, in reviewing/transcribing the written response portions of the CAEL tests, it was observed that some of the tests had written comments added by those judging their writing. While most of these comments had no impact on the transcription process, some tests were clearly marked to indicate that passages had been directly copied from either one of the two articles students were required to read before writing their response. Although this was not a common occurrence, whenever those evaluating the tests clearly marked passages as having been copied from either one of the two articles, these passages were not transcribed and were therefore not included as part
of the corpora. Since copied passages clearly do not represent a writer's own language abilities, including these passages in the corpora may have led to misleading results. While it is impossible to correctly identify all passages that may have been copied by test takers, where identified, judges comments were used to help eliminate these directly copied passages from the corpora. Despite leaving these passages out of the transcription, as we will see later on, many of the lexical bundles found in the corpora can still be found within at least one of the two reading articles. Even with the relatively small size of the corpora used in the present study (two corpora of roughly 75,000 words each), its highly specialized nature, along with the fact that it allows for a comparison between two groups of non-native writers, provides an interesting picture of how ESL/EFL writers use lexical bundles in their academic writing. Additionally, this analysis may provide insight into how ESL/EFL writers' use of lexical bundles changes as their academic English writing improves.

While compiling the corpora used for the present study, the most recently archived tests (2010) were transcribed first. The transcription process gradually worked backward chronologically with all of the available tests for any given year being typed out before moving on to the previous year's collection of tests. Following this path, the two corpora are primarily composed of tests from 2007-2010, with a small portion coming from 2006. The two corpora therefore consist of archived tests written between the period of 2006 and 2010.

5.1 **Identification criteria of Lexical Bundles**

Given the strong support for the investigation of 4-word lexical bundles in previous research (Chen 2008, Cortes 2002, Biber et al. 2004), the present study uses 4-
word bundles as a starting point in the analysis of the language used by writers in each of the two corpora (reasoning for the 4-word cut-off in previous lexical bundle research is discussed below).

The following sections begin with a discussion of how 4-word lexical bundles were identified within the lower-level and higher-level corpora used in this study (LLC and HLC). Following a detailed explanation of the reasoning behind both the frequency and range criteria used for the identification of 4-word lexical bundles, the criteria used to identify longer stretches of recurrent word clusters is also addressed.

5.1.1 Identification Criteria for 4-word Lexical Bundles in the LLC and HLC

Falling in contrast to the seeming majority of corpus studies of this kind that deal exclusively with 4-word bundles, in the present study, 4-word strings were used as a starting point in lexical bundle identification.

In her own work on lexical bundles, Cortes supports the use of the 4-word strings in studies of this kind since many 4-word bundles hold 3-word bundles within their structures (Cortes, 2004, p. 401). Additionally, Hyland echoes this sentiment, suggesting that 4-word bundles “offer a clearer range of structures and functions than 3-word bundles” (Hyland, 2008, p. 8).

In order to examine the distribution and function of lexical bundles present in both the LLC and HLC, it was first necessary to identify the most commonly occurring 4-word strings in each of these two corpora. In one of the earlier publications on lexical bundles, Biber et al. (1999) set the frequency cut-off point for 4-word lexical bundles at 10 occurrences per million words. Conversely, five years later in 2004, Biber et al. used
a more restrictive criterion by increasing the minimum frequency of occurrence to 40 times per million. Within these two extremes, it is also possible to see a trend among other researchers toward using a frequency cut-off point of 20 occurrences per million (Cortes 2004, 2008; Hyland 2008; Chen 2008).

While the more standard 20 occurrences per million words was used as a starting point in the present research, given the size of the corpus used in this study and its highly specialized nature, this would equate to a normalized frequency count of around 1.5 occurrences per 75,000 words. Using a frequency cut-off point of 1.5/75,000 words in the present study presents significant problems. First, the use of this criterion would allow nearly any 4-word sequence that appears within the corpora to qualify as a lexical bundle. Secondly, given that all of the texts contributing to the corpora were dealing with the same subject, it is likely that many structures would be repeated numerous times within the corpora and would therefore result in a situation where nearly every word string appearing in each of the two corpora would be classified as a lexical bundle.

Knowing that much of the previous research into lexical bundles involved the use of large sized corpora consisting of many hundreds of thousands, and even millions of words (Biber et al, 1999, Biber et al, 2002, Hyland 2008), these studies were not necessarily the most appropriate models to use when dealing with corpora of much smaller sizes. To get a better understanding of how lexical bundles are identified in smaller sized corpora, it was necessary to investigate how this problem had been addressed in previous research of this type. While there are admittedly fewer studies making use of smaller scale corpora, both Chen (2010) and Levy (2003) have used corpora of comparable size and specificity to those of the LLC and HLC. Although most researchers interested in this area of
corpus linguistics have mentioned that the frequency and range criteria are arbitrary choices, it is still necessary to provide some basis for the frequency and range criteria that were used in the present research. For this reason, both Chen (2010) and Levy (2003) will be discussed in turn, with the hope of providing a better understanding for the resulting decisions on the frequency and range used in this study.

Chen’s (2010) study of lexical bundles consisted of two corpora made up of English for Specific Purposes Textbooks (99,774 words) and Electrical Engineering Introductory Textbooks (247,346). In her study, whole chapters from textbooks deemed to be relevant to the topic of study were included in her corpora (along with one lab manual), and each of these chapters were counted as one “text” when entered into WordSmith Tools 4.0 for analysis. In terms of the frequency and range criteria used in the identification of lexical bundles in her study, Chen used a minimum frequency cut-off point of 20-occurrences per million words (adjusting for the size of her corpora, this translated to a minimum frequency of occurrence of 8 times within her corpora), and a range of 5 (must occur in at least 5 of the chapters/lab manual). While Chen’s study made use of a specialized corpus, in this case dealing exclusively with the topic of electrical engineering, it also dealt with multiple topics within this area of focus. For that reason, it can still be seen as a somewhat general corpus in comparison to the highly specialized nature of the material used for the compilation of the LLC & HLC. Again, since all of the written work included in the LLC & HLC was written in order to respond to one prompt, and is therefore centered on only one topic, these corpora have a much higher degree of specialization than that used by Chen (2010). Additionally, since each individual entry into the corpus used by Chen was no doubt substantially longer than
those in the LLC and HLC (entire chapters Vs. short essays), caution should be taken in applying the same range criterion to the present study.

Levy (2003) is an additional example where small-scale corpora were used in the investigation of lexical bundles. In her 2003 study on lexical bundles in professional and student writing, Levy compiled two separate corpora. The first corpus of professional essay samples consisted of 108 essays for a total of 127,000 words taken "from a variety of academic fields" (Levy, 2003, p. 28). The second corpus used in Levy’s study, the student essay corpus, was made up of a total of 110,000 words coming from 223 entrance essays written on one topic. Levy’s research presents potentially the closest comparison to the corpora used in the present study since they are both of comparable size, and with reference to the student essay corpus, deal with the same level of specificity in terms of content. Despite this high level of specificity however, Levy chose to implement extremely low frequency and range criteria to identify lexical bundles in her corpora. Setting both the frequency of occurrence and range criteria to 3 occurrences within her corpora, the criteria used in her study created an overly inclusive list of lexical bundles. Indeed, using the criteria described above, Levy identified 233 lexical bundles in the professional essay corpus and over 2,000 separate bundles in the highly specialized student corpus.

Using the information and results supplied by the two aforementioned studies, it was decided that in order to produce a manageable and valuable list of lexical bundles, the frequency and range criteria used in the studies mentioned above would need to be modified. Seeing as the frequency cut-off used by Levy (2003) was far too inclusive to be of value in the present highly specialized corpora, the criteria used by Levy were
eliminated, since they were judged to be too inclusive. Looking instead to Chen, and realizing that the corpora used in her study were far more general than those used in the present study, it was decided that a frequency of 20 occurrences per million (or a minimum of 8 occurrences in the EEITC of 247, 346) would need to be increased to compensate for the highly specialized nature of the corpora used in the present research. In order to produce a manageable set of lexical bundles used for later analysis, it was therefore decided that a minimum frequency cut off point of 16 occurrences would be used for the identification of lexical bundles within the LLC & HLC. This number was decided upon by taking the raw frequency of occurrence Chen (2010) used in her study and attempting to adjust for both the size and specialization of the corpora used in the present research. This represents a departure from much of the previous corpus based research dealing with lexical bundles, but was necessary in order to address the problems that would result from using the same selection criterion found in previous authors research on lexical bundles. As will be shown later on, this frequency cut-off point, although restrictive, still allowed for a high number of lexical bundles to be identified.

While frequency is one of the most important selection criteria used in the investigation of lexical bundles, range is also an important factor to consider in order to ensure we are not including lexical bundles that, although frequently occurring, only appear in a limited number of texts. Range deals with the spread of lexical bundles within a corpus, and helps to ensure that the lexical bundles identified through analysis will be common within the corpus as a whole. Again, there seems to be some support within previous research to use a range cut-off point of five texts (Cortes (2004, 2006), Biber 2004, Chen 2008), however due to the high number of small sized texts included in
the LLC & HLC, a more restrictive range criterion was also necessary. A range of 5 in the LLC & HLC would have little impact on the identification of lexical bundles in the present study. With the exception of Levy (2003), most of the previous studies on lexical bundles have included individual texts of larger sizes in their corpora than those used in the present research. Examples of this sampling method are seen in Chen (2010), where entire chapters of textbooks were used, Hyland (2008) where research articles/PhD dissertations/MA/MSc theses were included, and Biber et al. (2004) who used long stretches of classroom discourse as individual texts. Compared to the essays used in the present research, there is an obvious difference in terms of size of the individual texts used in the compilation of the corpora. The size of the individual texts included in corpora used for the purposes of lexical bundle identification will no doubt have an effect on the range of occurrence in which said bundles are found. Also, seeing as how all of these studies exist on a continuum in terms of the number of texts used in the creation of their corpora, it may be inappropriate to use the somewhat ‘standard’ method of implementing a minimum of 5 texts as the range criterion. For example, using the same range criterion of 5 in two corpora composed of a differing numbers of texts may not provide the most valid range criterion. While the use of an absolute number is still common in terms of defining the range criterion in corpus studies focusing on lexical bundle identification, Hyland (2008) breaks from this mold by implementing a percentage-based figure for identifying lexical bundles. This appears to be a more stable and potentially viable form of identifying the range criterion since it can allow for a more accurate comparison among corpora made up of differing numbers of texts. In Hyland’s 2008 study on lexical bundles in research articles, masters theses, and PhD dissertations,
he used a range criterion of 10% to ensure that the lexical bundles he identified appeared in a assortment of the texts that made up his corpora. Seeing as how the LLC & HLC in the present study are made up of differing amounts of text, a percentage based range figure may provide a more stable calculation than the more common cut-off point of 5 texts seen in the work of many previous researchers (Chen 2010, Cortes 2002, Biber et al. 2004). While the 10% cut-off is a good starting point for the identification of lexical bundles in the present study, due to the size of the texts being analyzed, the percentage based range criterion used by Hyland needed to be adapted so that it could better account for the smaller text sizes found in the LLC & HLC. Although the texts used in Hyland’s study were each composed of thousands of words, those of the LLC & HLC both average between 200-300 words per essay. Recognizing that a smaller percentage figure would be needed in order to compensate for the much smaller average length per text of the LLC & HLC, it was decided that the range criterion for the present study would be set at 4%. This 4% minimum range frequency equates to a minimum range of occurrence of 14 in the LLC, and 10 in the HLC. The range and frequency criteria for the identification of lexical bundles in the present study are summarized in the table below.

Table 5.1.1 Summary of the criteria used for the identification of lexical bundles

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Frequency</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLC</td>
<td>16</td>
<td>14*</td>
</tr>
<tr>
<td>HLC</td>
<td>16</td>
<td>10*</td>
</tr>
</tbody>
</table>

• Based on a range of 4% rounded to the nearest whole number
Once the identification criteria for 4-word lexical bundles had been finalized, WordSmith Tools 5.0 was used to create a list of lexical bundles in the LLC & HLC. In order to ensure the security of the writing prompt, and therefore ensuring that the test topic itself would not be revealed, selected content words were changed so that structures including the topic could still be brought into analysis. Appendices A and B show the initial lists of lexical bundles identified in the LLC and HLC. Using the above criteria, a total of 112 four-word lexical bundles were identified in the LLC, and 65 four-word lexical bundles were identified in the HLC.

5.1.2 Identifying lexical bundles of five or more words

While the use of 4-word lexical bundles seems to have become a standard unit length in this type of research, problems still persist. Just as not all units of meaning are represented by individual words, not all lexical bundles are of the 4-word variety. A major consequence of limiting the investigation of lexical bundles to those of a 4-word length is that longer, more complex structures are ignored. Additionally, due to the type of concordancing software commonly used in this type of analysis, a consistent problem has developed whereby longer bundles (more than four words) are displayed by the concordancing program as multiple 4-word entries that show significant overlap. This problem is shown in the work of many previous researchers where strings such as “at the end of” and “the end of the” are both included into analysis, even though they most likely combine to form the longer 5-word string “at the end of the”. This can potentially lead to inaccurate/misleading results since instead of one bundle being identified as serving a particular purpose, 2 or more slightly different bundles will be counted as serving identical, or potentially different roles in the discourse when functional analysis takes
place. Referring to Appendices B & C, we can also see evidence of this phenomenon in both the LLC & HLC, where bundles such as “absorbing most of the” and “by absorbing most of” were identified by WordSmith Tools 5.0. These two separate bundles most likely combine to form the longer, 5-word bundle “by absorbing most of the”, however, as of yet, there has been no adequate measure in place to deal with this problem. This issue has also been identified by Nekrasova (2009), whose own work on lexical bundles led to the following statement:

some of the bundles used in the study were simply parts of longer stretches that were cut off by corpus software into shorter individual sequences (e.g., what I want to, want to do is). This, however, is a common concern in research on lexical bundles that has not been resolved yet

More recently, Chen and Baker (2010) have made attempts to deal with the problem of overlapping structures when investigating 4-word lexical bundles. To deal with this issue, Chen and Baker (2010) identified two different kinds of overlap in their list of lexical bundles generated by WordSmith Tools 4.0, “complete overlap” and “complete subsumption”. “Complete overlap” refers to “two 4-word bundles which are actually derived from a single 5-word combination” (Chen & Baker, 2010, p. 33). In other words, 3 of the 4 words in two different 4-word bundles overlap, and each bundle occurs with equal frequency in the corpus. On the other hand, “complete subsumption” is used to refer “to a situation where two or more 4-word bundles overlap and the
occurrences of one of the bundles subsume those of the other overlapping bundles(s)” (Chen & Baker, 2010, p. 33). While “complete overlap” was used to identify 4-word bundles that occurred with similar frequencies, and subsequently combine these bundles to form longer lexical bundles, “complete subsumption” did not use frequency of occurrence as a factor in deciding to combine 4-bundles into longer stretches of recurrent language. The only example of “complete subsumption” given by the authors was the combination of as a result of which occurred 17 times, and a result of the which occurred only 5 times. While the combination of these two 4-word bundles to form the longer as a result of the seems appropriate, caution should be taken in overly extending this principle. Significant discrepancies in terms of frequency of occurrence may actually indicate the presence of two separate 4-word bundles, and should therefore not be combined to form a longer lexical bundle. For example, in the present study, the 4-word bundles is one of the and one of the most were both identified. Although they share three overlapping words (one of the), evidence from concordance lines shows that they often serve different functions within the text and should therefore not be combined to form the longer 5-word bundle – is one of the most.

While the approach used by Chen and Baker (2010) is a strong step forward in the investigation of lexical bundles, the present research takes a more conservative approach to the combination of partially recurrent 4-word lexical bundles. In order to address this issue in the present research, once the initial list of 4-word bundles had been generated by WordSmith Tools 5.0, manual analysis was undertaken with the aim of identifying which partially recurrent 4-word strings were actually longer stretches of repeated language. To this end, it was decided that any two or more 4-word strings that had 3 words in common
(i.e. "a marked increase in" and "marked increase in the"), and had both a frequency of occurrence and range of +/-4 would be combined to form a longer 5, 6, or even 7 word lexical bundle. These longer bundles were subsequently checked for frequency of occurrence within the corpus as a whole. If the figures for both frequency of occurrence and range were +/-4 of the frequency of occurrence and range of either of the two individual 4-word strings, the four word strings were removed from analysis and replaced with the new, longer lexical bundle that had been identified. Having a margin of error of +/-4 is both cautious and appropriate in this case since, as mentioned earlier, there may be small spelling/meaning errors within the two corpora used in this study. The margin of error represented by the +/-4 frequency/range helps to take this factor into account while still keeping a very strict guide for the identification of lexical bundles beyond the standard 4-word format. Using this method of identifying longer stretches of repeated language, and eliminating their partially recurrent 4-word constituent clusters, the original list of 177 recurrent strings was reduced to a total of 130 lexical bundles of varying lengths.

By extending the investigation to include not only 4-word bundles, but also longer 5, 6, and 7-word bundles, a higher percentage of the truly formulaic language present within the two corpora can be identified.

---

6 In order to more accurately identify longer stretches of recurrent discourse, lexical bundles occurring as few as 14 times were included into analysis when combining 4-word bundles to form 5-7 word lexical bundles.

7 One exception to the +/-4 rule was made in the HLC for the combination of most of the ultraviolet and of the ultraviolet radiation which were combined to create most of the ultraviolet radiation. This longer, 5-word bundle, had a +/-5 frequency of occurrence as a 5-word lexical bundle.
5.2 *WordSmith Tools 5.0 – Lexical Bundle Identification*

Originally released in 1996, as of late 2010 WordSmith Tools has now reached its 5th iteration. Over this time, the WordSmith Tools lexical analysis software has become one of the most popular and widely used concordancing programs available to corpus linguists. In an article discussing some of the most popular concordancing programs available at the time of its publication, Ari(2006) identified WordSmith Tools 3.0 as “the most efficient in its search for lexical bundles” (Ari, 2006, p. 39). Due in part to this review, along with its widespread use within corpus linguistic studies, WordSmith Tools 5.0 was chosen as the lexical analysis software used to search for lexical bundles within both the LLC & HLC.

In order to create the original lists of 4-word lexical bundles using WordSmith Tools 5.0, it was first necessary to produce an index file that contained all of the tests used in the two corpora. One index file was created for each of the two corpora. Once these index files had been created, the *wordlist* function was used to generate a list of all words appearing in the corpora. Using the *clusters* function, criteria were set to display only those lexical bundles that were 4-words in length and had a minimum frequency of occurrence which matched the criterion listed above. Given that WordSmith Tools 5.0 allows any output produced to be exported to Microsoft Excel files, all results were saved to Excel for easier manipulation and tracking. Once these excel files had been created, both lists were manually checked so that 4-word bundles that did not meet the minimum range criterion could be eliminated from future analysis.
5.3 Analytic Framework

In the analysis of the lexical bundles discovered through WordSmith Tools 5.0, several methods were used in order to provide as complete a picture as possible of the use and distribution of lexical bundles within both the LLC & HLC. While the use of Biber et al.'s (2004) functional taxonomy has provided interesting results in the past (Biber et al. 2004, Chen 2008/2010), additional types of frequency analyses were added to this functional analysis. These additional investigative methods include the analysis of differences in frequency of occurrence of specific lexical bundles between the two corpora, and the analysis of all lexical bundles for their appearance within the two articles that were read by all test takers before the commencement of their writing task. In the following sections each of these approaches will be elaborated on and further defined.

5.3.1 Functional Taxonomy – Biber et al. (2004)

To better understand the functions lexical bundles played within the discourse test takers had created, Biber et al.'s functional taxonomy from 2004 was used to classify all lexical bundles that were discovered. In using Biber et al.'s (2004) functional taxonomy, concordance lines of all 4, 5, 6 and 7-word lexical bundles were analyzed with the goal of identifying the purpose each of these bundles held within the text. Biber et al.'s (2004) taxonomy identifies three main functions of lexical bundles: Stance, Discourse, and Referential. Each of these main functions is also divided into numerous sub-categories that were also used to identify the function of lexical bundles in this study (refer to section 4.3 for further details on Biber et al.'s functional taxonomy).
5.3.2 Frequency of occurrence and distribution

As mentioned in the methodology section of this paper, normalized frequency counts based on the common lexical bundle cut-off point of 20 occurrences per million would not be appropriate for the highly specialized corpora used in this study. Seeing as very little previous research into lexical bundles has used such highly specialized corpora, the cut-off point used here is necessarily arbitrary. Recognizing this fact, it was decided that a more in depth form of analysis would be beneficial. As opposed to simply identifying which word-strings qualified as lexical bundles, it was decided that within the list of identified lexical bundles, an analysis of their distribution within the two corpora would also take place. For example, while *is one of the* qualified as a lexical bundle in both the LLC and HLC, by looking at the distribution of this bundle within each of these corpora, we can see that this lexical bundle is more than 2.5 times more likely to occur in the LLC than in the HLC. Additionally, a closer look at the distribution of lexical bundles of differing lengths was undertaken so that a better understanding of how non-native writers make use of longer lexical bundles could also be achieved. By better understanding the distribution of lexical bundles within the two corpora, we may also gain a better understanding of the path of progression non-native writers follow in their academic writing. An analysis of frequency of occurrence will take place in relation to both lexical bundle length, and required reading articles.

5.3.2.1 Lexical Bundles and the Required Readings

The use of the CAEL tests presents an interesting opportunity to researchers due to the fact that test takers are given articles to read before beginning their own writing. These articles are topic specific and are closely related to the writing students are
required to perform during the written response portion of their tests. While it would be likely that at least some of the lexical bundles found within the LLC and HLC would also appear in the assigned reading articles, it was decided that an investigation into how often this situation occurs would create a better understanding of how non-native writers make use of lexical bundles and any source material on which they are drawing.
6 Results and Discussion

Given the multiple methods implemented to investigate the use, distribution, and function of lexical bundles in both the LLC & HLC, this chapter is divided into three separate sections so that each approach can be discussed in turn. Each section begins with a brief introduction to the type of analysis being used, and is followed by the associated findings from both the LLC & HLC. Following the individual discussion of findings in the LLC & HLC, each section closes with a comparison of these findings, so that a better understanding of the differences between low-level and high-level non-native English writers can be achieved. Where appropriate, comparisons to previous studies are also presented in order to give some context to the findings uncovered through my analysis. Before launching into a detailed analysis of lexical bundles found in the writing produced by low level and high-level non-native English writers, the following section presents a brief overview of the lexical bundles identified in each of these two corpora.

6.1 Initial Overview of Lexical bundles in the LLC and HLC

Using the aforementioned frequency and range criteria, an initial list of 177 4-word lexical bundles was generated by WordSmith Tools 5.0 (Appendices B & C). This number was reduced to 130 by combining overlapping 4-word structures to create longer stretches of recurrent language (Appendices D & E). Each of these raw lists shows the most common lexical bundles found within the writing produced by test takers in this study. Looking at the two initial lists of lexical bundles generated by WordSmith Tools 5.0 (Appendices B & C), it is easy to see that the LLC contains significantly more 4-word
lexical bundles than the HLC using the frequency and range criteria outlined above (112 Vs. 65). Once this initial list had been revised to combine overlapping 4-word bundles to form 5, 6, and 7-word lexical bundles, these numbers dropped to 75 lexical bundles in the LLC and 55 lexical bundles in the HLC. While this still represents a noticeable difference, the revised lists show a less striking difference in terms of raw numbers of lexical bundles identified. As we will see later on however, these differences become more pronounced as we look more closely at the bundles by each group of writers.

Given that the test takers in this study were writing in response to an essay prompt which asked a question of the writers (i.e. "Does farm land reduction present a serious threat to life on earth?"), it is not surprising to see that many of the most common lexical bundles found in each of the two corpora are also found within the essay prompt itself. As is often encouraged by teachers who are asking students to respond to questions, the test takers in this study most often chose to provide ‘full’ answers to this question, including much of the original question in the answer itself (i.e. Yes, farm land reduction does present a serious threat to life on earth), and therefore lexical bundles that contain portions of the essay prompt are common in both corpora.

Now that a general overview of the lexical bundles identified in both the LLC and HLC has been given, the following sections provide detailed analysis along each of the analytical lines previously discussed: functional, frequency-based, and in relation to the reading articles.

6.2 Functional Analysis

To gain a better understanding of how the writing produced by low-level and high-level non-native English writers differed in the two corpora, it was necessary to
conduct an in depth functional analysis of all the lexical bundles found in each of the two corpora. To accomplish this task, Biber et al’s 2004 functional taxonomy was used as a guide in assigning functions to each of the lexical bundles generated by WordSmith Tools 5.0 (see section 4.3). While this type of functional analysis is useful in discovering the functions of lexical bundles in corpora of various kinds, it can also be viewed as a somewhat subjective measure. Read and Nation (2004), in their review of research methods used to identify and examine lexical bundles and other forms of formulaic language identify this issue as an area of concern in corpus linguistics of this type. Addressing the issue of internal reliability, Read and Nation (2004) recommend that “there should be at least two analysts (or raters) working independently through a substantial sample of the data, if not the whole set” (Read & Nation, 2004, p. 34). Keeping in mind Read and Nation’s recommendations, and recognizing the inherent subjectivity of assigning functions to lexical bundles, the functional classification of lexical bundles in the LLC & HLC was performed not only by the researcher, but also reviewed by an expert in the field of formulaic language. Following my initial classification of all lexical bundles into the categories defined by Biber et al 2004, these classifications were reviewed by the expert, to ensure agreement between at least two informed and expert judges was achieved.

All of the lexical bundles appearing in the LLC & HLC were classified according to the function they served based on ‘in context’ analysis – i.e. using concordance lines generated by WordSmith Tools 5.0. As discussed in Section 4.2, three major functional categories of lexical bundles were identified in Biber et al (2004), and often used in research of this kind (Cortes, 2004, Chen 2008, Chen & Baker 2010). Within each of
these major functional categories, more detailed functional sub-categories were also used for the classification of lexical bundles in this study.

In the following sections all of the lists of lexical bundles are presented in order of frequency with the most frequently occurring lexical bundles appearing at the beginning of the list.

6.2.1 Lexical Bundles in the LLC - Functional Analysis

A functional analysis of all of the 4-7 word lexical bundles identified in the LLC is provided in the following sub-sections. This functional analysis begins with the identification and listing of all stance bundles used by LLC writers. Following a discussion of the stance bundles, discourse-organizing bundles are the next functional category examined. Lastly, referential bundles used by lower-level writers in this study are presented.

6.2.1.1 Stance Bundles in the Lower Level Corpus (LLC)

A total of 22 stance bundles, falling into two distinct functional sub-categories were identified in the LLC. The two main functions of these lexical bundles are: epistemic and attitudinal. The following table shows all of the lexical bundles that were found to serve these functions in the LLC.

Table 6.2.1.1 Stance bundles in the LLC

Epistemic - Impersonal: a serious threat to
serious threat to life
threat to life on
present a serious threat
farm reduction present a
reduction present a serious
threat to the farm
farm reduction is a
the farm reduction present
the farm reduction is
serious threat to the
surface by absorbing most
is a serious threat
farm land is presented by the release
that farm reduction present
reduction is a serious
does farm reduction present
farm reduction presents a
presents a serious threat

\textit{Attitudinal – Prediction}: a marked increase in the
if the consumption of
consumption of CFCs continues at

Of the 22 stance bundles in the LLC, the majority (19) were used to express
impersonal epistemic stance. These bundles are most often used to express some kind of
opinion or judgment on the topic under discussion, as can be seen in the following
examples:

Farm land reduction does \textit{present a serious threat} to life on earth.
The farm reduction present a \textit{serious threat to life}.
Farm land \textit{reduction is a serious} problem for people.
Farm land reduction \textit{is a serious threat}.
We also know what could be a \textit{threat to the farm} land.

While most of the stance bundles were identified as serving impersonal epistemic
functions, the LLC also contains three bundles that serve to provide a ‘prediction’
function when viewed in context. These three bundles are: ‘a marked increase in the’, ‘if
the consumption of’, and ‘consumption of CFCs continues at’. Sample concordance lines
showing the function these ‘prediction’ lexical bundles play is provided below.
For example, a marked increase in the incidence of skin cancer is expected. Scientists have warned that if the consumption of CFCs continues at the present rate...

The two sample examples shown above both illustrate how writers in the LLC use bundles to show the reader potential consequences associated with various actions/events discussed in their writing. Given the nature of the essay prompt and the argumentative style responses it was meant to elicit, it is not surprising to see that these types of bundles were present in the writing test takers produced. Although numerous subcategories of stance bundles are defined in Biber et al's (2004) taxonomy, the LLC does not contain any bundles that serve either desire or obligation functions. Again, given that the essay prompt encouraged an argumentative style of writing, it is not surprising to see a lack of desire/obligation functioning bundles in test taker writing.

6.2.1.2 Discourse Organizing Bundles in the LLC

A total of six discourse organizing bundles, falling into two separate functional sub-categories were identified in the LLC. The two main functions of the identified bundles are: Topic Introduction/Focus and Elaboration/Clarification. The following table shows all of the lexical bundles that serve these functions in the LLC.

Table 6.2.1.2 Discourse organizing bundles in the LLC

*Topic Introduction/Focus:* according to the article
accordingly to the reading
according to reading (#)

*Topic Elaboration/Clarification:* is one of the
one of the most
on the other hand
Although only six discourse organizing lexical bundles were identified in the LLC, they cover two separate categories of discourse organizing functions: *topic introduction/focus*, and *topic elaboration/clarification*. The first sub-category, *topic introduction/focus*, was primarily used by lower-level writers to inform the reader that the writer was about to reference information found in either of the two reading articles test takers were asked to read before beginning their writing. While not always followed by a direct quote, these bundles nevertheless made use of information found in the reading articles.

*According to the reading 1* “The farm…”

*According to the reading* passage, destruction of farm makes…

*According to the article*, farm reduction is a serious threat

Conversely, the two topic elaboration/clarification bundles were mostly used to give the reader more detail concerning the topic being discussed. Additionally, these bundles can be seen to provide a higher level of detail used to support arguments put forward by the writer.

Which is *one of the* consequences of bad usage

Skin cancer in human *is one of the* predicted effects

Farm reduction *is one of the most* causes of global…

*One of the most* problem, that the…. 
6.2.1.3 Referential Bundles in the LLC

By far the category with the highest representation, a total of 47 referential bundles, were identified in the LLC. Again, these bundles can be further classified into sub categories, in this case all of the identified bundles fall into one of three sub-categories: identification/focus, specification of attributes, time/text/place/reference.

Table 6.2.1.3 Referential Bundles in the LLC

**Identification/Focus:** most of the ultraviolet radiation
skin cancer in humans
the release of pesticides
the incidence of skin cancer in
the ultraviolet radiation found in the
reduction of crop yields of
increase in the incidence of
food chain in the
cancer in humans is one of
crop yields of various plants
one of the predicted effects
that the farm reduction
a shield to earthly life
killing of certain forms of aquatic life

**Specification of Attributes:** to life on earth
the farm land is
farm land is a
by absorbing most of the
by the release of
form of an oxygen molecule
farm land serves as a shield
earth’s surface by absorbing
is a region of concentration of
protecting the forests and
synthetic compounds containing carbon fluorine and
produced by the action of solar radiation
as a shield protecting the forests
pesticides which are synthetic compounds containing
solar radiation on ordinary oxygen
the farm land serves
release of pesticides which are
the earth from the
protect the earth from
of concentration of the form of an

**Time/Text/Place Reference:** to the farm land
place in the farm
in the sun’s rays
in the farm land
of the farm land
found in the sun’s
the life on earth
life on the earth
chain in the seas
the forests and earth’s surface
of solar radiation on
of various plants and
and earth’s surface by

The majority of the lexical bundles identified in the LLC fall under the specification of attributes category which contains a total of 20 lexical bundles. These bundles were used to provide details concerning topics being discussed in test takers’ writing. Examples include:

Reduction causes a serious damaged *to life on earth.*

In short, *the farm land is* one the important.

A serious threat to the farm land is presented *by the release of* pesticides.

*Identification/focus bundles* and *time/text/place reference bundles* come next in order of frequency of occurrence with 14 and 13 occurrences respectively. As the name suggests, lexical bundles serving an identification/focus function, were most often used by writers to identify the subject they were writing about in any given sentence.
and earth’s surface by absorbing *most of the ultraviolet radiation*.

Other possible effects are *reduction of crop yields of plants*

So, my opinion is *that the farm reduction* does present a…

Conversely, the time/text/place category of lexical bundle was primarily used as place reference in the test takers writing.

Most of the ultraviolet radiation found *in the sun’s rays*.

Also know what could be a serious threat *to the farm land*.

The situation know *in the farm land* is different

**6.2.2 Lexical Bundles in the HLC – Functional Analysis**

A functional analysis of all of the 4-6 word lexical bundles (no 7-word bundles were found) in the HLC is provided in the following sections. As with the LLC, this functional analysis begins with the identification and listing of all stance bundles used by these writers. Following a discussion of the stance bundles, discourse-organizing bundles are the next functional category examined. Lastly, referential bundles used by higher-level writers in this study are presented.

**6.2.2.1 Stance Bundles in the Higher Level Corpus (HLC)**

A total of 14 stance bundles, all falling under the functional sub-category of impersonal epistemic stance were identified in the HLC. The following lists show all of the lexical bundles that serve this function in the HLC. As with all lists in this section,
these lists are sorted according to frequency, with the most frequently occurring lexical bundles appearing first.

Table 6.2.2.1 Stance Bundles in the HLC

*Epistemic – Impersonal*: a serious threat to
threat to life on
serious threat to life
present a serious threat
farm reduction present a serious
presents a serious threat
does present a serious threat
does present a serious threat
does present a serious threat

As with the LLC, these bundles were used primarily to give writers’ stance/opinion on the subject they were discussing in their writing. The following examples were taken from concordance lines produced by WordSmith Tools 5.0.

The farm reduction does present a *serious threat to life* on earth.

This presents *a serious threat* to the environment.

One serious *threat to the farm* land is presented by…

6.2.2.2 Discourse-Organizing Bundles in the HLC

Only two *discourse-organizing* bundles were identified in the HLC. Both of these *discourse-organizing* bundles can be classified under the *topic elaboration/clarification* sub-heading.
Table 6.2.2.2 Discourse-organizing bundles in the HLC

*Topic elaboration/clarification*: is one of the

one of the most

As can be seen through the concordance lines containing these lexical bundles, they are most often used to add clarification to topics being discussed in test taker writing.

Farm reduction *is one of the* major concerns human...

*Is one of the* reasons why farm...

Farm land reduction was considered *one of the most* menacing global threats.

but *one of the most* important causes is...

6.2.2.3 Referential Bundles in the HLC

In a similar fashion to the LLC, referential bundles in the HLC far outnumbered any other category of lexical bundle identified in my functional analysis. A total of 39 referential bundles falling into three separate categories were identified, and are provided in the lists below.

Table 6.2.2.3 Referential bundles in the HLC

*Identification/Focus*: food chain in the

farm land is a

most of the ultraviolet radiation

reduction of crop yields

skin cancer in humans

of the farm reduction

certain forms of aquatic life

crop yields of various plants

the incidence of skin cancer in
that the farm reduction
of farm land is
the ultraviolet radiation found in
killing of certain forms of
farm land is the
earth's surface by absorbing
the place in the
the loss of farm

**Specification of Attributes:** the farm land is
reduction of the farm
the reduction of the
reduction of farm land
the reduction of farm
to the life on
by absorbing most of the
surface by absorbing most
serves as a shield

**Time/Place/Text Reference:** to life on earth
of the farm land
the life on earth
in the farm land
to the farm land
place in the farm
chain in the seas
the farm land and
life on the earth
in the sun's rays
that the farm land
found in the sun's
forests and earth's surface

Contrary to the LLC, the majority of the stance bundles identified in the HLC fall
under the identification/focus heading which contains a total of sixteen lexical bundles.

These bundles were used to identify subjects being discussed in test takers' writing or
focus readers attention on a specific topic. Examples include:

Also, it is altering to *food chain in the seas*
When there is *reduction of crop yields*, there will be...

By absorbing *most of the ultraviolet radiation*

*Time/text/place reference* bundles and *specification of attributes* bundles come next in order of frequency of occurrence with thirteen and nine occurrences respectively. The time/text/place category of lexical bundle was primarily used as place reference in the test takers writing

This is important *to life on earth* as it protects the...

Recent studies showing the reduction *of the farm land*

Changes is the result of the reduction *in the farm land*

The *specification of attributes* lexical bundles were the least populated group of referential bundles identified in the HLC. These bundles were used to give further detail on subjects discussed in test taker writing.

The purpose of *the farm land* is to...

The other effects is *the reduction of the* crop yields

Protecting the earths *surface by absorbing most* of the ultraviolet radiation

**6.2.3 Functional Analysis - Comparing the LLC and HLC**

Given the discussion of individual corpus statistics in the previous sections, the following paragraphs provide a detailed description of the most notable differences between these two corpora as revealed by the aforementioned functional analysis. As a
recap of the findings discussed in sections 6.1.1 to 6.1.2, a table of summary statistics for each of the two corpora is provided below.

**Table 6.2.3 Summary of lexical bundles in the LLC & HLC**

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Stance Bundles</th>
<th>Discourse Organizing Bundles</th>
<th>Referential Bundles</th>
<th>Total # of lexical bundles identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLC</td>
<td>22 (29%)</td>
<td>6 (8%)</td>
<td>47 (63%)</td>
<td>75</td>
</tr>
<tr>
<td>HLC</td>
<td>14 (25%)</td>
<td>2 (4%)</td>
<td>39 (71%)</td>
<td>55</td>
</tr>
</tbody>
</table>

First, it is interesting to note that the LLC contains a total of 20 more lexical bundles than the HLC. It could therefore be said that the writing produced by lower-level non-native English writers in this study tends to be more formulaic than that of their higher-level counterparts. The LLC contains a higher number of bundles of all types and implies a greater reliance on formulaic language. This alone is an interesting statistic, however seeing as how differing numbers of lexical bundles were identified in the two corpora, it was decided that converting these raw statistics into percentage based accounts may provide a more valuable way of analyzing the functional distribution of lexical bundles in each of the corpora.

Looking specifically at the stance bundles used in each of these corpora, we can see that LLC also demonstrates a slightly greater reliance on bundles of this kind with a seeming decline in use as writers progress toward proficiency. Part of the reason for this difference in percentage usage is a result of LLC writers’ use of intention/prediction stance bundles. While the LLC contains three instances of bundles of this type, functional analysis of lexical bundles in the HLC revealed no bundles falling under this category.
Discourse-organizing bundles also seem to be over represented in the LLC (when compared to both the HLC and Biber et al. 2004). Looking more closely at the discourse organizing bundles found in the LLC and HLC, we can see that the major reason for this discrepancy in lower-level writers use of topic introduction/focus bundles that are used to introduce information taken from the reading articles (i.e. according to the article, according to the reading, according to reading #). Writers in the HLC, use very few discourse organizing lexical bundles (2), therefore the increased use of these types of bundles in the LLC, particularly those used to introduce information taken from the articles, may be interpreted as a sign of lower level non-native English writers. Potential reasons for this finding can be found in the Summary of Findings section below.

Perhaps the most apparent difference related to the functional distribution of lexical bundles in the LLC & HLC is their respective use of referential bundles. Although both the LLC & HLC show high numbers of referential bundles, percentage figures reveal that the LLC uses a comparatively small number of referential lexical bundles compared to the HLC (63% Vs. 71%). Again, potential reasons for this finding can be found in the Summary of Findings section found in Section 6.5.

The percentage-based analysis of the functional taxonomy of lexical bundles in the two corpora has demonstrated important differences in the ways lexical bundles are used by both groups of writers. However, these differences are given greater weight when shown in context with the findings of Biber et al’s 2004 functional analysis of lexical bundles in a large corpus of American and British English academic writing. Although Biber et al’s 2004 analysis of a large sized corpus of academic writing
produced only 19 lexical bundles, the proportion each functional category of lexical bundles held within the list of identified bundles can be compared to the results of the present study. In Biber et al (2004), three stance bundles, one discourse organizing bundle, and fifteen referential bundles were identified. Converting these raw statistics to percentages, we can see how these figures compare to those found in the LLC & HLC.

Table 6.2.5 Summary of lexical bundles in the LLC, HLC and Biber et al 2004 (academic prose)

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Stance Bundles</th>
<th>Discourse Organizing Bundles</th>
<th>Referential Bundles</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLC</td>
<td>29%</td>
<td>8%</td>
<td>63%</td>
</tr>
<tr>
<td>HLC</td>
<td>25%</td>
<td>4%</td>
<td>71%</td>
</tr>
<tr>
<td>Biber et al. 2004</td>
<td>16%</td>
<td>5%</td>
<td>79%</td>
</tr>
</tbody>
</table>

If we look at these three corpora as representing the progression non-native English writers aim to achieve in their academic writing (i.e. progressing from ‘low-level non-native English writers’ to ‘high-level non-native English writers’ and eventually achieving a level of writing proficiency similar to that of proficient English speaking academic writers) we can better understand the path non-native English writers seem to follow. Looking at each of the functional categories of lexical bundles in Table 6.2.5, we can see that, in every category, the HLC more closely resembles the functional category distribution of lexical bundles identified in Biber et al.’s 2004 study of English writers’ academic prose.

---

8 Although these bundles were originally identified in Biber et al. (1999), they were not given functional classifications at that time.
Using the statistics provided in this section we are able to come to the tentative conclusion that lower-level non-native English writers tend to make greater use of lexical bundles as a whole, in addition it seems that lower-level non-native English writers tend to underuse referential bundles and overuse discourse-organizing bundles. The percentage-based statistics provided by the results of Biber et al.’s 2004 findings also give greater context to the conclusions drawn from the comparison between the LLC & HLC and show the potential path of progression non-native English writers follow as they aim to become proficient second/tertiary language writers.

### 6.3 Frequency Analysis of lexical bundle lengths

While the functional classifications discussed in the previous section yielded interesting results, in order to get a better understanding of the writing produced by both high and low-level non-native English writers, frequency analysis was also implemented. Taking a look at the initial lists of 4-word lexical bundles generated by WordSmith Tools 5.0 (Appendices B & C), we can see that the LLC contains a significantly higher number of lexical bundles than the HLC. In fact, the initial list of 4-word lexical bundles in the LLC contains over 40% more bundles than the HLC (112 Vs. 65). Once the overlapping 4-word bundles had been replaced with longer stretches of recurrent language, the difference in number of lexical bundles found within the two corpora decreased to 75 Vs. 55 (27% more bundles in the LLC). Although the removal and combination of partially recurrent clusters in both the LLC and HLC decreased the statistical difference in numbers of lexical bundles identified, there was still a significant gap in terms of the number of lexical bundles in each of the two corpora. This section provides details on
the number of 4, 5, 6, and 7 word lexical bundles found within the two corpora used in this study. Summary statistics for the two corpora can be found in Section 6.2.3.

6.3.1 LLC – Frequency analysis of 4-7 word lexical bundles

Once the overlapping 4-word lexical bundles had been replaced with longer stretches of recurrent language, it was clear that the LLC made significant use of 5, 6, and 7 word lexical bundles. In fact 33% of the bundles identified in the Lower Level Corpus were over 4-words in length. A list of all the lexical bundles of various lengths identified in the LLC is provided below, results are grouped according to the length of the lexical bundle identified.

Table 6.3.1 4-7 word lexical bundles in the LLC

4-word Bundles: a serious threat
serious threat to life
threat to life on
present a serious threat
farm reduction present a reduction present a serious threat to the farm
farm reduction is a
the farm reduction present
the farm reduction is serious threat to the surface by absorbing most is a serious threat
that farm reduction present
one of the most reduction is a serious does farm reduction present
farm reduction presents a presents a serious threat
if the consumption of according to the article
according to the reading
according to reading #
is one of the
one the other hand
skin cancer in humans
the release of pesticides
food chain in the
that the farm reduction
to life on earth
the farm land is
farm land is a
by the release of
earth’s surface by absorbing
protecting the forests and
the farm land serves
the earth from the
protect the earth from
to the farm land
place in the farm
in the sun’s rays
in the farm land
of the farm land
found in the sun’s
the life on earth
life on the earth
chain in the seas
of solar radiation on
of various plants and
and earth’s surface by

5-word Bundles: a marked increase in the
consumption of CFCs continues at
most of the ultraviolet radiation
reduction of crop yields of
increase in the incidence of
crop yields of various plants
one of the predicted effects
a shield to earthly life
by absorbing most of the
form of an oxygen molecule
solar radiation on ordinary oxygen
release of pesticides which are
the forests and earth’s surface

6-word Bundles: the incidence of skin cancer in
the ultraviolet radiation found in the
cancer in humans is one of
farm land serves as a shield
is a region of concentration of
synthetic compounds containing carbon fluorine and
as a shield protecting the forests
pesticides which are synthetic compounds containing

**7-word Bundles:** farm land is presented by the release
killing of certain forms of aquatic life
produced by the action of solar radiation
of concentration of the form of an

Fifty, or 67% of the seventy-five lexical bundles identified in the LLC are 4-words in length. As we move forward to longer lexical bundles, we can see that the number of bundles drops as the length of the lexical bundle increases. Thirteen 5-word lexical bundles were identified in the LLC (17%), and eight 6-word bundles were found (11%). Four different 7-word bundles were found in the LLC (5%). As we will see later on, these longer lexical bundles were almost certainly copied from the reading articles since they were all found to also appear in those sources.

**6.3.2 HLC – Frequency analysis of 4-7 word lexical bundles**

After replacing the overlapping 4-word lexical bundles with longer stretches of recurrent language, it was clear that the HLC made the greatest use of 4-word lexical bundles. In fact 84% of the bundles identified in the Higher Level Corpus were 4-words in length. A list of all the lexical bundles identified in the HLC is provided below, results are grouped according to the length of the lexical bundle identified.

**Table 6.3.2 4-6 word bundles in the HLC**

**4-word Bundles:** a serious threat to
threat to life on
serious threat to life
present a serious threat
presents a serious threat
farm reduction is a
the farm reduction is
serious threat to the
is a serious threat
threat to the farm
threat to the life
does present a serious
one of the most
is one of the
food chain in the
farm land is a
farm land is the
reduction of crop yields
skin cancer in humans
of the farm reduction
that the farm reduction
of farm land is
earth’s surface by absorbing
the place in the
the loss of farm
the farm land is
reduction of the farm
the reduction of the
reduction of farm land
the reduction of farm
to the life on
surface by absorbing most
serves as a shield
to life on earth
of the farm land
the life on earth
in the farm land
to the farm land
place in the farm
chain in the seas
the farm land and
life on the earth
in the sun’s rays
that the farm land
found in the sun’s
forests and earth’s surface
5-word Bundles: farm reduction present a serious
farm reduction presents a serious
most of the ultraviolet radiation
certain forms of aquatic life
crop yields of various plants
the ultraviolet radiation found in
killing of certain forms of
by absorbing most of the

6-word Bundles: the incidence of skin cancer in

Forty-six, or 84% of the fifty-three lexical bundles identified in the HLC are 4-
words in length. As we move forward to longer lexical bundles, we can see a sharp drop
in the number of bundles identified, with eight 5-word bundles found (15%), and only
one 6-word bundle found in the HLC (2%). Contrary to the LLC, no 7-word bundles
were identified in the HLC.

6.3.3 Frequency Analysis – comparing lexical bundles of multiple
lengths in the LLC and HLC

With 29% more 4-7 word lexical bundles appearing in the LLC than in the HLC,
we are able to say that, in general, less proficient writers in this study tended to produce
more lexical bundles. In addition, by using the aforementioned approach to identifying
lexical bundles of 5,6, and 7 word lengths, we can see that less proficient writers also
seem to make greater use of longer lexical bundles in their writing. Although the number
of 4-word bundles in the LLC and HLC are comparable (50 Vs. 46), the HLC contains
fewer 5-word bundles, and only one 6-word bundle. This lies in stark contrast to the
LLC, which has thirteen 5-word bundles, and eight 6-word bundles; in addition the LLC is the only corpus where any 7-word bundles were found.

**Table 6.3.3 Summary of 4-7 word lexical bundles in the LLC and HLC**

<table>
<thead>
<tr>
<th>Corpus</th>
<th>4-word Lexical Bundles</th>
<th>5-word Lexical Bundles</th>
<th>6-word Lexical Bundles</th>
<th>7-word Lexical Bundles</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLC</td>
<td>50 (67%)</td>
<td>13 (17%)</td>
<td>8 (11%)</td>
<td>4 (5%)</td>
</tr>
<tr>
<td>HLC</td>
<td>46 (84%)</td>
<td>8 (15%)</td>
<td>1 (2%)</td>
<td>-</td>
</tr>
</tbody>
</table>

As with the functional analysis, these figures can also be converted into percentages to offer a more accurate comparison of the distribution of lexical bundles in each of the two corpora. With these figures in mind, it is therefore plausible to say that lower-level writers in this study tended to use longer lexical bundles than their higher-level counterparts, and that longer lexical bundles may be a sign of less proficient non-native English writing.

Again, previous studies can be referenced to give context to this conclusion. In a 1998 study on the phrasicon of EFL speakers, De Cock, Granger, Leech and McEnery used a corpus made up of interviews with advanced English language learners to come to the conclusion that non-native speakers used significantly more 4 and 5-word recurrent word combinations than native English speakers⁹. "It is particularly striking that the number of longer word combination tokens above frequency threshold (four and five-word recurrent combinations) is significantly higher in NNS" (De Cock, Granger, Leech & McEnery, 1998, p.72). Additionally, the authors cite Milton and Freeman (1996) along with Milton (1998) as additional sources of research that support the finding that

---

⁹ While this study looks at lexical bundles of up to 7-words in length, De Cock et al. (1998) only looked at recurrent word clusters of 2-5 words.
non-native speakers tend to use longer word sequences with greater frequency than native English speakers. The present study seems to support the findings reported in each of these studies, as well as adding additional information. In the aforementioned research, a distinction between native and non-native speakers use of lexical bundles was identified, the present research however, shows how these findings can be extended to the analysis of non-native writers of differing levels of proficiency. We can see from the results in Table 6.3.3 that the aforementioned distinction between native and non-native English speakers is also apparent when comparing low and high level non-native English writers. While non-native English speakers/writers as a whole may overuse lexical bundles, especially of longer lengths, this observation is even more prominent in the writing of lower-level non-native English writers.

The higher incidence of 5,6 and 7-word lexical bundles in the LLC may be at least partially related to a higher incidence of quotes and citations used in lower level writing in this study, as evidenced by the use of the three discourse organizing bundles that were used to reference information provided in the reading articles, i.e. *according to the article, according to the reading, according to reading #*. Although prominent in the LLC, none of these lexical bundles appeared commonly enough in the HLC to qualify as a lexical bundle and may help to explain some of the longer length recurrent word clusters found in the LLC. In light of this observation, an analysis of lexical bundles, with reference to the reading articles and the essay prompt, are discussed in the following section.
6.4 Lexical Bundles and the Assigned Reading Articles/Essay Prompt

The highly specialized nature of the corpora used in this study has allowed for a rare form of analysis. Given the fact that each of the texts included in both the LLC & HLC were written following the reading of two specific articles, and in response to one essay prompt, I was able to analyze how writers made use of structures present in each of these sources. While the majority of the lexical bundles identified in each of the two corpora also appear in either the essay prompt or at least one of the two reading articles (75% in the LLC Vs. 59% in the HLC), numerous bundles in each corpus were also found to be ‘original’ productions. The term ‘original’ is used here to refer to the fact that these word clusters, although common in test taker writing, do not appear in either the readings or the essay prompt. While it is not surprising that the majority of the lexical bundles identified in the LLC and HLC also appeared in the reading articles, it is important to see which bundles used in test taker writing were most likely to be sourced from the writers own language repertoire.

6.4.1 LLC bundles and the reading articles/essay prompt

Of the 75 lexical bundles identified in the LLC, 56 (75%) were also found to appear in the essay prompt, or either of the two reading articles. This leaves only 19, or 25% of the identified lexical bundles as original productions. These 19 lexical bundles are provided in the list below.
Table 6.4.1 Originally produced lexical bundles in the LLC

**Stance Bundles:** farm reduction is a
the farm reduction present
the farm reduction is
is a serious threat
that farm reduction present
reduction is a serious
farm reduction presents a
presents a serious threat

**Discourse Organizing Bundles:** according to the article
according to the reading
according to reading #
on the other hand

**Referential Bundles:** food chain in the
that the farm reduction
the earth from the
protect the earth from
the life on earth
life on the earth
chain in the seas

While the LLC was found to contain the highest percentage of lexical bundles over 4-words in length (33%), none of these 5, 6, and 7-word bundles can be classified as original productions since they were all found to appear in either the essay prompt or the reading articles. In total, eight stance bundles, four discourse organizing bundles, and seven referential bundles were judged to be originally produced by lower-level writers in this study.

Each of the eight originally produced stance bundles identified in the LLC are of the impersonal epistemic stance variety. Since the three intention/prediction bundles identified during the functional taxonomy in section 6.1 also appear in the reading articles, it is likely that these bundles are borrowed passages taken from the readings. It
is also interesting to note that 3 of the 4 originally produced discourse-organizing bundles identified in the LLC, were used exclusively by writers to reference material taken from the reading articles (according to the article, according to the reading, according to reading #). The originally produced referential bundles identified in the LLC are fairly evenly spread, with two serving identification/focus functions (food chain in the, that the farm reduction), two being used to specify attributes (the earth from the, protect the earth from), and three place reference bundles (the life on earth, life on the earth, food chain in the).

6.4.2 HLC bundles and the reading articles/essay prompt

Of the 55 lexical bundles identified in the HLC, 23 (42%) could be classified as original productions. These 23 original productions are provided in the lists below.

Table 6.4.2 Originally produced lexical bundles in the HLC

Stance Bundles: presents a serious threat
farm reduction is a
the farm reduction is
farm reduction presents a serious
is a serious threat
threat to the life
does present a serious

Discourse Organizing Bundles: -

Referential Bundles: food chain in the
of the farm reduction
that the farm reduction
of farm land is
the place in the
the loss of farm
reduction of the farm
the reduction of the
reduction of farm land
the reduction of farm
to the life on
the life on earth
chain in the seas
the farm land and
life on the earth
that the farm land

While the HLC was found to contain a lower percentage of lexical bundles greater
than 4-words in length, it is the only corpus that contained originally produced lexical
bundles greater than 4-words (i.e. farm reduction presents a serious). In total, seven
stance bundles, zero discourse organizing bundles, and sixteen referential bundles were
judged to be originally produced by higher-level writers in this study.

As with the LLC, all of the originally produced stance bundles in the HLC were
of the impersonal epistemic stance variety. Contrary to the LLC however, the HLC was
found to contain no originally produced discourse-organizing bundles. With sixteen
bundles, the referential function category was the most populated group of originally
produced bundles in the HLC. The originally produced referential bundles in the HLC
were fairly evenly split between each of the functional categories, with six falling under
the identification/focus classification, and five in each of the specification of attributes
and place reference sub-categories.

6.4.3 Originally Produced Bundles - Comparing the LLC and HLC

While numerous distinctions between higher-level and lower-level writers in this
section could be made, it is important to focus on some of the most salient distinctions
between these two corpora. One of the most obvious differences in the tables above is
the number of ‘originally’ produced bundles used by each of the two groups of writers. Although analysis of writing included in the LLC yielded a higher number of lexical bundles, fewer of these bundles were found to be original productions. While 75 lexical bundles were identified in the LLC, only 19 could be classified as original productions. In contrast, of the 55 lexical bundles identified in the HLC, 23 were classified as original productions. This alone is an important distinction, however a look at the functional composition of these bundles reveals another important fact. In the HLC, there were a total of sixteen ‘originally produced’ referential lexical bundles, while less than half this amount, seven, were identified in the LLC. As mentioned in the previous discussion on functions of lexical bundles in the LLC & HLC, higher-level writers in this study were found to use a greater percentage of referential lexical bundles in their writing, which is also consistent with the findings of Biber et al 2004 (as a proportion of the total number of lexical bundles identified). By eliminating all bundles found in each of the two corpora that also appear in either of the two reading articles or the essay prompt, we can see that higher-level writers were able to create more originally produced referential bundles than the lower-level writers in this study. Given that greater use of referential lexical bundles was also found to be indicative of higher-level academic writing (Biber et al. 2004, as well as Section 6.1), it is important to note that higher-level non-native English writers were able to produce more than twice as many referential bundles from their own language repertories than lower-level non-native English writers in this study.

As it relates to stance bundles, it is interesting to note that although ‘presents a serious threat’ appears as a lexical bundle in both of the corpora, its frequency of occurrence in each of the corpora seems to show large differences in usage. Looking at
the lists above, we can see that while ‘presents a serious threat’ is the most frequently occurring originally produced stance bundle in the HLC (40 instances), it is the least frequently occurring originally produced stance bundles in the LLC (only 18 occurrences). Although less frequently occurring in the LLC, an examination of concordance lines revealed that when used by these writers, in all instances, it was correctly incorporated into their writing. Considering the fact that the LLC used on average more lexical bundles, and used these bundles with greater frequency than in the HLC, this is an important finding. In order to elicit student writing, the prompt used in this test asked students “Does farm land reduction present a serious threat to life on earth?” If we are to answer in the affirmative, it seems as though there are two main options:

Yes, farm land reduction does present a serious threat.

Yes, farm land reduction presents a serious threat.

While both groups of writers preferred the first option when answering this question, it is evident that lower-level students relied more heavily on the ‘present a serious threat’ structure that was provided in the prompt (169 occurrences in the LLC Vs. 94 occurrence in the HLC). Both answers are valid, however it seems as though, yet again, lower-level non-native English writers were much more reliant on structures which appeared in the prompt and reading articles, while higher-level writers were able use alternate structures to express their thoughts. Taking a look at the remaining ‘originally’ produced lexical bundles used by each group of writers, we can also see that despite the
fact that the LLC was found to contain more lexical bundles, and in general a higher frequency of occurrence for these lexical bundles, the ‘originally’ produced lexical bundles were generally found to occur with more frequency in the HLC. The few exceptions to this finding are: farm reduction is a, life on the earth, is a serious threat, that the farm reduction. Table 6.4.3 displays all of the ‘originally’ produced lexical bundles in each of the two corpora along with frequency of occurrence figures for each bundle. Discourse-organizing bundles used by LLC writers to reference material taken from the reading articles are shown in italics.

Table 6.4.3 Originally produced lexical bundles in the LLC & HLC

<table>
<thead>
<tr>
<th>Rank</th>
<th>HLC Bundle</th>
<th>Freq.</th>
<th>Rank</th>
<th>LLC Bundle</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>THE LIFE ON EARTH REDUCTION OF THE FARM</td>
<td>54</td>
<td>18</td>
<td>FARM REDUCTION IS A THE FARM REDUCTION PRESENT</td>
<td>39</td>
</tr>
<tr>
<td>10</td>
<td>THE REDUCTION OF THE PRESENTS A SERIOUS THREAT</td>
<td>51</td>
<td>20</td>
<td>THE FARM REDUCTION IS</td>
<td>39</td>
</tr>
<tr>
<td>12</td>
<td>PRESENTS A SERIOUS THREAT</td>
<td>43</td>
<td>24</td>
<td></td>
<td>36</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>40</td>
<td>31</td>
<td>THE LIFE ON EARTH</td>
<td>29</td>
</tr>
<tr>
<td>14</td>
<td>FARM REDUCTION IS A THE FARM REDUCTION IS</td>
<td>39</td>
<td>32</td>
<td>LIFE ON THE EARTH</td>
<td>27</td>
</tr>
<tr>
<td>15</td>
<td>FOOD CHAIN IN THE REDUCTION OF FARM LAND</td>
<td>36</td>
<td>33</td>
<td>FOOD CHAIN IN THE IS A SERIOUS THREAT THAT FARM REDUCTION PRESENT</td>
<td>28</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>34</td>
<td>38</td>
<td>IS A SERIOUS THREAT THAT FARM REDUCTION IS</td>
<td>24</td>
</tr>
<tr>
<td>17</td>
<td>FARM REDUCTION PRESENTS A SERIOUS THE REDUCTION OF FARM</td>
<td>33</td>
<td>41</td>
<td>REDUCTION PRESENT REDUCTION IS A SERIOUS THAT THE FARM REDUCTION</td>
<td>24</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>31</td>
<td>44</td>
<td>ACCORDING TO THE ARTICLE</td>
<td>22</td>
</tr>
<tr>
<td>21</td>
<td>TO THE LIFE ON CHAIN IN THE SEAS</td>
<td>31</td>
<td>46</td>
<td>THAT THE FARM REDUCTION</td>
<td>22</td>
</tr>
<tr>
<td>23</td>
<td></td>
<td>29</td>
<td>47</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>27</td>
<td>THE FARM LAND AND OF THE FARM REDUCTION</td>
<td>27</td>
<td>53</td>
<td>CHAIN IN THE SEAS ACCORDING TO THE READING</td>
<td>19</td>
</tr>
<tr>
<td>29</td>
<td></td>
<td>27</td>
<td>56</td>
<td></td>
<td>19</td>
</tr>
<tr>
<td>30</td>
<td>LIFE ON THE EARTH</td>
<td>26</td>
<td>59</td>
<td>ON THE OTHER HAND</td>
<td>19</td>
</tr>
<tr>
<td>32</td>
<td></td>
<td>25</td>
<td>65</td>
<td>PRESENTS A SERIOUS</td>
<td>18</td>
</tr>
</tbody>
</table>
Based on the results displayed in Table 6.4.3, we can say that HLC writers in this study were able to make greater use of lexical bundles that were not sourced from the reading articles. Additionally, looking at the italicized lexical bundles in the above table, we see further support for the LLC writers’ reliance on the reading articles as a source to complement their own writing.

Using the analysis presented in this section, we can say that lower-level writers in this study made greater use of structures and phrasings presented in the essay prompt and the reading articles, while higher-level writers were able to create their own structures to express the ideas and thoughts they wanted to convey to the reader.

### 6.5 Summary of findings and Discussion

The analysis of low-level and high-level non-native English test taker writing was performed using an empirical, corpus-based approach which has allowed for the identification of potential differences in the way lexical bundles are used by low-level and high-level non-native English writers. Using multiple forms of analysis, it has become evident that there are indeed significant differences in how this aspect of
language is used by non-native English test takers of differing levels of proficiency.

Distinguishing characteristics between the two corpora were found in each of the forms
of analysis used in this study, and have yielded important results that add to our
knowledge base on how non-native English test takers make use of lexical bundles in
their writing.

A brief review of the original list of 4-word lexical bundles generated by
WordSmith Tools 5.0 revealed that lower-level writers in this study relied more heavily
on lexical bundles in the discourse they had produced. In order to get a more accurate
picture of the lexical bundles being used by non-native English test takers in this study
however, it was decided that overlapping 4-word bundles would be combined (where
appropriate) to form longer 5, 6, and 7-word lexical bundles. Upon completion of this
task, it was evident that, while there still existed a significant difference in terms of the
number of lexical bundles identified in each of the two corpora, this discrepancy was less
pronounced (75 lexical bundles in the LLC Vs. 55 in the HLC). Once the
removal/combination of overlapping lexical bundles in the two corpora had been
completed, three major forms of analysis were undertaken. In the following paragraphs
each of these approaches will be discussed in turn, along with some potential reasons for
these findings.

First of all, assigning functional roles to each of the 4, 5, 6, and 7-word lexical
bundles identified, using Biber et al.’s 2004 functional taxonomy, showed that low-level
writers in this study tended to rely more heavily on lexical bundles that serve both stance
and discourse-organizing functions when compared to both high-level non-native English
writers and English writers of academic prose (when taking a percentage based view of
lexical bundles used by these groups). Conversely, low-level writers in this study were also found to under use referential bundles when similar comparisons were made to the other two aforementioned groups of writers. A percentage based comparison of the functional roles of lexical bundles used by each of these groups of writers demonstrated that the functional distribution of lexical bundles in the LLC was furthest from the functional distribution of lexical bundles used by English writers of academic prose found in Biber et al. (2004). Comparing these figures to the HLC showed that the higher level writers fell somewhere in between these two extremes.

The greater use of discourse organizing bundles by low-level writers in this study can be interpreted as a sign of low-level writing. Potential reasons for this discovery can be found by taking a closer look at the actual discourse organizing bundles used by each of the groups of writers. Looking more closely at the discourse-organizing bundles identified, we can see that, while both groups of writers used discourse-organizing bundles of the topic elaboration/clarification subcategory (3 in the LLC Vs. 2 in the HLC), the LLC was the only group to use any topic introduction/focus discourse-organizing bundles. This is an important distinction since each of these introduction/focus discourse-organizing bundles were used to reference information found in the reading articles. This finding suggests that writers in the LLC were more reliant on information and structures found in the reading articles. This may be a strategy used by less proficient writers to cope with their limited vocabulary and store of formulaic sequences. Similar to the “teddy-bear principle” (Hasselgren, 2000) in which learners cling to words or phrases with which they feel comfortable using, writers in the LLC may have viewed the structures taken from the reading articles as safe zones which
they could use whenever they felt it was necessary. By incorporating passages and sentence fragments from the readings, lower-level writers in this study could supplement their limited language resources with structures they could be confident would be considered "proper" English. In other words, students may be incorporating structures taken from the readings, in their own writing, to ensure that what they are producing will be expressed in an acceptable form.

Another major finding discovered through the functional analysis of lexical bundles in the two corpora was that HLC writers used more referential bundles, as a percentage of total bundles used, than writers in the LLC. The main function of referential bundles is to "identify an entity or single out some particular attribute of an entity as especially important" (Biber et al. 2004, p. 393). Given that the writing task used in this study asked test takers to write an argumentative style essay with the implicit goal of providing a convincing argument for any conclusions they had come to, greater use of referential bundles may be taken as a sign that HLC writers relied more heavily on facts and evidence to support their position than LLC writers. This is also supported by the finding that LLC writers used slightly more stance bundles in their writing, something that suggests LLC writers preferred to argue their position using opinion statements.

Given the limited scope and size of the corpora used in this study however, these are only tentative conclusions that are in need of further research.

By taking a closer look at the original list of 4-word lexical bundles generated by WordSmith Tools 5.0, and eliminating/combining overlapping structures using the criteria identified in section 5.1.2, the present study was also able to investigate the differences in the frequency of occurrence of lexical bundles of varying lengths in each of
the corpora. By combining overlapping lexical bundles to create longer stretches of recurrent language (i.e. certain forms of aquatic and forms of aquatic life were combined to create the longer lexical bundle certain forms of aquatic life), we were able to see that lower-level writers also tend to use more lexical bundles of greater length than their higher-level counterparts. As mentioned previously, this has also been found to be a general sign of NNS discourse (DeCock et al. 1998). While the research performed by DeCock et al (1998) has shown that increased use of longer lexical bundles is a sign of non-native speaker discourse, the present study suggests that this feature is even more pronounced in the writing lower level non-native English writers produce. This finding brings us to the tentative conclusion that the increased use of longer lexical bundles in non-native English writers is a sign of lower-proficiency levels.

Analyzing the reading articles test takers read prior to beginning their writing, the final form of analysis shown above, allowed for the discovery that lower-level non-native English writers in this study used a higher proportion of lexical bundles that could also be found in the reading articles, and were most likely copied from this source. A total of seventy-five 4-7 word lexical bundles were identified in the LLC, however only nineteen could be classified as ‘original’ productions. In contrast, the HLC contained only fifty-five 4-6 word lexical bundles (no 7-word bundles were identified), yet nearly half of these bundles (23) were classified as ‘original’ productions. Therefore, the original list of lexical bundles generated by WordSmith Tools 5.0 may be slightly misleading if we are interested in investigating test takers’ actual repertoires of formulaic language. While initial indications showed that LLC writers actually used more lexical bundles than HLC writers, looking at the reading articles revealed that HLC writers actually used more
originally produced lexical bundles than LLC writers. Additionally, by taking a close
look at these originally produced lexical bundles, we can see that, while the LLC
contained more lexical bundles of 5, 6, and 7-words, all of these bundles were most likely
sourced from the reading articles, and therefore should not be considered as ‘originally’
produced sequences. Once again, this finding can be seen as an indication that LLC
writers may use copying as a strategy to compensate for their limited language abilities.
Additionally, the fact that those writers judged to have produced writing of higher
proficiency by the CAEL test evaluators were also found to have produced more
‘original’ lexical bundles than the less proficient writers may be taken as a sign that this
feature of writing does indeed play an important role in the way writers structure their
discourse. Looking again at Appendix A, we can see that one of the scoring criteria for a
limited writer is “writing is not predictable”. While the HLC writers used less lexical
bundles in general than the LLC writers, the fact that they used more ‘originally’
produced bundles (i.e. not sourced from the reading articles) may have helped contribute
to a more predictable style of writing that was graded more favourably by the CAEL test
judges.

7 Conclusions

Having summarized the findings uncovered through the analysis of the identified
lexical bundles in each of the corpora, this section discusses some of the implications
these findings may hold. Implications to both teaching & learning, along with language
testing & assessment are given before a brief section stating final conclusions in relation
to this research.
7.1 Teaching and Learning Implications

Formulaic sequences have been identified by numerous researchers as an important part of the language native speakers/writers produce. Not only do their use aid in the production of quick and accurate discourse, they also help to improve comprehension and aid in the reading/understanding of texts (Wray, 2002). While primarily demonstrated as a tool used by native-speakers of a language, lexical bundles are an area of formulaic language that may prove to be particularly useful in ESL/EFL classrooms in years to come. While the lower-level writers in the current research were actually found to produce more formulaic language than higher-level writers, a closer investigation showed that the vast majority of the lexical bundles identified were actually sourced from the reading articles they were given prior to beginning their writing. Initial analysis showed that lower-level writers produced more lexical bundles overall, eliminating those bundles which also appeared in the reading articles/prompt however, showed that lower-level writing actually contained fewer ‘originally produced’ lexical bundles. This is an important finding that suggests lower-level non-native English writers have a limited resource of lexical bundles available to them. As Boers et al’s 2006 study has shown, using formulaic sequences as a teaching focus for non-native English speakers can improve overall language fluency. Increased use of formulaic sequences in their study was seen to improve fluency, provide greater range of expression, and improve accuracy when compared to the control group (as evaluated by outside judges). In a similar manner, the increased use of ‘originally’ produced lexical bundles in this study was correlated with higher proficiency levels (as evaluated by CAEL test judges).
The traditional approach to language teaching in the past has generally viewed syntax and grammatical knowledge as the keys to language learning, however there is growing evidence, suggested in part by usage-based models of language, that formulaic sequences have been undervalued under this approach. While the ability to create sentences “from scratch” will never be seen as unnecessary, it is also important to understand the important role formulaic language plays in the construction of discourse. In light of this knowledge, language classrooms should make attempts to incorporate activities and learning tasks that involve a greater focus on this aspect of language. As tools for both teachers and students alike, corpora can be a useful resource in language classrooms and can be used to help students better understand the role formulaic sequences play in academic discourse. The use of native English writer corpora, coupled with concordancing software programs can be used by students to identify important lexical bundles and see how they function within the text. Common usage patterns and functional distribution of lexical bundles within different corpora may also prove useful. Alternatively, by using non-native English writer corpora, students may be able to better identify deficiencies in their own writing as it relates to formulaic language and identify areas for improvement. Corpora made up of students’ own writing could be particularly advantageous in this respect as it may offer students a higher level of engagement with the material they are studying. Learner corpora of this type may also be useful for language learners themselves as a tool to investigate common patterns indicative of both low-level and high-level non-native English writing. Using these types of corpora in the classroom may help students to better identify weak areas in their writing, and contribute to a better understanding of the issues they are facing in their writing advancement. By
better understanding the errors they are making, these techniques may be able to help these students focus their efforts on improving aspects of their language that will create the greatest benefit to their overall development.

Without knowledge of the role formulaic language holds in academic writing, ESL/EFL students may perceive a greater freedom to construct sentences that are grammatical, yet perceived as non-native like when used. The use of corpora made up of actual language use may help to alleviate such problems.

### 7.2 Language Testing and Assessment Implications

The potential implications this research holds for language testing and assessment should also be mentioned. Seeing as how the corpora used in this study are made up of the written discourse produced for a language test (i.e. the CAEL test), studies of this kind may hold some influence on how such tests are evaluated in the future. Obvious differences in lexical bundle use were discovered through the investigation presented in the previous sections, and these differences could potentially be used to provide a clearer guide for judges evaluating future language tests. Potential alterations to grading rubrics may include the inclusion of common phrasings used by both high and low level writers, and other common distinguishing characteristics.

The use of computers and computer programs that help evaluate written discourse has also become more popular in recent years with various websites now offering detection tools to help identify plagiarized work, plagiarism.org, for example. Technology similar to that used to identify plagiarism may also be beneficial in language testing as a high proportion of the lexical bundles identified in the LLC seem to have been sourced from the reading articles. Using a computer program to search through test
taker writing and identifying all the lexical bundles that also appear in the reading articles could potentially be used to help evaluate test taker writing. Perhaps the two most striking differences between high-level and low-level writers in this study related to the length of lexical bundles used, and the percentage of these bundles that could be labeled 'originally produced'. Using a program to help identify each of these features in test taker writing may help to more accurately identify which scoring band each writer falls into. Since this is only a small scale-study however, a larger sample of writing would need to be used before implementing these measures.

7.3 Limitations of the present study

Although the compilation of the two corpora used in this study and the resulting analysis have yielded interesting and important results, this study is not without limitations. Perhaps most importantly, it should be noted that the two corpora used in this study, although similar in size, vary in terms of the number of texts of which they are composed. Ideally the two corpora used in comparisons of this type would not only have similar word counts, but also numbers of texts. Unfortunately, this was an unlikely feat given the nature of the data. On average, lower-level test takers in this study tended to write shorter essays in response to the prompt they were given (223 words per essay in the LLC Vs. 298 words per essay in the HLC) and it was therefore necessary to include more lower-level writing samples so that similar word counts between the two corpora could be achieved.

Additionally, the relatively small size of the two corpora used in this study (around 75,000 words each) is diminutive in comparison to many of the large scale corpora used by other researchers in the investigation of lexical bundles. For this reason,
the present study may be seen as more of a pilot study that has revealed areas in need of further exploration and not a complete assessment of the writing produced by non-native English writers. Corpora of greater sizes can be viewed as providing greater reliability in terms of their results, as corpora of greater sizes tend to reveal trends and statistics concerning a larger population of writers. While large-scale corpora are often seen as an ideal to which researchers should strive to achieve, primary concern should be given to ensuring the writing being included fit the sampling frame defined by the researcher(s). Given the strict sampling frame used in the present study, and the fact that there were no publicly available corpora that could adequately address the research questions, the creation of large-scale corpora that could adhere to the sampling frame was impossible to achieve during the research period.

It should also be noted that the functional taxonomy of the lexical bundles identified in this study is a necessarily subjective process. While the identification of the lexical bundles themselves is based on statistical analysis, the classification of any identified bundles must be performed based on the researchers own intuition and interpretation of the bundles as seen through concordance lines (a problem seen in both small and large sized corpora). As a result, different researchers may interpret the functions of lexical bundles in different ways. As mentioned earlier, attempts were made to compensate for this limitation by having two separate researchers sort through the lexical bundles generated by WordSmith Tools 5.0, while this was an important step toward achieving a higher level of internal reliability, it is unlikely all researchers would classify these bundles in exactly the same manner. Evidence of this problem can be seen in the work of other researchers who have given functional classifications of lexical
bundles in their own studies. For example, 'one of the most' is a lexical bundle identified in both Biber et al. (2004) and Chen & Baker (2010), however each of these studies have assigned different functional roles to this bundle. In Biber et al. (2004), 'one of the most' is classified as a referential bundle serving an identification/focus function, in contrast, Chen and Baker (2010) label this bundle as a discourse-organizer serving the same identification/focus function. While these two studies did use slightly different functional sub-categories to classify the lexical bundles they identified, they both chose to implement the same three broad categories common to research of this type in their analysis: referential, discourse-organizing, and stance. Although this is an issue that may affect the ability of researchers to compare their findings with the work of others, in general this problem does not present much of an issue since 'problem' bundles that may be may easily be misinterpreted, or interpreted differently by different researchers, generally make up a relatively small portion of the classified bundles. Additionally, the implementation of internal reliability measures such as that used in the present research (i.e. making use of multiple researchers to classify any identified bundles) can be used to help alleviate this problem.

7.4 Final Conclusions

Although some researchers have chosen to investigate different populations of L2 English writers based on their language backgrounds and mother tongue, L2 English speakers/writers are most often defined as a uniform group, so that comparisons can be

---

10 In my own study, this bundle was assigned a discourse organizing function, similar to Chen & Baker 2010. This did not affect my comparison to findings of Biber et al (2004) however, since the bundle was not common enough to appear in their final functional analysis which I used as a comparison to my own findings.
made to native-English speakers/writers discourse. The present study has chosen to move away from this trend by investigating the writing produced by two distinct populations of non-native English writers. By comparing the lexical bundles used by low-level and high-level non-native English writers in this study, we have been able to see how each of these populations uses lexical bundles to structure their discourse. Performing such an analysis has revealed numerous differences between these two populations and given potential areas where these findings may be applied to aid in the teaching and learning of writing for non-native English students. Formulaic sequences have previously been identified by researchers as an important area of language production that serves numerous advantages to the speaker/writer making use of these structures. Additionally, it has been suggested that these word sequences may also be beneficial to the hearer/reader of said structures. This second point is supported by in the analysis of lexical bundles provided in the preceding sections. While not listed as one of the criteria important in grading CAEL test taker writing, the differences in lexical bundles usage between the two groups of writers in this study suggests that lexical bundles are not only a significant aspect in the way writers structure their discourse, but that they may also play an important role in how evaluators judge the writing produced by non-native writers.
### Appendix A: Writing Performance Band Score Criteria

<table>
<thead>
<tr>
<th>Band Score</th>
<th>Score Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-20</td>
<td><strong>Very Limited Writer:</strong>&lt;br&gt;Is generally unable to express ideas effectively&lt;br&gt;Uses very restricted and/or ungrammatical language&lt;br&gt;Uses words randomly and without overall coherence</td>
</tr>
<tr>
<td>30</td>
<td><strong>Limited Writer:</strong>&lt;br&gt;Attempts to write something which is related to the topic but the writing is not predictable&lt;br&gt;Uses restricted and/or ungrammatical language&lt;br&gt;Seems to understand the topic, but is unable to develop ideas because language constrains or distorts expression</td>
</tr>
<tr>
<td>40</td>
<td><strong>Marginally Competent Writer:</strong>&lt;br&gt;Makes links among ideas and addresses the topic but the writing lacks clarity and cohesiveness&lt;br&gt;Displays elements of control in the writing (e.g. a thesis statement, an introduction and conclusion) but internal coherence is lacking&lt;br&gt;Uses little or no support (i.e. quotations, examples, etc.) to develop the thesis</td>
</tr>
<tr>
<td>50</td>
<td><strong>Competent but Limited Writer:</strong>&lt;br&gt;Addresses the topic to a degree but with somewhat limited clarity and cohesiveness&lt;br&gt;Uses some support to develop the thesis&lt;br&gt;Control of the argument is limited by poor comprehension of the readings and lecture, and/or poor understanding of the requirements of academic writing</td>
</tr>
<tr>
<td>60</td>
<td><strong>Competent Writer:</strong>&lt;br&gt;Develops a thesis using a range of support&lt;br&gt;Uses language that is generally accurate but is constrained by a somewhat limited vocabulary&lt;br&gt;Demonstrates general understand of the requirements of academic writing</td>
</tr>
<tr>
<td>70</td>
<td><strong>Adept Writer:</strong>&lt;br&gt;Responds readily to the demands of the topic and presents information clearly and logically&lt;br&gt;Uses the readings and lecture effectively to support the thesis&lt;br&gt;Demonstrated understand of the requirements of academic writing</td>
</tr>
<tr>
<td>80-90</td>
<td><strong>Expert Writer:</strong>&lt;br&gt;Demonstrates mastery of appropriate, concise, and persuasive academic writing&lt;br&gt;Writes with authority and style</td>
</tr>
</tbody>
</table>
Appendix B: 4-word lexical bundles – Lower Level Corpus

<table>
<thead>
<tr>
<th>N</th>
<th>Word</th>
<th>Frequency</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A SERIOUS THREAT TO</td>
<td>243</td>
<td>165</td>
</tr>
<tr>
<td>2</td>
<td>TO LIFE ON EARTH</td>
<td>225</td>
<td>162</td>
</tr>
<tr>
<td>3</td>
<td>SERIOUS THREAT TO LIFE</td>
<td>214</td>
<td>149</td>
</tr>
<tr>
<td>4</td>
<td>THREAT TO LIFE ON</td>
<td>214</td>
<td>153</td>
</tr>
<tr>
<td>5</td>
<td>PRESENT A SERIOUS THREAT</td>
<td>169</td>
<td>123</td>
</tr>
<tr>
<td>6</td>
<td>FARM REDUCTION PRESENT A</td>
<td>141</td>
<td>110</td>
</tr>
<tr>
<td>7</td>
<td>REDUCTION PRESENT A SERIOUS</td>
<td>137</td>
<td>106</td>
</tr>
<tr>
<td>8</td>
<td>THE FARM LAND IS</td>
<td>134</td>
<td>93</td>
</tr>
<tr>
<td>9</td>
<td>IS ONE OF THE</td>
<td>61</td>
<td>53</td>
</tr>
<tr>
<td>10</td>
<td>TO THE FARM LAND</td>
<td>56</td>
<td>43</td>
</tr>
<tr>
<td>11</td>
<td>PLACE IN THE FARM</td>
<td>47</td>
<td>44</td>
</tr>
<tr>
<td>12</td>
<td>IN THE SUN'S RAYS</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>13</td>
<td>IN THE FARM LAND</td>
<td>44</td>
<td>41</td>
</tr>
<tr>
<td>14</td>
<td>OF THE ULTRAVIOLET RADIATION</td>
<td>44</td>
<td>42</td>
</tr>
<tr>
<td>15</td>
<td>SKIN CANCER IN HUMANS</td>
<td>44</td>
<td>43</td>
</tr>
<tr>
<td>16</td>
<td>THREAT TO THE FARM</td>
<td>44</td>
<td>36</td>
</tr>
<tr>
<td>17</td>
<td>OF THE FARM LAND</td>
<td>43</td>
<td>36</td>
</tr>
<tr>
<td>18</td>
<td>MOST OF THE ULTRAVIOLET</td>
<td>42</td>
<td>41</td>
</tr>
<tr>
<td>19</td>
<td>FARM REDUCTION IS A</td>
<td>39</td>
<td>33</td>
</tr>
<tr>
<td>20</td>
<td>FARM LAND IS A</td>
<td>39</td>
<td>35</td>
</tr>
<tr>
<td>21</td>
<td>THE FARM REDUCTION PRESENT</td>
<td>39</td>
<td>34</td>
</tr>
<tr>
<td>22</td>
<td>THE RELEASE OF PESTICIDES</td>
<td>39</td>
<td>37</td>
</tr>
<tr>
<td>23</td>
<td>INCIDENCE OF SKIN CANCER</td>
<td>38</td>
<td>36</td>
</tr>
<tr>
<td>24</td>
<td>FOUND IN THE SUN'S</td>
<td>37</td>
<td>37</td>
</tr>
<tr>
<td>25</td>
<td>OF SKIN CANCER IN</td>
<td>36</td>
<td>35</td>
</tr>
<tr>
<td>26</td>
<td>THE INCIDENCE OF SKIN</td>
<td>36</td>
<td>34</td>
</tr>
<tr>
<td>27</td>
<td>THE FARM REDUCTION IS</td>
<td>36</td>
<td>32</td>
</tr>
<tr>
<td>28</td>
<td>ABSORBING MOST OF THE</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>29</td>
<td>BY ABSORBING MOST OF THE</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>30</td>
<td>ULTRAVIOLET RADIATION FOUND IN</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>31</td>
<td>RADIATION FOUND IN THE</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>32</td>
<td>SERIOUS THREAT TO THE</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>33</td>
<td>REDUCTION OF CROP YIELDS</td>
<td>31</td>
<td>28</td>
</tr>
<tr>
<td>34</td>
<td>BY THE RELEASE OF</td>
<td>30</td>
<td>29</td>
</tr>
<tr>
<td>35</td>
<td>IN THE INCIDENCE OF</td>
<td>30</td>
<td>29</td>
</tr>
<tr>
<td>36</td>
<td>OF CROP YIELDS OF</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>37</td>
<td>THE LIFE ON EARTH</td>
<td>29</td>
<td>25</td>
</tr>
<tr>
<td>38</td>
<td>THE ULTRAVIOLET RADIATION FOUND</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>39</td>
<td>FOOD CHAIN IN THE</td>
<td>28</td>
<td>24</td>
</tr>
<tr>
<td>40</td>
<td>INCREASE IN THE INCIDENCE</td>
<td>28</td>
<td>27</td>
</tr>
<tr>
<td>41</td>
<td>LIFE ON THE EARTH</td>
<td>27</td>
<td>24</td>
</tr>
<tr>
<td>42</td>
<td>THE FORESTS AND EARTH'S</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>43</td>
<td>FORESTS AND EARTH'S SURFACE</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>44</td>
<td>ONE OF THE PREDICTED</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>45</td>
<td>FARM LAND IS PRESENTD</td>
<td>25</td>
<td>24</td>
</tr>
<tr>
<td>46</td>
<td>SURFACE BY ABSORBING MOST</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>
47 CANCER IN HUMANS IS 24 24
48 CROP YIELDS OF VARIOUS 24 23
49 HUMANS IS ONE OF 24 24
50 IS A SERIOUS THREAT 24 23
51 IS PRESENTD BY THE 24 23
52 LAND IS PRESENTD BY 24 23
53 OF AN OXYGEN MOLECULE 24 24
54 PRESENTD BY THE RELEASE 24 23
55 THAT FARM REDUCTION PRESENT 24 22
56 IN HUMANS IS ONE 23 23
57 ONE OF THE MOST 23 20
58 SERVES AS A SHIELD 23 23
59 CONTAINING CARBON FLUORINE AND 22 22
60 REDUCTION IS A SERIOUS 22 20
61 OF THE PREDICTED EFFECTS 22 22
62 FARM LAND SERVES AS 22 22
63 THAT THE FARM REDUCTION 22 16
64 YIELDS OF VARIOUS PLANTS 22 21
65 ACCORDING TO THE ARTICLE 21 17
66 EARTH'S SURFACE BY ABSORBING 21 21
67 FORM OF AN OXYGEN 21 21
68 IS A REGION OF 21 21
69 LAND SERVES AS A 21 21
70 PROTECTING THE FORESTS AND 21 21
71 SYNTHETIC COMPOUNDS CONTAINING CARBON 21 21
72 A REGION OF CONCENTRATION 20 20
73 BY THE ACTION OF 20 20
74 CHAIN IN THE SEAS 20 19
75 COMPOUNDS CONTAINING CARBON FLUORINE 20 20
76 REGION OF CONCENTRATION OF 20 20
77 THE ACTION OF SOLAR 20 20
78 A SHIELD PROTECTING THE 19 19
79 A SHIELD TO EARTHLY 19 14
80 ACCORDING TO THE READING 19 14
81 ACTION OF SOLAR RADIATION 19 19
82 DOES FARM REDUCTION PRESENT 19 19
83 PESTICIDES WHICH ARE SYNTHETIC 19 19
84 ON THE OTHER HAND 19 17
85 PRODUCED BY THE ACTION 19 19
86 SHIELD TO EARTHLY LIFE 19 14
87 SOLAR RADIATION ON ORDINARY 19 19
88 THE FARM LAND SERVES 19 19
89 A MARKED INCREASE IN 18 18
90 AS A SHIELD PROTECTING 18 18
91 IF THE CONSUMPTION OF 18 18
92 OF SOLAR RADIATION ON 18 18
93 PRESENTS A SERIOUS THREAT 18 16
94 WHICH ARE SYNTHETIC COMPOUNDS 18 18
95 ARE SYNTHETIC COMPOUNDS CONTAINING 17 17
MARKED INCREASE IN THE
OF CONCENTRATION OF THE
OF VARIOUS PLANTS AND
FARM REDUCTION PRESENTS A
RADIATION ON ORDINARY OXYGEN
RELEASE OF PESTICIDES WHICH
THE EARTH FROM THE
THE FORM OF AN
ACCORDING TO READING #
AND EARTH'S SURFACE BY
CERTAIN FORMS OF AQUATIC
CONSUMPTION OF CFCS CONTINUES
FORMS OF AQUATIC LIFE
OF CERTAIN FORMS OF
OF THE FORM OF
SHIELD PROTECTING THE FORESTS
### Appendix C: 4-word lexical bundles – Higher Level Corpus

<table>
<thead>
<tr>
<th>N</th>
<th>Word</th>
<th>Freq.</th>
<th>Texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TO LIFE ON EARTH</td>
<td>185</td>
<td>130</td>
</tr>
<tr>
<td>2</td>
<td>A SERIOUS THREAT TO</td>
<td>177</td>
<td>131</td>
</tr>
<tr>
<td>3</td>
<td>THREAT TO LIFE ON</td>
<td>165</td>
<td>122</td>
</tr>
<tr>
<td>4</td>
<td>SERIOUS THREAT TO LIFE</td>
<td>149</td>
<td>112</td>
</tr>
<tr>
<td>5</td>
<td>THE FARM LAND IS</td>
<td>113</td>
<td>79</td>
</tr>
<tr>
<td>6</td>
<td>OF THE FARM LAND</td>
<td>106</td>
<td>64</td>
</tr>
<tr>
<td>7</td>
<td>PRESENT A SERIOUS THREAT</td>
<td>94</td>
<td>78</td>
</tr>
<tr>
<td>8</td>
<td>REDUCTION PRESENT A SERIOUS</td>
<td>58</td>
<td>48</td>
</tr>
<tr>
<td>9</td>
<td>FARM REDUCTION PRESENT A</td>
<td>58</td>
<td>48</td>
</tr>
<tr>
<td>10</td>
<td>THE LIFE ON EARTH</td>
<td>54</td>
<td>37</td>
</tr>
<tr>
<td>11</td>
<td>REDUCTION OF THE FARM</td>
<td>51</td>
<td>34</td>
</tr>
<tr>
<td>12</td>
<td>IN THE FARM LAND</td>
<td>44</td>
<td>32</td>
</tr>
<tr>
<td>13</td>
<td>THE REDUCTION OF THE</td>
<td>43</td>
<td>30</td>
</tr>
<tr>
<td>14</td>
<td>PRESENTS A SERIOUS THREAT</td>
<td>40</td>
<td>36</td>
</tr>
<tr>
<td>15</td>
<td>TO THE FARM LAND</td>
<td>40</td>
<td>31</td>
</tr>
<tr>
<td>16</td>
<td>FOOD CHAIN IN THE</td>
<td>34</td>
<td>32</td>
</tr>
<tr>
<td>17</td>
<td>REDUCTION OF FARM LAND</td>
<td>33</td>
<td>19</td>
</tr>
<tr>
<td>18</td>
<td>REDUCTION PRESENTS A SERIOUS</td>
<td>33</td>
<td>30</td>
</tr>
<tr>
<td>19</td>
<td>FARM REDUCTION IS A</td>
<td>33</td>
<td>25</td>
</tr>
<tr>
<td>20</td>
<td>FARM LAND IS A</td>
<td>33</td>
<td>32</td>
</tr>
<tr>
<td>21</td>
<td>FARM REDUCTION PRESENTS A</td>
<td>32</td>
<td>28</td>
</tr>
<tr>
<td>22</td>
<td>PLACE IN THE FARM</td>
<td>31</td>
<td>25</td>
</tr>
<tr>
<td>23</td>
<td>THE REDUCTION OF FARM</td>
<td>31</td>
<td>24</td>
</tr>
<tr>
<td>24</td>
<td>THE FARM REDUCTION IS</td>
<td>31</td>
<td>24</td>
</tr>
<tr>
<td>25</td>
<td>MOST OF THE ULTRAVIOLET</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>26</td>
<td>OF THE ULTRAVIOLET RADIATION</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>27</td>
<td>TO THE LIFE ON</td>
<td>29</td>
<td>24</td>
</tr>
<tr>
<td>28</td>
<td>REDUCTION OF CROP YIELDS</td>
<td>28</td>
<td>25</td>
</tr>
<tr>
<td>29</td>
<td>SKIN CANCER IN HUMANS</td>
<td>28</td>
<td>27</td>
</tr>
<tr>
<td>30</td>
<td>ABSORBING MOST OF THE</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>31</td>
<td>CHAIN IN THE SEAS</td>
<td>27</td>
<td>25</td>
</tr>
<tr>
<td>32</td>
<td>SERIOUS THREAT TO THE</td>
<td>27</td>
<td>25</td>
</tr>
<tr>
<td>33</td>
<td>THE FARM LAND AND</td>
<td>27</td>
<td>23</td>
</tr>
<tr>
<td>34</td>
<td>BY ABSORBING MOST OF</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>35</td>
<td>OF THE FARM REDUCTION</td>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>36</td>
<td>FORMS OF AQUATIC LIFE</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>37</td>
<td>LIFE ON THE EARTH</td>
<td>25</td>
<td>21</td>
</tr>
<tr>
<td>38</td>
<td>CROP YIELDS OF VARIOUS</td>
<td>23</td>
<td>22</td>
</tr>
<tr>
<td>39</td>
<td>IN THE SUN'S RAYS</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>40</td>
<td>INCIDENCE OF SKIN CANCER</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>41</td>
<td>IS A SERIOUS THREAT</td>
<td>23</td>
<td>20</td>
</tr>
<tr>
<td>42</td>
<td>IS ONE OF THE</td>
<td>23</td>
<td>22</td>
</tr>
<tr>
<td>43</td>
<td>THAT THE FARM REDUCTION</td>
<td>23</td>
<td>20</td>
</tr>
<tr>
<td>44</td>
<td>YIELDS OF VARIOUS PLANTS</td>
<td>23</td>
<td>22</td>
</tr>
<tr>
<td>45</td>
<td>OF SKIN CANCER IN</td>
<td>22</td>
<td>22</td>
</tr>
</tbody>
</table>
THREAT TO THE FARM
CERTAIN FORMS OF AQUATIC
OF FARM LAND IS
THAT THE FARM LAND
THE INCIDENCE OF SKIN
THREAT TO THE LIFE
FOUND IN THE SUN'S
SURFACE BY ABSORBING MOST
ULTRAVIOLET RADIATION FOUND IN
does present a serious
OF CERTAIN FORMS OF
ONE OF THE MOST
FARM LAND IS THE
THE ULTRAVIOLET RADIATION FOUND
EARTH'S SURFACE BY ABSORBING
KILLING OF CERTAIN FORMS
SERVES AS A SHIELD
THE PLACE IN THE
THE LOSS OF FARM
FORESTS AND EARTH'S SURFACE
## Appendix D: 4-7 word lexical bundles – Lower Level Corpus

<table>
<thead>
<tr>
<th>N</th>
<th>BUNDLE</th>
<th>FREQ</th>
<th>RANGE</th>
<th>FREQ AS 5-WORD</th>
<th>FREQ AS 6-WORD</th>
<th>FREQ AS 7-WORD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A SERIOUS THREAT TO</td>
<td>243</td>
<td>165</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>TO LIFE ON EARTH</td>
<td>225</td>
<td>162</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>SERIOUS THREAT TO LIFE</td>
<td>214</td>
<td>149</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>THREAT TO LIFE ON</td>
<td>214</td>
<td>153</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>PRESENT A SERIOUS THREAT</td>
<td>169</td>
<td>123</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>FARM REDUCTION PRESENT A SERIOUS</td>
<td>141</td>
<td>110</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>REDUCTION PRESENT A SERIOUS</td>
<td>137</td>
<td>106</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>THE FARM LAND IS</td>
<td>134</td>
<td>93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>IS ONE OF THE</td>
<td>61</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>TO THE FARM LAND</td>
<td>56</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>PLACE IN THE FARM</td>
<td>47</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>IN THE SUN'S RAYS</td>
<td>45</td>
<td>45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>IN THE FARM LAND</td>
<td>44</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>MOST OF THE ULTRAVIOLET RADIATION</td>
<td>44/42</td>
<td>42/41</td>
<td>39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>SKIN CANCER IN HUMANS</td>
<td>44</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>THREAT TO THE FARM</td>
<td>44</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>OF THE FARM LAND</td>
<td>43</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>FARM REDUCTION IS A</td>
<td>39</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>FARM LAND IS A</td>
<td>39</td>
<td>35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>THE FARM REDUCTION PRESENT</td>
<td>39</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>THE RELEASE OF PESTICIDES</td>
<td>39</td>
<td>37</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>THE INCIDENCE OF SKIN CANCER IN</td>
<td>38/36/36</td>
<td>36/35/34</td>
<td>33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>FOUND IN THE SUN'S</td>
<td>37</td>
<td>37</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>THE FARM REDUCTION IS</td>
<td>36</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>BY ABSORBING MOST OF THE ULTRAVIOLET RADIATION</td>
<td>34/34</td>
<td>34/34</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>THE ULTRAVIOLET RADIATION FOUND IN THE</td>
<td>34/32/29</td>
<td>34/32/29</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>SERIOUS THREAT TO THE Way</td>
<td>32</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>REDUCTION OF CROP YIELDS OF THE</td>
<td>31/29</td>
<td>28/28</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>BY THE RELEASE OF</td>
<td>30</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>INCREASE IN THE INCIDENCE OF SKIN CANCER IN</td>
<td>30/28</td>
<td>29/27</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>THE LIFE ON EARTH</td>
<td>29</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>LIFE ON THE EARTH</td>
<td>27</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>FOOD CHAIN IN THE</td>
<td>28</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>THE FORESTS AND EARTHS SURFACE</td>
<td>26/26</td>
<td>26/26</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>SURFACE BY ABSORBING MOST</td>
<td>25</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>CANCER IN HUMANS IS ONE OF THE</td>
<td>24/24/23</td>
<td>24/24/23</td>
<td>23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Page</td>
<td>Text</td>
<td>Numbers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>CROP YIELDS OF VARIOUS PLANTS IS A SERIOUS THREAT</td>
<td>24/22 23/21 22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>FARM LAND IS PRESENTED BY THE RELEASE</td>
<td>24 23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>FORM OF AN OXYGEN MOLECULE</td>
<td>24/21 24/21 20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>THAT FARM REDUCTION PRESENT ONE OF THE MOST</td>
<td>24 22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>FARM LAND SERVES AS A SHIELD REDUCTION IS A SERIOUS</td>
<td>23 20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>ONE OF THE PREDICTED EFFECTS</td>
<td>22 22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>THAT THE FARM REDUCTION ACCORDING TO THE ARTICLE</td>
<td>22 16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>REDUCTION PRESENTS A SERIOUS THREAT PRESENTS A</td>
<td>21 17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>EARTH'S SURFACE BY ABSORBING IS A REGION OF CONCENTRATION OF</td>
<td>21 21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>PROTECTING THE FORESTS AND SYNTHETIC COMPOUNDS CONTAINING CARBON FLUORINE AND PRODUCED BY THE ACTION OF SOLAR RADIATION</td>
<td>21/20/20 21/20/20 18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>CHAIN IN THE SEAS AS A SHIELD PROTECTING THE FORESTS</td>
<td>20 19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>A SHIELD TO EARTHLY LIFE ACCORDING TO THE READING</td>
<td>19/16 19/18/16 15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>DOES FARM REDUCTION PRESENT PESTICIDES WHICH ARE SYNTHETIC COMPOUNDS CONTAINING</td>
<td>19/18/17 19/18/17 15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>ON THE OTHER HAND</td>
<td>19 17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>OF SOLAR RADIATION ON SOLAR RADIATION ON ORDINARY OXYGEN</td>
<td>18 18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>THE FARM LAND SERVES</td>
<td>19 19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>A MARKED INCREASE IN THE IF THE CONSUMPTION OF</td>
<td>18/17 18/17 17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>PRESENTS A SERIOUS THREAT</td>
<td>18 16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>OF VARIOUS PLANTS AND</td>
<td>17 16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>FARM REDUCTION PRESENT A</td>
<td>17 15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>RELEASE OF PESTICIDES WHICH ARE THE EARTH FROM THE PROTECT THE EARTH FROM OF CONCENTRATION OF THE FORM OF AN ACCORDING TO READING # AND EARTH'S SURFACE BY KILLING OF CERTAIN FORMS OF AQUATIC LIFE CONSUMPTION OF CFCS CONTINUES AT</td>
<td>17/15</td>
<td>17/15</td>
<td>13</td>
<td>17</td>
<td>17</td>
</tr>
</tbody>
</table>
### Appendix E: 4-7 word lexical bundles – Higher Level Corpus

<table>
<thead>
<tr>
<th>N</th>
<th>BUNDLE</th>
<th>FREQ</th>
<th>RANGE</th>
<th>FREQ AS 5-WORD</th>
<th>FREQ AS 6-WORD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TO LIFE ON EARTH</td>
<td>185</td>
<td>130</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>A SERIOUS THREAT TO</td>
<td>177</td>
<td>131</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>THREAT TO LIFE ON</td>
<td>165</td>
<td>122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SERIOUS THREAT TO LIFE</td>
<td>149</td>
<td>112</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>THE FARM LAND IS</td>
<td>113</td>
<td>79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>OF THE FARM LAND</td>
<td>106</td>
<td>64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>PRESENT A SERIOUS THREAT</td>
<td>94</td>
<td>78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>FARM REDUCTION PRESENT A SERIOUS</td>
<td>58/58</td>
<td>48/48</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>THE LIFE ON EARTH</td>
<td>54</td>
<td>37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>REDUCTION OF THE FARM</td>
<td>51</td>
<td>34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>IN THE FARM LAND</td>
<td>44</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>THE REDUCTION OF THE</td>
<td>43</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>PRESENTS A SERIOUS THREAT</td>
<td>40</td>
<td>36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>TO THE FARM LAND</td>
<td>40</td>
<td>31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>FARM REDUCTION IS A</td>
<td>39</td>
<td>33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>THE FARM REDUCTION IS</td>
<td>36</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>FOOD CHAIN IN THE</td>
<td>34</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>REDUCTION OF FARM LAND</td>
<td>33</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>FARM REDUCTION PRESENTS A SERIOUS</td>
<td>33/32</td>
<td>30/28</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>FARM LAND IS A</td>
<td>33</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>PLACE IN THE FARM</td>
<td>31</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>THE REDUCTION OF FARM</td>
<td>31</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>MOST OF THE ULTRAVIOLET RADIATION</td>
<td>30/30</td>
<td>30/30</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>TO THE LIFE ON</td>
<td>29</td>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>REDUCTION OF CROP YIELDS</td>
<td>28</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>SKIN CANCER IN HUMANS</td>
<td>28</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>BY ABSORBING MOST OF THE</td>
<td>27/26</td>
<td>27/26</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>CHAIN IN THE SEAS</td>
<td>27</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>SERIOUS THREAT TO THE</td>
<td>27</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>THE FARM LAND AND OF THE FARM REDUCTION CERTAIN FORMS OF AQUATIC LIFE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>27</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>26</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>25/21</td>
<td>25/21</td>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>25</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>23/22</td>
<td>23/22</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>23</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>23/22/21</td>
<td>23/22/21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>23</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>23</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>23</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>22</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>21</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>21</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>21</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>20</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>19</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>18/17</td>
<td>18/17</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>17</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>17/16</td>
<td>17/16</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>17</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>17</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>16</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>16</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>16</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>16</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>16</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References


