# LEXICAL BUNDLES IN L2 ENGLISH ACADEMIC WRITING

## Proficiency level differences

By Randy Appel, Concordia University

**Abstract**

Research on formulaic sequences, frequently occurring multiword units, has seen significant growth in recent decades. However, relatively few studies have focused on how non-native English users make use of formulaic sequences in their academic English writing and how these structures contribute to assessments of linguistic ability. In order to better understand how second language English writers make use of this aspect of language, a collection of argumentative essays written by test takers of the Canadian Academic English Language Assessment (CAEL) was analyzed for the use of lexical bundles, a frequency driven sub-category of formulaic sequences. Dividing this collection of writing into three corpora, based on assessed proficiency, revealed marked differences in how lexical bundles were used by each group of writers. Implications for the teaching and assessment of second language English academic writing are discussed.

Second language (L2) learners face a seemingly monolithic task when confronted with the challenge of attaining a high-level of proficiency in their target language. Not only do L2 learners need to develop general-level abilities related to the processing and production of their L2, but they are also frequently required to develop genre- and register-specific skills. These skills include, among others, knowledge of discipline specific jargon, culturally appropriate organizational patterns, and conventionalized forms of expression that may vary substantially from one genre or register to the next. As L2 learners progress, they quickly realize that their task is not simply to acquire one version of the target language, but multiple versions that all serve different purposes (i.e., conversational English, written English, academic English, business English). If L2 English learners are to be successful in their goals, they need to attain an adequate level of proficiency in each of the target genres and registers they will be using to communicate.

One particular genre that merits special attention for many L2 English learners is academic writing. Due to the fact that the academic success, and therefore eventual career aspirations, of many L2 English learners is reliant on their ability to effectively communicate in written academic English, high proficiency in this area has become an important goal for many L2 English learners.

Additionally, with the continued internationalization of postsecondary institutions in Canada expected to result in increased numbers of L2 English learners (AUCC, 2010), there is a growing need for quality instruction of academic English writing by English as a Second Language (ESL) teachers. Unfortunately, while the importance of academic writing for ESL teachers and learners continues to grow, our understanding of this genre of English and the factors that lead to differences in perceived proficiency within it, particularly from the perspective of multi-word utterances, remains underdeveloped.

Corpus research, and the ability to effectively analyze large collections of L2 discourse, presents a valuable tool that can be used to provide new insights and further our understanding of L2 English academic writing from various perspectives. Although corpora of learner English are a relatively new addition to applied linguistics research (Granger, 1998), corpus-informed studies have seen substantial growth in recent years and are now used to help study first language (L1) and L2 discourse in a variety of genres and registers (e.g., Granger & Rayson, 1998; Virtanen, 1998). While these and other related studies have helped to identify specific points of differentiation between L1 and L2 writers, comparatively few research endeavours have attempted to focus on differences within populations of L2 writers possessing differing levels of linguistic ability.

The present study aims to further understanding of L2 English academic writing by using a corpus-informed approach to analyse L2 English academic writing by test takers of the Canadian Academic English Language (CAEL) assessment. Since the use of conventionalized expressions, otherwise known as formulaic sequences, has been identified as an important aspect of English language ability (Bamber, 1983; Boers Eyckmans, Kappel, Stengers & Demecheleer, 2006; McCully, 1985; Pawley & Syder, 1983, Wray, 2002), these multi-word sequences were targeted as a way of better understanding differences in perceived proficiency in L2 English academic writing.

## Formulaic Sequences

Dating back to at least Firth (1935), the study of formulaic sequences (FSs) holds a long history within linguistic inquiry. Defined as prefabricated sequences that are "stored and retrieved whole from memory at the time of use" (Wray, 2002, p. 9), FSs are considered an important part of native language users' linguistic competence that help facilitate production of quick and accurate discourse (Nattinger & DeCarrico, 1992; Pawley & Syder, 1983, Wray, 2002). From the perspective of the listener or reader, experimental research has demonstrated important advantages associated with the use of formulaic language that can be linked to improved processing speed. For example, Underwood, Schmitt, and Galpin (2004) investigated the number and duration of eye fixations during reading tasks. Comparing time spent fixated on terminal words in formulaic and non-formulaic contexts, it was found that both L1 and advanced L2 participants spent less time fixated on terminal words in FSs than when the same words appeared in non-formulaic contexts.

With the importance of FSs increasingly recognized, attempts to better understand this aspect of language, and identify how it is used in various settings, have begun to grow. However, due to the fact that FSs can come in many different forms (e.g., idioms, collocations,

proverbs), the identification of FSs is an on-going challenge with no single solution. As a result, numerous methods of identification have been introduced. For instance, in oral discourse, phonological coherence, intonation contour, and speed of delivery have all been used to indicate formulaic status (Altenberg & Eeeg-Olofsson, 1990; Bybee & Scheibman, 1999; Kuiper, 1996). Similarly, for written discourse, multiple methods of identification are also available; however, a quantitative approach based primarily on minimum frequency and range requirements has emerged as the leading indicator of formulaic status in many corpus-driven studies. This method, originally developed by Biber, Conrad, and Reppen (1999) is labelled the lexical bundle approach.

## Lexical Bundles

Defined simply as "the most frequently recurring sequences of words" (Biber & Barbieri, 2007, p. 264), the term lexical bundle refers to a specific subset of formulaic language that is statistically defined on the basis of adherence to minimum frequency and range criteria. While lexical bundles occur in multiple word lengths, research using this methodology generally focuses on the identification of four-word sequences (e.g., Biber & Barbieri, 2007; Biber, Conrad, & Reppen, 1999; Chen, 2008; Chen & Baker, 2010; Cortes, 2004; Hyland, 2008). This focus on four-word structures is based in the observation that shorter sequences are contained within four-word lexical bundles (Cortes, 2004), and that they offer a greater range of functional roles that are more easily identifiable than three-word sequences (Hyland, 2008).

In terms of identification, two main criteria are used in the lexical bundle approach: frequency and range. Frequency, refers to the number of occurrences of a particular structure within the collection of discourse being analysed. Although minimum frequency cut-offs are often viewed as arbitrary (Hyland, 2008) and necessarily influenced by the size and specificity of the corpus being analysed, a common trend has emerged that sets minimum frequency for the identification of four-word sequences at 20–25 occurrences per million words (Adel & Erman, 2012; Chen, 2008; Chen & Baker, 2010; Cortes, 2004, 2008; Hyland, 2008).

Range, the second main criterion, is used to measure the spread of identified sequences within the corpus as a whole. Here, the goal is to ensure identified structures are not confined to a limited number of writers, or a limited number of texts. Consequently, range is used to reduce the chance of including idiosyncratic tendencies of individual or small groups of writers in the resulting list of lexical bundles. Range requirements also vary from study to study, yet a minimum of five texts has emerged as a common trend (Biber, Conrad, Reppen, 2004; Chen, 2010; Cortes, 2004). However, given that each study implements a differing number of texts in the analysis, it may prove more beneficial to use a percentage based approach (i.e., Hyland, 2008).

### Research Questions

Although existing research has examined the use of lexical bundles within various genres and registers, research across proficiency levels is lacking. In order to better understand how L2 English users of differing proficiency levels make use of this aspect of language in their academic writing, the present study targeted lexical bundles as a way of investigating differences in L2 English academic writing ability. Comparisons along proficiency levels were performed using a combination of quantitative and qualitative methods to assess differences in lexical bundle use by writers of varying levels of L2 English academic writing ability. The main question guiding this research was:

1. How do L2 English writers of differing proficiency levels make use of lexical bundles in their academic writing?

Answers to this question will help provide a greater understanding of how L2 English users make use of lexical bundles in their academic writing and which lexical bundles may be beneficial to L2 English learners and teachers aiming to increase target language proficiency. Consequently, an additional goal of the current study was to develop a list of lexical bundles that could be used in English for Academic Purposes (EAP) programs to help learners become more proficient academic writers. In order to analyse potential correlations between lexical bundle use and L2 English academic writing ability, the archives of the Canadian Academic English Language (CAEL) assessment were used as a data source.

### Canadian Academic English Language Assessment (CAEL)

CAEL is an integrated, topic-based test that evaluates L2 English learners' ability to use academic English as it is used in Canadian post-secondary institutions (CAEL, 2011). The test is frequently taken by L2 English learners aiming to gain entrance into English-medium universities and colleges throughout Canada. Composed of four sections (reading response, lecture response, oral language response, and written response), each test version focuses on one central theme or topic. In the final section of the test, the written response, test takers are allocated 45 minutes to handwrite an argumentative essay of between one and three pages that addresses a given prompt. During the composition of this essay, test takers are able to draw on information from two related readings and any previously encountered material to help strengthen the position they take.

Grading of the written response takes place during group marking sessions using a collaborative read-aloud protocol. These sessions consists of one rater who is assigned to read the essay aloud while two others listen and take notes. Once the essay has been read in its entirety, the three raters independently decide on a grade for the essay using one of nine possible grading bands (Appendix A). After raters independently decide on a grade, these scores are revealed to the group. If all raters are in agreement, the score is finalized. If independently assigned scores differ, a discussion of the reasoning for each score is initiated and continues until a consensus has been reached.

## Corpora

The main corpus used in this study is composed of argumentative essays written by test takers of the CAEL assessment. All essays used in this study were written in response to the same prompt, and therefore focus on the same main subject matter. By keeping essay type and topic stable, a better evaluation of proficiency level differences was possible since these potentially confounding factors had already been controlled for.

Essays used in this study were divided into three main corpora based on the assigned score each essay received. The first corpus, the Lower Level Corpus (LLC) is composed of essays that received a grade of either 20 or 30; their authors are therefore considered limited or very limited writers. The second corpus, the Medium Level Corpus (MLC), is composed of essays that received a grade of either 40 or 50; their authors are therefore evaluated as intermediate level writers. Finally, the third corpus, the High Level Corpus (HLC), is composed of essays rated between 60 and 90 on the CAEL grading scale; their authors are therefore considered upper-intermediate and advanced writers (full descriptions of each scoring band can be found in Appendix A). Although the lowest-level scoring band available on the CAEL assessment is 10, no essays at this level were included in the present study since these essays were often extremely short and contained extended passages that were directly copied from the reading articles included as part of the test. Descriptive statistics for each of the corpora are provided in Table 1.

Table 1: Descriptive Statistics

| | LLC | MLC | HLC | TOTAL |
|---|---|---|---|---|
| Words | 41,316 | 63,869 | 41,893 | 147,078 |
| Essays | 185 | 243 | 134 | 562 |
| Words/Essay | 223 | 263 | 313 | 262 |

## Methodology

Since all CAEL test essays are hand-written, the first step in preparing for analysis was to transcribe all essays so that they could be stored and analysed digitally. Once this step had been completed, it was necessary to decide on the length of lexical bundles that would be extracted. Although previous corpus driven research into the use of recurrent word sequences has primarily focused the identification of four-word lexical bundles (Chen, 2010; Cortes, 2004, Underwood, Schmitt, Galpin, 2004), the present study extended this focus to investigate the use of three- to five-word structures. This decision was made to more fully capture the complete range of formulaic language being used by each group of writers and investigated whether lexical bundle length could be related to assessed proficiency.

**Identification Procedure: three- to five-word lexical bundles**

Although a minimum frequency of occurrence in previous lexical bundle research has often been set at 20–25 occurrences per million words, the highly specialized nature of the corpora used in the present study necessitated substantial modification to previously used criteria. This was because the use of a high number of relatively small texts on the same topic was likely to result in greater overlap than the more standard book-length texts used in many previous studies. Therefore, in order to create manageable lists of lexical bundles that could be analysed in sufficient detail, the minimum frequency criterion in the present study was increased. For the LLC, the smallest corpus in the present study, frequency of occurrence was set at a minimum of nine instances. While this equates to relatively high 217 occurrences per million words, it was necessary in order to create a manageable list of sequences that could be used for subsequent analysis. As each corpus in the present study consisted of a differing number of total words, the minimum frequency of occurrence for the LLC was normalized to the remaining corpora to retain consistency. This process resulted in a minimum frequency of occurrence of 16 in the MLC and nine in the HLC.

The minimum range criterion was also influenced by the size and specificity of the corpora used in the present study. Although previous research has often implemented a raw range of five texts, a percentage-based approach was considered more suitable in the present study due to the large number of relatively short essays. Consequently, to ensure that the lists of lexical bundles identified in each corpus would not be negatively impacted by idiosyncratic tendencies representative of a few individual writers, and that identified sequences would be more representative of general trends within each collection of writing, a minimum range requirement of 7% was used. This criterion was chosen to help produce a sufficient, yet manageable, list of sequences that could be analysed in greater detail.

Once frequency and range requirements had been decided upon, WordSmith Tools 6.0 (Scott, 2011) was used to extract all three- to five-word recurrent word sequences that met the given criteria. Following this step, the most topic-dependent bundles were removed using a list of 12 content words that were considered potential indicators of test topic. Since the version of the CAEL test being analysed is still in use, this step helped to ensure the topic of the test would not be revealed.

Once overly topic-dependent lexical bundles had been removed, each list was examined for partially overlapping structures (e.g., *a marked increase*, *a marked increase in*, *a marked increase in the*) with the goal of discovering if any shorter lexical bundles contained within longer structures could be eliminated, thereby reducing overlap and leading to more accurate lexical bundle identification. To achieve this goal, in each case where partially overlapping structures were identified, frequency and range criteria were checked. If frequency and range for overlapping sequences varied by a maximum of plus or minus three, the shorter overlapping sequences were eliminated, and only the longer sequence was kept. For example, in the LLC there were 13 occurrences of each of the following sequences: *a marked increase*, *a marked increase in*, and *a marked increase in the*. Since

the frequency of occurrence and range for each overlapping bundle varied by a maximum of plus or minus three, these bundles were joined to create only one listing (*a marked increase in the*).

A final preliminary step in preparing for analysis was to normalize frequency of occurrence in each corpus so that standardized comparisons across each group of writers could be made. In order to achieve this goal, frequency of occurrence for each lexical bundle was normalized to the size of the smallest corpus used in this study, the LLC. In this way, the different sizes of each corpus could be controlled for.

## Analysis & Results

After revising the lists from each corpus to remove overly context-dependent structures and partially overlapping sequences, as well as normalizing frequency of occurrence so that more accurate comparisons between each corpus could be made, it was possible to begin analysis. In order to better understand how each group of writers made use of lexical bundles in their essays, three forms of analysis were implemented: i) number and length of lexical bundles used in each corpus, ii) degree of overlap between lexical bundles and source texts, iii) frequency comparisons for individual lexical bundles within each corpus with a focus on identifying proficiency level differences.

### Number and Length of Lexical bundles

The first form of analysis used in this study focused on the number and length of lexical bundles identified in each corpus. In terms of raw numbers, a total of 64, 51, and 73 lexical bundles were identified in the LLC, MLC, and HLC, respectively. The high number of lexical bundles in the HLC supports the notion that higher level writers tend to make greater use of this feature in their academic writing. However, the fact that MLC writers used the fewest total number of lexical bundles of any group of writers was unexpected. Although numerous potential reasons for this finding are possible, the analysis of lexical bundles and source texts, the second form of analysis, helps to clarify and explain these differences. Consequently, this will be discussed in greater detail within that section.

Since this study focused on three- to five-word lexical bundles, it was possible to compare how each group of writers made use of structures of varying lengths in their essays. In comparing the length of lexical bundles used by each group of writers it was found that MLC and HLC writers tended to make greater use of shorter lexical bundles in their writing, with 87% and 88% of all lexical bundles in the MLC and HLC being of the three-word variety. Conversely, LLC writers displayed a tendency toward the use of longer structures, with four- to five-word lexical bundles accounting for 40% of the identified lexical bundles. Frequency statistics for all lexical bundles extracted from each corpus are provided in Table 2.

Table 2: Lexical bundles by length

| | LLC | MLC | HLC |
|---|---|---|---|
| three-word | 38 (59%) | 45 (87%) | 64 (88%) |
| four-word | 10 (17%) | 3 (6%) | 4 (6%) |
| five-word | 15 (23%) | 4 (8%) | 5 (7%) |
| Total | 64 | 51 | 73 |

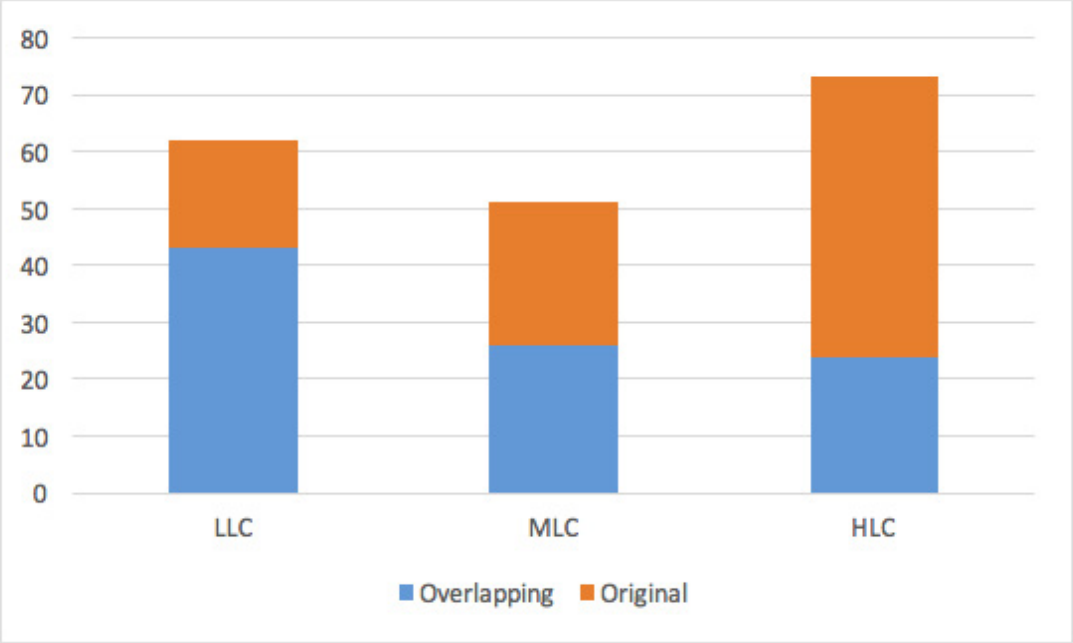
### Lexical Bundles and the Source Texts

Given that the CAEL assessment uses two reading articles related to the topic of the test to help writers become familiar with the subject matter, and that these articles can be used as reference material during the writing task, it was possible to analyse overlap between identified lexical bundles in each corpus and the two source texts. Using this type of analysis, we can achieve a better understanding of the level of reliance on source texts by each group of writers, and which sequences may, or may not, be part each group of writers' internalized linguistic competence. To distinguish between the two types of bundles identified, the term OVERLAPPING is applied to all bundles also appearing in the source texts, and ORIGINAL is used to refer to all other identified structures.

Although a significant portion of the lexical bundles identified in each corpus contained overlap with the reading articles, this tendency was most pronounced in the LLC. Of the 64 lexical bundles identified in the LLC, 43 (67%) were also found in at least one of the two reading articles. In comparison, only 26 (50%) of the lexical bundles identified in the MLC, and 24 (33%) of the lexical bundles identified in the HLC were also found in the source texts (see Figure 1). This finding suggests a greater reliance on source texts in LLC writing and that many of the lexical bundles identified in the LLC may not have been part of the internal lexical repertoire on these writers. While plagiarism is one explanation for this finding, it should be noted that alternative explanations are also possible, including a greater reliance on direct quotations, and a lack of knowledge regarding appropriate academic writing conventions.

By examining overlap between source texts and lexical bundles in each corpus, we can better explain the finding from the first form of analysis that suggested MLC writers produce the fewest total number of lexical bundles of any group of writers. If we remove the overlapping lexical bundles from these lists and focus instead on originally produced structures, we can see that MLC writers actually produce more original lexical bundles than their LLC counterparts. These modified lists show that the use of original lexical bundles increases with proficiency as 21 original bundles can be identified in the LLC, 25 in the MLC, and 49 in the HLC.

Figure 1: Overlapping and Original Lexical Bundles




**Individual Frequency Comparisons between Corpora**

In light of the significant overlap present between the lexical bundles identified in each corpus and the two reading articles included as part of the test, the third form of analysis focused on frequency comparisons for original lexical bundles in each of the three corpora. These frequency comparisons revealed a total of 13 lexical bundles that could be identified as more frequently occurring in the LLC, and 33 lexical bundles that could be identified as more frequently occurring in the HLC. Table 3 provides a list of all originally produced lexical bundles more frequently used by LLC writers in this study.

While numerous conclusions can be drawn from the list provided in Table 3, there are several important items that can be highlighted. First, the frequent use of *a lot of* by LLC writers is noteworthy for the fact that this bundle is more commonly associated with casual spoken English, rather than formal written discourse represented by academic writing. Therefore, the frequent use of this bundle may be an indication that LLC writers have not yet begun to recognize the particular form of English that is needed when writing academic discourse. Secondly, the more frequent use of *according to the* and *according to the article* by LLC writers once again suggests a greater dependence on outside sources in writing at this level. Combined with the previous finding that LLC writers tend to make greater use of lexical bundles appearing in the reading articles, it is clear that LLC writers have a greater dependence on these sources than the other two groups of writers.

Table 3: Originally produce lexical bundles more frequently used by LLC writers.

| | LLC | MLC | HLC |
|---|---|---|---|
| *A lot of* | 49 | 40 | 13 |
| *According to the* | 52 | 17 | 15 |
| *According to the article* | 15 | - | - |
| *First of all* | 20 | 23 | - |
| *In the future* | 23 | - | - |
| *In the world* | 24 | - | - |
| *Is very important* | 16 | - | - |
| *Life on the* | 22 | 16 | 13 |
| *The most important* | 22 | 16 | - |
| *The other hand* | 16 | 12 | - |
| *There are many* | 30 | 21 | 12 |
| *There are some* | 13 | - | - |
| *There will be* | 13 | - | - |

In terms of lexical bundles appearing more frequently in higher level writing, a substantially larger number of lexical bundles were identified. This is at least partially a result of the fact that a greater number of originally produced lexical bundles were identified in higher level writing in this study. Table 4 provides a list of all originally produced lexical bundles that occur with greater frequency in the HLC.

Table 4: Originally produce lexical bundles more frequently used by HLC writers.

| | LLC | MLC | HLC |
|---|---|---|---|
| *A result of* | - | 7 | 14 |
| *A threat to* | - | - | 9 |
| *As a result* | - | 16 | 23 |
| *As well as* | - | - | 19 |
| *Because of the* | 18 | 20 | 28 |
| *Caused by the* | - | - | 11 |
| *Does pose a serious threat* | - | - | 10 |
| *Due to the* | - | 19 | 24 |
| *Fact that the* | - | - | 9 |

| | | | |
|---|---|---|---|
| *In conclusion the* | - | - | 10 |
| *In order to* | - | - | 15 |
| *In the eyes* | - | 10 | 14 |
| *It is a* | - | 14 | 18 |
| *It is not* | - | - | 12 |
| *More and more* | - | 23 | 24 |
| *Of life on* | - | - | 11 |
| *Of the food* | - | - | 9 |
| *Poses a serious threat to* | - | 16 | 25 |
| *That it is* | - | - | 11 |
| *The amount of* | - | - | 15 |
| *The destruction of* | - | - | 17 |
| *The fact that* | - | - | 15 |
| *The increase of* | - | 11 | 12 |
| *The life on* | 27 | 31 | 40 |
| *The loss of* | - | - | 10 |
| *The surface of* | - | - | 12 |
| *The thinning of* | - | - | 11 |
| *This essay will* | - | - | 12 |
| *This is a* | - | - | 11 |
| *To the environment* | - | - | 11 |
| *To the life on* | - | - | 16 |
| *We need to* | - | - | 15 |
| *Which is the* | - | 12 | 9 |

While numerous individual lexical bundles can be highlighted, there appears to be at least one clear pattern in the type of lexical bundles being used by higher level writers in this study, with 'the (_____) of" emerging as a common trend (e.g., *the amount of*, *the destruction of*, *the loss of*, *the surface of*). A total of seven (21%) of the bundles identified as more frequently occurring in the HLC exhibit this pattern. Additional examples that can be highlighted as more frequent in HLC writing include the discourse organizing bundles *in conclusion the*, *this essay will*, *as a result*, and *the fact that*. Each of these bundles can be seen as helping guide the reader by providing greater cohesion and coherence to the essay.

As such, these lexical bundles may lead to improved perceived proficiency by creating a more apparent organizational pattern.

## Discussion

This study used a corpus-driven approach to identify differences in the way recurrent word combinations are used by low-, medium-, and high-level L2 English academic writers. Based on the presented results, it appears that although LLC writers made frequent use of repeated word sequences, many of these structures were likely copied from the source texts included as part of the CAEL assessment. Combined with the fact that LLC writers also tended to use longer lexical bundles than MLC and HLC writers, the presented findings suggest that LLC writers are more reliant on information and structures found in these sources. Consequently, the use of these sources may be a strategy implemented by less proficient writers to help cope with their limited store of formulaic language. In other words, by incorporating passages and sentence fragments from the source texts, LLC writers may have been attempting to supplement their limited linguistic resources with structures they were confident would be considered "proper" academic English.

Based on these findings, caution should be taken when associating increased use of recurrent word sequences in L2 English academic writing with greater proficiency, and more in-depth analysis should be conducted before any firm conclusions can be made. Consequently, the distinction between skilled use of recurrent word sequences and plagiarism becomes an important issue that should be addressed in future research on this topic. While formulaic sequences may hold important benefits for L2 English academic writers, these learners need to be made aware of the potential pitfalls associated with an overreliance on extended sequences from outside sources, particularly without proper citation, since this may be viewed as plagiarism and result in negative consequences for the writer.

Despite the potential danger associated with the use of extended words sequences from source texts, it is clear that, when properly evaluated, this aspect of language can be associated with proficiency level differences in L2 English academic writing. This is evidence by the varying number of lexical bundles identified in each corpus, the differing lengths of lexical bundles used by each group of writers, and the numerous lexical bundles that were found to be more frequently associated with LLC and HLC writing, respectively. As a result, making students aware of common tendencies toward the preference or avoidance of these structures, and their in-context use, may prove beneficial in academically oriented ESL classrooms.

## Implications and Conclusion

Formulaic expressions have been identified by numerous researchers as an important part of native language users' linguistic competence (e.g., Wray, 2002). With gathering evidence of the importance of formulaic expressions to L1 users, it seems this element of language may also prove beneficial to L2 learners aiming to increase proficiency and general linguistic ability in their target language. Therefore, for teachers and students

alike, understanding how formulaic expressions are used in the target language genre and register is an important step. For L2 English academic writing, the lists presented here can be used to achieve this goal by helping teachers and students better understand usage tendencies for specific formulaic expressions associated with higher perceived linguistic ability, and eventually helping these learners feel confident and capable enough of producing appropriate formulaic expressions on their own. Conversely, these lists can also be used to help students better understand which structures they may want to avoid in order to appear more linguistically competent in their academic English writing.

Although this study has focused specifically on L2 English academic writing, lists of common formulaic expressions in other genres and registers can also be explored (e.g., Biber & Barbieri, 2007; Chen, 2008; Cortes, 2008; Hyland, 2008; Wood & Appel, 2013). Wherever possible it is important to closely align the formulaic expressions being taught to the target genre and register of the learner. To investigate specific usage patterns for existing lists of formulaic expressions, several free on-line corpora can easily be searched and used as in-class tools (e.g., British National Corpus, Corpus of Contemporary American English, etc.). Teachers and students can use these corpora to explore the frequency with which specific formulaic expressions occur and better understand how these structures contribute to the effectiveness of each piece of discourse. For example, formulaic expressions from the lists presented here could be searched for in online corpora and used as examples for in class discussion.

Despite the important benefits associated with the correct use of formulaic expressions, it is important to recognize the potential pitfalls that may result from teaching these sequences to students since their use can also be viewed as academic plagiarism (see above). In addition to reviewing definitions of academic plagiarism with students, writing samples could also be used in class as a way of highlighting how effective use of formulaic expressions differs from academic plagiarism. By further developing and using corpora of L2 English writing in ESL and EAP classrooms, students may be able to better identify deficiencies in their own writing and develop ways of remedying these issues.

By continuing to investigate the use of recurrent word sequences of various lengths, as opposed to the more standard approach that focuses solely on four-word structures (Chen 2008; Cortes, 2004, 2008; Hyland, 2008), it will be possible to develop a more complete picture of how formulaic expressions are used by L2 English users of varying proficiency levels and better understand the difference between skilled use of formulaic language and academic plagiarism. This is an important area that deserves increased attention, and the growing availability of computerized corpora should provide useful means of exploring these issues.

## References


Adel, A. & Erman, B. (2012). Recurrent words combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*(2), 81–92.

Altenberg, B. & Eeg-Olofsson, M. (1990). Phraseology in spoken English: Presentation of a project. In J. Aarts & W. Meijis (Eds.), *Theory and practice in corpus linguistics* (pp. 1–26). Rodopid: Amsterdam.

Association of Universities and Colleges of Canada (2010, August). *Canada's universities: Contributing to a better future*. Pre-budget submission to the House of Commons Standing Committee on Finance. Ottawa: AUCC.

Bamber, B. (1983). What makes a text coherent? *College Composition and Communication, 34*(4), 417–429.

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

Biber, B., Conrad, S., & Reppen, R. (1999). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research*, *10*(3), 245–261.

Bybee, J. & Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics*, *37*(4), 575–596.

Canadian Academic English Language Assessment (CAEL). (2010). *Scoring criteria, methods and reliability*. Retrieved from http://www.cael.ca/pdf/C4.pdf

Chen, L. (2008). An investigation of lexical bundles in electrical engineering introductory textbooks and ESP textbooks (Master's thesis). Carleton University, Ottawa, Ontario.

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

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

Cortes V. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora, 3*(1), 43–57.

Firth, J. (1935). The technique of semantics. *Transactions of the Philological Society, 34*, 36–77.

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

Granger, S. & Rayson, P. (1998). Automatic profiling of learner texts. In. S. Granger (Ed.), *Learner English on computer* (pp. 119–131). New York, NY: Longman.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4–21.

Kuiper, K. (1996). *Smooth talkers*. Mahwah, NJ: Erlbaum.

McCully, G. (1985). Writing quality, coherence, and cohesion. *Research in the Teaching of English, 19*(3), 269–282.

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

Pawley, A. & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–226). New York, NY: Longman.

Scott, M. (2011). WordSmith Tools Version 6.0. Oxford, UK: Oxford University Press.

Underwood, G., Schmitt, N., Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, processing and use* (pp.153–172). Philadelphia, PA: John Benjamins.

Virtanen, T. (1998). Direct questions in argumentative student writing. In. S. Granger (Ed.), *Learner English on computer* (pp. 94–106). New York, NY: Longman.

Wood, D. & Appel, R. (2013). Lexical bundles in 1st year university business and engineering textbooks: A resource for EAP. In H.M. McGarrell & D. Wood, (Eds.). *Special Research Symposium Issues of CONTACT*. Refereed Proceedings of TESL Ontario Research Symposium, October 2012. Vol. 39, No. 2 (pp. 92–102).

Wray, A. (2002). *Formulaic language and the lexicon*. New York, NY: Cambridge University Press.

**Appendix A: Writing Performance Band Score Criteria**

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