Investigating the Lexical Bundle Use of Novice and Upper-Level Writers Within a University Discipline

By

Max Weinstein

A Thesis

Submitted to the Faculty of Graduate Studies
In Partial Fulfillment of the Requirements
For the Degree of

Master of Arts

School of Linguistics and Language Studies

Carleton University
Ottawa, Ontario

May 24, 2011

Your file Votre référence
ISBN: 978-0-494-83101-4
Our file Notre référence
ISBN: 978-0-494-83101-4

Abstract

This was a quantitative dominant, mixed method study, examining how novice and upper-level student writers (native and non-native speakers) in an applied linguistics and discourse studies (ALDS) program used lexical bundles. A novice and upper-level corpus of ALDS student writing were analyzed for overall bundle counts as well as discourse functions. The results showed that novice writers use a larger number of lexical bundles than their upper-level counterparts, and that lower-level novice writers use fewer of the bundles found in the upper-level writing. In the second phase of the study, questionnaires and interviews were conducted with four ALDS students to inquire about the meaningfulness of lexical bundles as a component of language support. One implication of the study is that lexical bundle analysis can be used to highlight differences in the writing of novice and upper-level students within a single discipline. Results also suggest that the use of lexical bundles, especially text organizing bundles, is an important component of successful ALDS writing.

Acknowledgements

It is my great pleasure to thank all of the friends, peers, students, faculty, and family that made this thesis possible. I first need to thank my supervisor, Janna Fox, for the time and care she showed me not only through the writing of my thesis, but throughout my two years in the ALDS Masters program. Witnessing the dedication and passion that Dr. Fox brings to her work has been inspiring, and made my learning experience immeasurably richer.

I also must thank David Wood. His willingness to share his knowledge about formulaic language and offer critical and creative advice has made this piece of research possible. Dr. Wood's interest in my work and consistent support were key contributors to any success I achieved in this program. I have greatly appreciated his clear perspective and sense of humour throughout this process.

I would also like to thank Jennifer Gilbert, who has offered me great opportunities for learning at the Centre for Initiatives in Education, and shown great kindness to me throughout my final year at Carleton. Her feedback on my thesis, and genuine interest in my work has been a great help.

Finally, thank you to my external examiner Tom Cobb for his insightful feedback and great encouragement with my research and future studies.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

ALDS – Applied linguistics and discourse studies
ALDSI – Applied linguistics and discourse studies intermediate corpus
ALDSN – Applied linguistics and discourse studies novice corpus
ALDS-UL – Applied linguistics and discourse studies upper-level corpus
BNC – British National Corpus
CANCODE – Cambridge and Nottingham corpus of discourse in English
CPALS – Carleton Papers in Applied Linguistics
DELNA – Diagnostic English language needs assessment
ESL – English as a second language
EAP – English for academic purposes
ESP – English for specific purposes
L1 – First language
L2 – Second language
LGSWE – Longman Grammar of Spoken and Written English
MA – Master of Arts
TA – Teaching assistant
T2K-SWAL - The TOEFL 2000 Spoken and Written Academic Language Corpus
WTS – Writing Tutorial Services

Chapter 1 Introduction

The activity of academic writing has been examined and re-examined from various perspectives using a range of research methodologies within applied linguistics, genre and writing studies (Hyland 2009, Artemeva, in press). This thesis is a quantitative dominant, mixed methods investigation (Tashakorri and Teddlie, 2010) of academic writing in two phases: The first phase involves corpus research, with a focus on formulaic language, and more specifically lexical bundle usage in Carleton's Applied Linguistics and Discourse Studies program (ALDS). The second phase is a questionnaire and interview of student writers in the same discipline. This first phase of research (the primary part of the study) involved analyzing corpora of novice and upper-level writing from ALDS at Carleton University to find out which frequently occurring, widely distributed, pragmatically specialized sequences of words (lexical bundles) were used by both group of writers, and how they might vary in discourse functions. In the second phase of the study, a questionnaire and interview with student writers was analyzed in order to deepen the understanding of the corpus results by triangulating lexical bundle trends and patterns with knowledge and experience from current ALDS students. The questionnaire was administered to four ALDS student writers (both novice and upper-level native and non-native speakers) and included several questions about academic writing experience, as well as demographic information about language background and year of study. The questions regarding academic writing experience were informed by past research (McCarthy, 1987; Freedman and Medway, 1994; Haas, 1994; Russell, 1997; Sommers and Saltz, 2004; Beaufort, 2004; Artemeva and Fox, 2010;

Rogers, 2010) as well as results from the corpus analysis in the current research. A short interview immediately following the questionnaire allowed participants to elaborate on the answers they provided. A mixed method approach was chosen for this research because lexical bundle analysis alone was not deemed sufficient to address questions about why novice or upper-level writers were using specific bundles in a specific way, or how lexical bundle usage might be indicative of one's success in a discourse community. As Tashakorri and Teddlie (2010) say in their guide to mixed methods research, one should use "the best techniques available to answer research questions that frequently evolve as a study unfolds" (p. 8). Although very few studies in corpus linguistics have combined lexical bundle analysis with other types of qualitative or quantitative measures, Hyland (2009) championed a mixed method approach in his corpus informed study on academic engagement:

> Corpus analysis is a method which moves away from individual preferences to focus on community practices, dematerializing texts and approaching them as a package of specific linguistic features employed by a group of users...Corpus studies also require a focus on 'action' to balance the focus on 'language', however, and this necessitates rematerializing these features to understand how and why writers make the choices they do when they write (p. 110)

Underpinning my investigation of lexical bundles and writing development are theories about formulaic language, which will be discussed next (as background to the research questions) along with information about the three corpora built for the study.

1.1 Introduction to formulaic language

Formulaic language is commonly defined as a sequence of words that is stored in the brain as a prefabricated chunk and is produced without the generative functions of grammar and syntactic rules. Lexical bundles are a subset of formulaic language, commonly defined as "recurring combinations of three or more words which are identified in a corpus of natural language" (Wood, 2010, p. 92). These word sequences are found by setting frequency and distribution parameters in a corpus search program, and sequences identified in the corpus must have an identifiable function(s) in the discourse to be classified as lexical bundles. While there have been more and more studies emerging in the field of corpus linguistics that examine lexical bundles, few studies compared levels of student writing (both L1 and L2) within a specific discipline. Studies by Tribble and Scott (2006) as well as Hyland (2008b) were similar, but these studies have compared beginner writing with non-student expert writing. It is valuable for students to see not only how sequences are used by academics in a field, but also how upper-level students in the same discipline construct their texts.

A study comparing novice undergraduate writing from a discipline with upper-level writing in the same discipline has not been examined with a focus on lexical bundles. This study addresses the gap in the research by compiling a primary corpus of upper-level student writing (ALDS-UL), and a comparison corpus from novice student writing (ALDSN) in the ALDS discipline. The upper-level corpus was scanned with Wordsmith Tools 4.0 (Scott, 2004) and Antconc 3.2.0m in order to find and analyze occurrences of lexical bundles. The function of these bundles was organized into the taxonomy informed by the work of Biber (1999), Biber, Conrad

and Cortes (2004), Cortes (2004), Hyland (2008b) and Chen and Baker (2010), and compared to the bundles found in the novice corpus. In addition to the two main corpora, a third corpus of 'intermediate' ALDS writing (ALDSI) was used as a control to investigate some of the findings from the ALDN and ALDS-UL comparison. The ALDSI corpus was not analyzed as extensively as the primary corpora were.

There have been several studies that focus on lexical bundles in university registers that were useful in predicting results from the ALDS-UL and ALDSN corpora. Firstly, it was expected that the total number of bundles as compared to registers of conversation or classroom teaching would be quite low. This was based on Biber's (2006) findings from the T2K-SWAL corpus, which showed that registers of academic writing that allow for greater amounts of planning and editing use fewer bundles than spoken registers. It was also expected that the papers from novice and upper-level writing would contain different numbers of bundles, and that they would differ in their functions in the text. This is to say that both novice and upper-level writers rely to some extent on lexical bundles, but the mastery of their use and selecting the appropriate variety makes a noticeable difference to the reader. Based on past research, it seemed likely that the bundles present in both corpora would be made up mostly of referential bundles, with a relatively small number of stance and discourse organizers. To measure the importance of each functional category, bundles from each category were examined in isolation from other categories[1] and compared between the two corpora to see whether there was any notable difference.

---

[1] For example, each type of stance bundles was measured for frequency and context, and compared within and between novice and expert papers.

1.2 Rationale for the research

Overall, it was expected that there would be a sufficient number of bundles found in the writing, and sufficient difference in pragmatic functions to conclude that lexical bundles are an important aspect of student academic prose, as well as an important aspect of university writing development. Therefore, this research could have numerous benefits for students and teachers alike.

Firstly, getting a picture of which lexical bundles upper-level students are using within ALDS would give EAP/ESP teachers a chance to provide L2 learners with more authentic language instruction and awareness-raising activities. Research from specific academic disciplines (in this case Carleton's ALDS program) is also crucial since there is considerable variety between the bundles used in different disciplines (Biber, 2006; Hyland, 2008; Oakey, 2002). This type of discipline specific language data is especially important for students who have difficulty with language or writing because,

"leaving [their] learning to chance encounters with lexical bundles in texts is not a reliable way to build knowledge" (Byrd and Coxhead, 2010, p. 56).

Exercises that present chunks in listening and noticing activities, fill-in-the-blank dialogues, and even isolated drills on lexical bundles could be effective ways of assisting L2 learners (O'Keefe, McCarthy and Carter, 2007, p. 77); however, these materials/activities will be much more effective if students can work with bundles taken from a register(s) they are striving to engage in. As Byrd and Coxhead (2010) emphasize the need for more instruction and analysis of lexical bundle in the classroom (specifically with L2 speakers): "Putting lexical bundles back together post analysis and using them accurately and appropriately in speaking and writing

are not easy tasks for learners. This is particularly true when they are participating in short courses of instruction" (p. 55). Although L2 students are often those most in need of such language instruction and could benefit from working with frequent sequences taken from successful students, native English speakers could also benefit greatly from this type of research. Students at a variety of levels could receive feedback about language use that is informed by corpus research and students who master these important bundles early may have an easier time reading and writing academically since "the use [and knowledge] of chunks 'frees up' the cognitive processing load so that mental effort can be allocated to other aspects of production such as discourse organization" (O'Keefe et al., 2007, p. 77). Although this study was not conducted at a scale that affords a great deal of generalization from the findings, results could be used to inform future studies in this area and raise teacher and student awareness on the role and importance of lexical bundles in academic writing.

1.3 Research questions

Lexical bundles form the basis of the first set of research questions below for the following reasons: (1) there have been no (known) studies that look specifically at lexical bundles in a spectrum of university student writing in a single discipline (first year to graduate level) (2) lexical bundles have been shown as an important component of genre/register knowledge since many sequences are community and register specific (Biber, 2006; Hyland, 2008, 2008b; O'Keefe et al. 2007; Wray, 2002; Cortes, 2004; Tribble and Scott, 2006; Byrd and Coxhead, 2010; Oakey 2002) (3) Past research has shown that native-like control and knowledge of formulaic

sequences/lexical bundles is important for academic writers to express ideas and content in a register appropriate fashion (Biber et al. 2004; Cortes, 2004; Hyland, 2008; Oakey, 2002). 4) Identifying lexical bundles may also provide students with linguistic knowledge that is not otherwise transparent or intuitively learned (Cortes, 2004; Cortes, 2006; Byrd and Coxhead, 2010).

In the first phase of the study, there is one key question: Are there any noticeable patterns in the way novice or upper-level ALDS student writers use lexical bundles and can these patterns be used to characterize novice or upper-level ALDS writing in any way? This main question is broken down into the following parts: 1) Are there any noticeable differences in the number of bundles or distribution of bundles across texts between novice and upper-level ALDS writers? 2), Are there key bundles in ALDS that occur to a significant extent in either high or low level apprentice ALDS writing? 3) Based on the comparison of upper-level and novice ALDS writing, what importance might lexical bundles have for helping students improve their writing?

The next set of questions relating to the qualitative phase of the study are posed in order to better assess the importance of lexical bundles in the larger context of writing development and transition between academic levels. 1) What aspects of writing (content, format, and language expression) are reported as being the most challenging by novice and upper-level writers respectively? 2) How do novice and upper-level writers conceive of the ALDS genre in terms of importance and purpose of writing, importance of language command, and importance of content? 3) Do novice or upper-level ALDS writers express any desire for increased

instruction or practice with ALDS writing tasks, and if so what form should writing support take?

In the following three chapters I will discuss the evolution of corpus research and some relevant corpus studies, review the literature on formulaic language and how it is defined and understood in spoken and written contexts, and present a more specific review of lexical bundles and the corresponding research. After reviewing the literature relevant to the first phase of this study, some research concerning academic writing development, genre, and case studies on university writers will be considered in chapter 5. Literature review chapters will be followed by methodology (chapter 6), results and discussion (chapter 7), and finally pedagogical implications, limitations, and conclusions (chapter 8).

Chapter 2 Corpus research and construction of corpora

This chapter will discuss the evolution and history of corpus research, before reviewing common corpus construction methods and findings from previous corpus studies. These studies relate to research questions and methodology for this study concerned with how a specialized corpus should be built as well what patterns and data one could expect to find doing lexical bundle analysis of academic student writing.

2.1 Evolution of the corpus: Generalized vs. specialized corpora

Before computers became as powerful and commonplace as they are today, any study of language that intended to look at authentic samples of speech or writing had to compile hard copies of all materials and examine them by hand (Sinclair, 1991; Wray, 2002). This allowed researchers to do some qualitative analysis of a corpus, but there was still no practical way to analyze enough text to make quantitative generalizations about the language as a whole. This great limitation to exploring language questions caused linguistics research to become far more introspective, relying on intuition and deductive reasoning as opposed to examination of data gathered from living language production. As computer technology advanced in the mid to late 20th century, new opportunities availed themselves in the field. With greatly increased access to texts and easy electronic storage opportunities, researchers could search hundreds or thousands of texts containing hundreds of millions of words (Tribble and Scott, 2006, p. 4) to examine new theories and commonly held beliefs about the structure and use of language.

This proliferation of technology has allowed linguists to construct very large generalized 'reference corpora' such as the BNC and CANCODE[2], that contain hundreds of millions of words from sources including conversation, essays, movies, books, etc. These reference corpora are intended to reflect the state of a language as a whole, and thus aim for as much language from as many mediums and genres as possible. Another great advance in the technology of computers has been the increasing transparency and intuitive design of programs and operating systems. While one had to be familiar with programming languages and complex procedures to search or compile corpora in the early years, corpus research has now become feasible for researchers, teachers and students who have even the most minimal computer literacy. This has not only allowed for wide ranging pedagogical applications of corpora, but has also allowed many people who are interested in language to build their own specialized databases of language.

These much smaller, "specialized corpora" (such as the one intended for this study) are restricted to certain registers or speech communities, and do not aim to be representative of the language as a whole. These corpora are often constructed by insiders of the register who are experienced with how texts are produced and evaluated in their field. Although general reference corpora are much larger, they are often not better tools for exploring language in specific registers. Because specialized corpus designers can select texts with special care and expertise brought from their field of study, these corpora tend to have more representative coverage

---

[2] The acronyms BNC and CANCODE stand for British National Corpus and Cambridge and Nottingham corpus of discourse in English.

their target register than sub-groups of reference corpora, which have been designed with much different purposes in mind.

In addition to general reference and specialized corpora, there are several other labels used to describe corpora of different types. A Learner corpus is used to identify any corpus that has been compiled entirely of learners' work. The three corpora used in this study can be thought of as Learner corpora as well as specialized corpora (as they are referred to in chapter 6) because they are made up of ALDS student (learner) writing. Corpora can also include texts in multiple languages; commonly referred to as 'bilingual' or 'multilingual' parallel corpora (Gentil, 2009, Personal communication)

With the exponential growth in the past decades of corpus construction (especially with smaller corpora), it has becoming increasingly possible to build a corpus without much background knowledge of corpus linguistics methods or theory. It is crucial that some guidelines and theoretical background about how to build a proper corpus be followed if that corpus is going to have any value. Some of these guidelines concerning general reference and specialized corpora are considered next.

2.2 Methods of corpus construction

To begin, it is important to differentiate between methods for constructing a general reference corpus from a smaller specialized corpus. While there are many elements of design applicable to both, there are significant differences in addressing representation and variety of data. Each of these corpus types will be discussed in turn.

In undertaking the construction of a reference corpus, there are two orders of concern: the first order considers issues arising before texts are selected, and the second with how to properly select and manipulate texts for inclusion in the corpus. Before work on a corpus begins, a researcher should consider the following questions: (1) What purpose is the corpus going to serve? (2) What time period will the texts cover? (3) What modes of text will be considered (Sinclair, 1991)? It is especially important that a goal for the corpus, or some intended use is made clear from the outset; if not designed with purpose, choices concerning, mode, typicality, representation, and balance of texts will be impossible to address. It is also worth considering whether one's status as an expert or insider in a particular area of language might make one a better or worse choice for corpus builder. Sinclair (1991) suggests that linguists and corpus analysts can make poor corpus builders because they are not able to create a true random sample of text for inclusion. However, linguists and corpus analysts sometimes are the more logical choice for specialized corpora since intimate knowledge of language and genre is important in their design.

Once these initial concerns have been addressed, one must carefully consider methods of text sampling and collection. Typicality: how typical is each text of the language or register in question? Representativeness and balance: texts selected for a reference corpus must represent the target language as broadly and evenly as possible. This means that modes (spoken, written) as well as genre must be broadly sampled. Put another way, "corpora should not be concerned with the specific language of the text, but the communicative function the text plays within the language community" (Sinclair, 2005, p. 5). While these issues of balance and

representativeness of texts must be considered for specialized corpora, they are undertaken differently because the purpose of the database is different. For example, the upper-level ALDS corpus built for this research did not need to represent the English language as a whole, or serve as a reference point for graduate student language in general; it was only necessary that the ALDS-UL represent a balanced sample of upper-level writing in a single ALDS program (papers from a variety of courses and writing assignments) that was large enough to analyze and compare with other levels of student writing.

Studies with specialized corpora have increased in recent years, generating a great deal of data on language in many different registers. For example, the MLC and MICASE[3] corpora are specified to academic learner language, and are used only to infer language patterns for those specific groups and contexts represented by the source texts. Specialized corpora also have very different targets in terms of size. While a corpus should generally be "as big as possible" (Sinclair, 1991, p. 4) to increase the usefulness of the data, specialized corpora can operate at much smaller sizes (often in the high thousands or hundreds of thousands of words) and still generate valuable results. Sinclair's (2005) work has shown that while specialized corpora do not match larger corpora for variety of lexical items with frequencies under 20 times per million, once above that threshold they become quite comparable. Because "this characteristic vocabulary of the special area is prominently featured in the frequency lists [...] much smaller corpus will be needed for typical studies" (Sinclair, 2005, p. 13).

---

[3] MLC is the Multilingual Learner Corpus, MICASE is the Michigan Corpus of Academic Spoken English.

2.3 Interpreting Corpus Results

While corpora are invaluable tools for discovering how language is actually used, the results they generate must be interpreted with a proper amount of expertise and skepticism. Firstly, a researcher must carefully examine entries on word lists that have been falsely inflated due to irregularities or imbalances in the source text. Some examples of these imbalances include oft-repeated interjections, "fifteen minutes of fame expressions"[4], and other tokens not representative of the target discourse. Researchers using specialized corpora will often remove/separate words or sequences that involve proper nouns or topic specific jargon (Oakey, 2002). This issue of token/cluster selection from my corpus data will be addressed in methodology and results.

It is also important to note that when doing research with a general reference corpus, the absence of an idiom or other formulaic sequence does not mean that this idiom is not used in natural language. In Moon's (1998) work on identifying discourse functions of phrasal lexemes, she uses an 18 million-word corpus and still finds very low frequencies or no occurrences of many opaque metaphors or true idioms[5]. Moon states that her "eighteen million words [...] are not enough to describe adequately and in detail many of the phrasal lexemes I was looking at" (p. 81). A corpus such as Moon's (1998) does not inform the researcher how often the pragmatic situation for each phrasal combination occurred. Gaining this type of more textured information from a corpus, such as how often a lexical

---

[4] These expressions are ones that arise from media or popular culture and experience very high rates of usage over a relatively short period of time (Wray, 2002, p. 27).

[5] True idioms are sequences that have almost no variability or semantic compositionality; i.e. 'red herring'.

item occurs in relation to a certain pragmatic situation is possible with the use of mark-up language and tagging. One can add tags to parts of speech, pragmatic categories of a text (stance bundles, topic organizers, etc.) or large sections of discourse (introductions, conclusions). Although this type of mark-up requires higher computer literacy and can be an overwhelming task for one or two researchers, it is often a valuable tool for maximizing a corpus' utility.

2.4 Influence of Corpora

Since the first significant corpus study conducted nearly five decades ago (1961), there has been a great deal of advancement both in the field of corpus research, and linguistics both applied and theoretical. The following is a brief overview of corpus research and its influence.

The first large corpus to be studied and reported on was the Brown corpus or the "Standard Corpus of Present-Day American English", and had 1,014,312 running words[6] when it was created in 1961 (Francis & Kucera, 1979). The Brown corpus, typical of many early corpora, included a wide array of text from written mediums (newspaper, novel, letter, business report, etc.), but did not include much if any transcribed spoken data. The Brown corpus, and other large reference corpora were primarily used as dictionary building tools (Sinclair, 2005). Lexicographers could test their intuition about a new word to not only see how often it occurred in the corpus, but what contexts it normally occurred in. While the impact on lexicography

---

[6] When texts are compiled into a corpus, analyzing software will differentiate between running words and individual words in the text. Every string of letters separated by a space on either side is considered a running word (similar to the count given by word processors) and individual words give a count of how many different words are in the corpus. This means that an individual word count would count *the* as one word-type, even if it occurred thousands of times in text.

was profound, there were also broad implications for linguists and grammarians from this new searchable source of 'living language' (Sinclair, 1991). Perhaps the greatest discoveries that came from this research according to Biber (2001) were "the centrality of register for studies of language use" and "the unreliability of intuitions about use" (p. 332). The former point refers to the fact that when texts are organized and searched by register, they yield significantly different results of language use than the global corpus and other registers. The latter point Biber makes about the unreliability of intuition has threatened to radically redefine the way linguists and language teachers approach grammar and the teaching of language. The possibilities afforded by this new technology allowed Biber, Johansson, Leech, Conrad, & Finegan (1999) the opportunity to construct a manual of English grammar based on corpus text from a range of disciplines. The Longman Grammar of Spoken and Written English used samples from real academic writing and speech to provide authentic examples of grammatical forms and functions. Collocations and colligations[7] of words and clusters were also used to provide better guides for learners in how to use these grammatical rules and structures. In Biber's (2001) study he runs basic frequency counts of selected texts from the Longman Grammar of Spoken and Written English database to show that basic assumptions about aspect and verb choices (i.e. progressive aspect as the unmarked choice in conversation) are not supported by corpus data. This use of corpora to reshape textbooks and pedagogy in the field of language teaching has generated a great deal of research and controversy in the past two decades. Because it is relatively easy to

[7] Colligation refers to the statistical preference of words or groups of words to co-occur with grammatical forms or choices. For example, an impersonal third person pronoun might be found to *colligate* with the present perfect aspect in some texts.

build a corpus and make sweeping generalizations from the results, many have correctly pointed out that these corpus results are often not good representations of language in general or specific registers and genres. Even the biggest, and most carefully built corpora such as the BNC or CANCODE are far from perfect representations of English and require frequent maintenance due to the dynamic nature of language and culture[8] (Sinclair, 2005).

Despite challenges from linguists and teachers in favour of more traditional language investigation methods, there has been a push to exploit corpora instead of simply explaining their results (Flowerdew, 1998, p. 542). One of the most influential uses and 'exploitations' of corpus tools has come from the work by Avril Coxhead (2000) in compiling the New Academic Word List. From a corpus of academic texts from a variety of disciplines (28 subject areas), containing over 3 million words, Coxhead (2000) produced a list of 570 key word families. These words accounted "for 10% of the total tokens in the Academic Corpus, and more than 94% of the words in the list occur in 20 or more of the 28 subject areas of the Academic Corpus" (Coxhead, 2000, p. 226). There were more general word lists constructed for ESL learners, as well as a University Wordlist created by Xue and Nation (1984), but the AWL had better coverage while using considerably fewer words. This Academic Word List has been widely influential in EAP/ESP, but is not without some skeptics. Hyland (2008) pointed out with his study of lexical bundles across academic registers that the variety in language, even between academic

[8] It is also important to consider the serious issues of power and representation that exist in the building of corpora. Because published text makes up the vast majority of reference corpora and this text is produced by a select few (often privileged) people, prescriptive corpus approaches risk misrepresenting, devaluing, or ignoring many registers of speech and text.

genres, is large enough to shed doubt on the usefulness of small academic word lists. Hyland (2008) argues that his study emphasizes the usefulness of lexical bundles as the target for EAP corpus research, and "helps undermine the widely held assumption that there is a single core vocabulary needed for academic study" (p. 20). While Hyland makes a strong case against one core academic vocabulary, it is arguably still important to pursue these kinds of informed lists for EAP students. Coxhead and Byrd (2010) point out that Hyland's findings for lexical bundles across academic disciplines in his 2008 study found that they occurred in only 2.2% and 1.1% of each other's academic corpora respectively. Considering that "learners need many encounters with a word or phrase before it becomes part of their lexicon", most EAP students will not get sufficient exposure to lexical bundles since they "will read fewer than the 15,000 words needed to encounter *on the basis of* [an example bundle] even twice" (Coxhead, 2000, p. 47).

2.5 Corpora and Formulaic Language

As discussed in the section above, corpora can be valuable in many areas of applied linguistic research, but they may be especially useful to someone interested in exploring formulaic language and word clusters. A corpus can be used to measure the frequency of word strings and give an excellent starting point to identify what is or is not formulaic in a set of data. Frequency is an important aspect of formulaic sequences; as Wray and Perkins (2000) point out, "there is undoubtedly some relationship between frequency and formulaicity, both in the sense that some formulaic sequences are used frequently, and formulaic output is frequently called upon" (p. 7). Studies focused on cognitive grammar have also supported the

relationship between formulaicity and frequency; "the formulaic nature of the adult language system comes about via a process of schematization, [...] through the reinforcement and progressive entrenchment of recurring commonalities, as well as the `cancellation' (non-reinforcement) of features that do not recur" (Langacker, 1987, p. 107).

Researchers must be conscious that corpus data and frequency counts are not sufficient to infer that a sequence is stored as a chunk in the mental lexicon. In assessing the psycholinguistic validity of their findings, Schmitt, Grandage and Adolphs (2004) warn "that corpus data on its own is a poor indicator of whether those clusters are actually stored in the mind as wholes" (p. 147). The creator of Wordsmith Tools is also careful to differentiate clusters (lexical bundles) from phrases or prefabricated units, emphasizing that despite frequency and distribution of clusters, "simply being found together in software doesn't guarantee they are true multi-word *units*" (Scott, 2011). Although Wordsmith is capable of generating lists of the most frequent clusters within specified parameters, there is still a great deal of intuition that must be used by the researcher in pulling out clusters that have some pragmatic function and pedagogical purpose (Nattinger & DeCarrico, 2002). From his study of lexical bundles in academic disciplines, Oakey (2002), who also used a version of Wordsmith Tools, laid out the following two conditions for including a word-string as a lexical bundle: "Its form must approximate to the specific string, and then the pragmatic function of the phrase must be the one specified" (p. 116). In any study involving lexical bundle identification, "it is still the researcher's intuition which makes the judgement based on the context in which the string occurs" (Oakey, 2002, p. 116).

Having looked at the role and development of corpora in academic writing research, I will take a closer look at formulaic language next.

Chapter 3 Formulaic Language

In the first part of this chapter, definitions and evidence for identifying formulaic sequences are taken from research and theory related to spoken language. Although this research focus is on academic writing, evidence from speech provides valuable background for several reasons: evidence from speech helps present a more complete picture of how research on formulaic language has developed, there are several similarities in how language is identified and classified as formulaic in both speech and writing (i.e. frequency of use, lexicalization within a speech community, specific pragmatic purpose (O'Keefe et al., 2007, Biber et al., 2004)), and some of the rationales for why we use formulaic language can apply to both spoken and written production (i.e. reduced processing effort for both producer and listener/reader, ability to sound native like in a specific genre or register). These connections between spoken and written formulaic language will be made more explicit in the sections that follow.

Formulaic language can be defined rather succinctly as two or more independent morphemes that are stored as a single unit in a person's lexicon, and are not subject to analysis by grammatical rules. In Weinert's (1995) words,

> Generally, they [formulaic language formulas] are expressed in terms of processes, and refer to multi-word *(How do you do?)* or multi-form strings *(rain-ed, can-'t)* which are produced or recalled as a whole chunk, much like an individual lexical item, rather than being generated from individual lexical items/forms with linguistic rules (p. 182)

Taking this rather broad and inclusive definition, how much of our lexicon and linguistic knowledge is made up of this prefabricated material? Bolinger (1976) made a famous proclamation that, "speakers do at least as much remembering as

they do putting together" (cited in Erman, 2001, p. 1353). In their seminal study seven years later, Pawley and Syder (1983) suggest that while the "the number of single morpheme lexical items known to the average mature speaker is quite small, the number of morphologically complex items is much greater, running [...] into the hundreds of thousands" (p. 210). The belief that English (especially spoken) has a large proportion of formulaic language has continued to gain traction with others since Pawley and Syder's (1983) work, including Erman (2001) who says that "40-60% of an appreciable body of spoken and written texts consists of more or less ready-made, idiosyncratic combos of words" (p. 1353). The linguist Jackendoff (1997) also suggests that, "the number of fixed structures in a speaker's mental lexicon is of at least the same magnitude as in single words of the vocabulary" (cited in Erman, 2001, p. 1353).

These claims about the proportion of formulaic language have brought scrutiny and doubt to the models of language production championed by Chomsky (1965) and his followers. Although Chomsky (1965) acknowledged the presence of prefabricated lexical items and memorized language, he largely discarded the problems they posed to his theory of generative grammar.

Chomsky's groundbreaking work in the 1960s set the paradigm for how linguists understand language production and first language acquisition. Chomsky (1965) deduced that we use a finite number of grammatical rules and knowledge, often understood only tacitly, to produce an infinite range of utterances. The principle that guides our language (both spoken and written) is thus one of open choice, allowing us to place words together with generative rules. While few would disagree with the claim that the open choice principle is a crucial part of our

linguistic competence, it is the role of the idiom principle that creates disagreement among linguists. The idiom principle "brings about the selection of two or more words together, on the basis of their previous and regular occurrence together" (Sinclair 1991, p. 110). Although Chomsky acknowledged the problem posed by idioms and other sequences of language that do not seem to be constructed with straight-forward syntax, the data were ignored because they were considered too small and inconsequential to challenge the primacy of an open choice model. However Wray (2002, 2008), Nattinger and Decarrico (1992), Pawley and Syder (1983), Weinert (1995), Van Lankar Sidtis (2009), Wood (2001), and Sinclair (1991) argue for a more balanced dual system model that favors the idiom choice principle and uses syntax as a sort of glue and repair kit when necessary. According to Wood, "It is likely that the basis of fluent speech is an intricate interweaving of formulaic and newly constructed segments" (p. 580). Because writing allows one to produce language more slowly with a more analytical approach to building text, it often contains a lower proportion of formulaic language and perhaps more of these "newly constructed segments" than spoken registers. However, the idiom principle likely plays an important role in writing production due to the high number of lexical bundles, idioms, and other collocations used in many registers of writing (Cortes, 2004; Hyland, 2008; O'Keefe et al. 2007). It is also fair to assume that speed of production and time constraints also play a role in dictating how much formulaic language is used in writing, just as they do in speech (Kuiper, 1996).

The idiom and open choice principles help explain linguistic competence, but the picture becomes more complicated when differences between child and adult learners are considered. In his analysis of the open choice principle, Sinclair (1991)

describes generative grammar and prefabricated language choices as diametrically opposed systems that function in turns (stop/start) instead of simultaneously. Although no specifics are provided concerning adult language and second language learning, the idiom choice principle is described as the dominant model we use to produce language (regardless of native/non-native speaking status). The dominance of this model is inferred from evidence including internal lexical variation of phrases, flexible word order of many phrases, phrases that attract other words in collocation, and the power of context in predicting where a phrase will be used (Sinclair, 1991, p. 112). Sinclair (1991) uses this evidence to postulate that "the open choice principle could be imagined as an analytical process which goes on in principle all the time, but whose results are only intermittently called for (p. 114). While children seem to rely very heavily on the use of the idiom principle and formulaic chunking for language development, the situation for adults is less clear. There is relative consensus among researchers including Schmitt, Grandage and Adolphs (2004), Weinert (1995), Wray (2002), Wood (2001, 2002), Gatbonton and Segalowitz (2005) Bolanger (1989), and Yorio (1989) that formulaic sequences do not play as important a role in language development for adult second language learners as they do for children; however, Yorio (1989) admits that "the studies that do exist do not present a clear picture of the issues [formulaic language in adult learning] involved" (p. 56), and that formulaic language helps adults "economize effort and attention in spontaneous communication" (Wood, 2002, p. 5). There is agreement on the importance of formulaic sequences to language production and fluency, but the identification and classifications of formulaic language can be much more vexing. The following section, providing a review of psycholinguistic research

and prosodic features of formulaic language, will turn away from the topic that primarily concerns this research (writing), to focus on spoken language. This allows for a wider view of how formulaic language is studied, and provides a rationale and explanation for how lexical bundles and corpus research fit into the larger picture of formulaic production and storage of language.

3.1 Psycholinguistic evidence

There is ample psycholinguistic evidence that supports claims that formulaic language assists fluency and reduces the cognitive workload of speakers. This evidence will be examined through findings relating to fluency and speech rate, as well as phonological studies that highlight the prosodic features of formulaic language.

Intuitively it makes sense that the faster one is speaking the more one will rely on memorized or prefabricated chunks of language to assist them. Research by Kuiper (1996) on auctioneers and sports broadcasters confirmed this as individuals in these professions were found to use a higher quantity of formulas and fixed expressions. Wood (2001) also highlights the connection of prefabricated language to fluency noting that speakers "who have a greater repertoire of automatized chunks of language" are better able to "balance skills, attention and planning during speech" (p. 578).

The effect that formulaic language has on fluency is most often demonstrated by examining speech rate. From looking at breaks and pauses in speech, Pawley and Syder (1983) proposed the 'one clause at a time constraint' which states that "the most skilled or fluent talkers regularly pause or slow down at the end of each clause

of four to ten words...although they rarely do so in mid-clause" (p. 202). Even though Pawley and Syder say that one cannot encode a clause of more than 8-10 words in one encoding act, "some clauses are entirely familiar, memorized sequences...which the speaker or hearer is capable of consciously assembling or analyzing, but which on most occasions of use are recalled as wholes or as automatically chained strings" (p. 15); this explains how some strings can be much longer than 10 words and not include pauses or other disfluencies. Wood's (2001) research supports the notion that these "chained strings" of language are crucial for speakers to produce fluent sounding speech; Wood (2001) found that the "fluent units" between clauses are "complete grammatical structures" more than 50% of the time (p. 577), and that fluency is defined not only by these temporal distinctions of pause and hesitation, but through "their connection with pragmatics and structure" (p. 574).

In addition to the evidence from speech rate, phonological analysis of stress, rhythm, and intonation are also effective in demonstrating how our brains store some language in a formulaic manner. Lin (2010) conducted a survey of research done on phonology of formulaic language finding that "high frequency words and phrases undergo phonetic reduction at a faster rate than low and mid frequency sequences" (p. 177). As expected, sequences that are more frequent allow the speaker more chances to practice, and the more chances they have, "the more fluent the articulation is" (Lin, 2010, p. 180).

Another good indicator of prefabrication is when a sequence outstrips the grammatical competence of the speaker, especially in child learners. Young children who can produce complex multi-word utterances with proper articulation and

stress support this "complexity argument" (Lin, 2010, p. 179). If a child is able to produce speech that is phonologically coherent, and has articulatory and intonational accuracy, "it is highly probable that the whole sequence has been learned as an unanalyzed chunk and stored as a holistic unit in the child's mental lexicon" (Lin, 2010, p. 179). Idioms also show evidence of formulaic processing based on stress and intonation patterns. Ashby (2006) identified idioms by their accentual patterns and even-tone choices, which are often highly constrained (Ashby cited in Lin, 2010, p. 178). This means that idioms "have an accentual pattern different from the corresponding literal expression [but] the same as the least-marked literal version[9]. Experiments by Van Lancker (1981) show listeners are able to identify the difference in articulation and stress from idiomatic sayings (85% correctly) when the speaker emphasized the intended meaning (Weinert, 1995, p. 187).

3.2 Fixedness and non-compositionality

Formulaic sequences are also often identified by their degree of fixedness and semantic non-compositionality[10]. The non-compositionality of formulaic strings is derived from frequent use, which makes detailed semantic processing of each word in a sequence redundant. In terms of fixedness, many idioms, quotations, proverbs or other deliberately memorized sequences will not tolerate much alteration at all

---

[9] An example of accentuating the least marked literal version; 'have a BONE to pick' versus 'have a SCAB to pick'. An example of different stress in the corresponding literal expression; '*RUN* up the score' versus 'run *UP* the mountain'

[10] Semantic compositionality is the degree to which individual words in a string contribute to the meaning of the whole. For example, the expression "red herring" has low compositionality because neither "red" nor "herring" contributes literally to the meaning: 'a distracting and unimportant piece of information'.

to their form (or stress and intonation patterns). This can prove quite problematic for second language learners who might assume that normal generative grammatical rules apply to these sequences when many do not. However, fixedness is best seen as a cline or continuum. While idioms like 'red herring' might not take any variation, others like 'spill the beans' can be altered in many ways (she was a bean spiller, he spills beans often, she's full of beans) (Wray, 2002, p. 50). Wray (2002) argues that the level of compositionality in an idiom, or how much literal meaning the individual words retain, will help dictate the variability, since it is easier to manipulate a saying where each word has a clear semantic and grammatical role (i.e. beans=information, spill=to tell others).

3.3 Form and function

Through corpus research, and a great deal of intuition, researchers like Butler (1997) have suggested structural guidelines for identifying formulaic sequences such as "the majority of longer repeated sequences... begin with conjunctions, articles, pronouns, prepositions or discourse markers" (p. 76). However, the picture becomes less clear when one considers the range of fixedness within these formulaic sentence 'frames'. While some idioms like "kick the bucket" are structurally fixed (VP+NP) others can substitute words, have open-class words, or have 'empty frames' altogether. It becomes very difficult to pin down formulaic structure when the first word in the sequence might be an open-class and highly variable one, or there may be very few predictable words at all (i.e. the frame for responding to obvious questions (is+NP (the pope)+Adj (a catholic)), after all, "what precisely, is being stored, when all the words can be novel?" (Wray, 2002, p. 32).

This means that researchers must not rely entirely on corpus data since the boundaries between formulaic and non-formulaic cannot be decided by computer software alone.

Taxonomies for formulaic language have been many and varied. Becker (1975) offered categories of *polywords, phrasal constraints, meta-messages, sentence builders, situational utterances* and *verbatim texts*. Nattinger and Decarrico (1992) refined Becker's taxonomy into functions of *social interactions, necessary topics, discourse devices*, which fit into three formal categories of *polywords, phrasal constraints*, and *sentence builders*, Moon (1998) offers categories of *situational, hyper-positional, evaluative,* or *informational,* and Howarth (1998) postulates a hierarchy of phraseological combinations with *functional expressions* and *composite units* as the main nodes that branch into collocations and idiomatic/non-idiomatic expressions. While these taxonomies can be helpful, Wray (2002) warns that "definitions of formulaic sequences based on function tend either to become the victim of complex sub-categorizations by form...or else formal specifications are let aside, leaving lists which look as though they need organizing" (p. 52). Based on the difficulties of creating exhaustive formal taxonomies as mentioned earlier, it seems more reasonable to offer general functional categories that strike a balance between interactional and referential functions. No taxonomy will be appropriate for all examples or offer perfect criteria for organizing corpus data; the researcher must consider their text source carefully, since certain taxonomic models may work better for certain types of data (written discourse will undoubtedly display different functions than conversational speech, etc.).

While Chapter Three addressed theory, definitions and research related to formulaic language in general, the next section covers lexical bundles more specifically. Lexical bundles are not always formulaic sequences, but they share many characteristics of the formulaic language characteristics discussed in chapter 3. Chapter 4 will examine some definitions and common types of lexical bundle analysis, followed by a review of findings from lexical bundle research conducted over the past two decades.

Chapter 4 Lexical Bundles

In chapter 3, the formulaic sequences presented shared several key characteristics; they were highly idiomatic (low internal semantic compositionality, highly fixed) and they occurred quite infrequently when searched in large reference corpora. Lexical bundles on the other hand, are not judged with these criteria of semantic composition or fixedness, but consist of highly transparent, frequent collocations that "offer insight into the language used by writers in particular contexts" (Hyland, 2008b, p. 44). As corpus technology has evolved, the research on formulaic language has followed suit, exemplified by recent work on lexical bundles. While fixed expressions and idioms that make up the bulk of formulaic language in traditional phraseological research are classified based on their compositionality, and fixedness (as mentioned above), lexical bundles are found by using frequency and distribution parameters alone. Many lexical bundles are not metaphorical or semantically opaque; they are common, transparent expressions that have been institutionalized, meaning they have been repeated enough to gain lexicalization within a speech community, and serve a specific pragmatic purpose (O'Keefe et al., 2007; Biber et al., 2004). In their influential study Nattinger and DeCarrico (1992) see the lexical bundle as "a pedagogically applicable unit of formulaic language [...] with categorical form and discourse function" (Oakey, 2002, p. 113). It is not always straightforward to associate a bundle with a specific function and pedagogic value, but in general this is a criterion that will be enforced since otherwise the results from this study will have very little applicable value.

To begin identifying lexical bundles in a text, one must decide which frequency cut-off to use for the search. Since frequency is the key component in what qualifies a multi-word string as a lexical bundle, researchers are careful in justifying the parameters they set. For example, a study using a relatively small corpus often use a lower frequency rating, for example 10 times per million words, whereas a study using a larger corpus will use a cut-off of around 40 times per million words (Biber, 2006, p. 134). Despite the variation, it is standard for a study with a corpus of larger than 500,000 words to use a 20 times per million-word cut-off. Studies by Cortes (2004, 2006, 2008), Hyland (2008), Chen (2008) use this cut-off, but they note that it is common for the most frequent bundles to occur much more frequently than 20 times per million (60-200 times per million for the top 10 most frequent bundles in Hyland, (2008), and many top bundles were over 200 times per million in Biber, 2006). While it is a crucial part of finding bundles, frequency alone does not demonstrate a bundle's formulaic status. As Biber, Conrad, and Cortes (2004) said, "Frequency is only one measure of the extent to which a multi-word sequence is prefabricated. [...] We do not regard frequency data as explanatory. In fact we would argue the opposite: frequency data identifies patterns that must be explained" (p. 376). Wray (2002) is also skeptical about reliance on frequency, saying that "if a string is required regularly, it is likely to be stored whole for easier access [...] but it does not have to be"" (p. 31).

In addition to frequency, distribution parameters also help ensure that the sequences highlighted for explanation are not superfluous to the target register. Typically, studies will use a cut-off of 5 texts (a text can be a chapter, a book, or some other selection of discourse) or 10% of all texts. It is common for bundles to

occur in far more than 5 texts, and since some studies use hundreds of texts in their corpus, it is often more reliable to use a 10% distribution parameter. The point of using distribution in addition to frequency as a means of identifying bundles is to enforce the notion that these sequences are not idiosyncratic within the target register(s) (Biber, 2006, p 134).

Register is a key notion when it comes to lexical bundles; Biber et al. (2004) define lexical bundles as "simply the most frequent recurring lexical sequences in a *register*" (p. 376). Sometimes synonymous with genre, "both terms have been used to refer to varieties [of academic language] associated with particular communicative purposes" (Biber, 2006, p. 10). While there is some grey area between genre and register, in general register studies "have focused on lexico-grammatical features" whereas "genre studies have usually focused on socio-cultural actions" within a "culturally recognized 'message type'" (Biber, 2006, p. 11).

4.1 Grammatical and pragmatic features of lexical bundles

Lexical bundles differ from other prefabricated phraseological units considerably when it comes to grammaticality. Unlike idioms or fixed expressions, lexical bundles are made up mostly of incomplete grammatical structures. In Biber et al.'s (2004) work, they found that "only 15% of the lexical bundles in conversation can be regarded as complete phrases or clauses, while less than 5% of the lexical bundles in academic prose represent complete structural units. These sequences then might best be thought of as common frames that allow one to bridge other structural elements together when they are speaking or writing. Because these lexical bundles exist across the boundaries of multiple grammatical units, past research on

phraseological language has not considered them, even though they make up a significant part of discourse in certain registers.

Although not complete structural units, lexical bundles can be classified in terms of their grammatical parts. Biber et al. (2004) use a taxonomy that differentiates lexical bundles that incorporate *verb phrase* (VP) fragments, *dependent clause* (DC) fragments, and *noun phrase and prepositional phrase* (NP/PP) fragments (p. 381)[11]. Biber et al. (2004) found that these 3 structural groups of lexical bundles (which are subdivided further) differ depending on the register. For example, text from classroom teaching showed the highest number of total bundles, as well as the most NP/PP and DC fragments; textbooks had a very small number of VP and DC based fragments, but a higher proportion of NP/PP based bundles.

In addition to the structural taxonomy, Biber et al. (2004) also laid out three functional categories to organize lexical bundles: (1) stance expressions, (2) discourse organizers, and (3) referential expressions. Informed by Halliday's (1994) analysis of linguistic functions, these three functional categories can be defined in the following ways: referential bundles "perform an ideational function; they help writers structure their experience and determine their way of looking at things" (Cortes, 2004, p. 401). These bundles provide reference to physical, spatial, and temporal elements in the text. Text organizing bundles "are word combinations used to express textual functions which are concerned with the meaning of the sentence as a message in relation to the surrounding discourse" (Cortes, 2004, p. 401). This category includes bundles with framing, contrasting, resultative, and text-

[11] Some examples of bundles for these structural categories can be found in Appendix A

structuring functions. Finally, stance bundles are those that "express attitudes that frame some other proposition [in the text]" (Cortes, 2004, p. 401). These bundles allow writers to express epistemic certainty and attitude/modality, serving an interactional function between text and reader.

As with the structural categories, Biber et al.'s (2004) study shows that the discourse functions assigned to lexical bundles varies between registers. The study found that classroom teaching again has the highest proportion of total bundles, with referential and stance expressions composing the majority of the bundles, while textbooks and academic prose have the least number of bundles, and the bundles they do have tend to be referential. Many other studies including Chen and Baker (2010) and Cortes (2004, 2008) report using the same taxonomy as Biber's original taxonomy from 1999 (also the basis for the Biber et al. (2004) taxonomy). Despite this assertion, many bundles and categories in the Cortes (2004) study are classified differently, most likely due to different subjective evaluation of concordance lines and different registers of text (academic writing as opposed to speech and textbook writing).

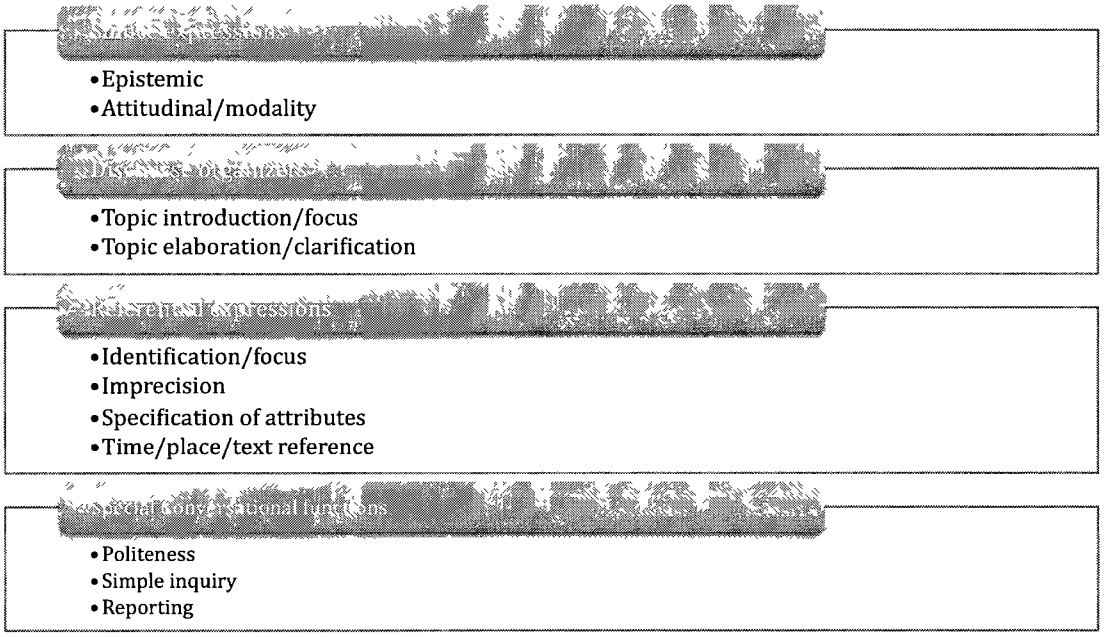

Figure 3.1. Biber et al.'s (2004) functional taxonomy[12]

Hyland (2008b) presents a different functional taxonomy that was "loosely based on Halliday's (1994) linguistic macrofunctions" (p. 49). Hyland's study which looked at the importance of clusters in several academic disciplines (including ALDS) and how bundle use varied in each discipline, provided the closest match in terms of research questions and text types to the present study. For this reason, Hyland's (2008b) taxonomy served as the primary guide for the functional classification of ALDS bundles. The final version of the taxonomy used for the upper-level and novice ALDS corpora is presented in figure 3.2. The justification for this choice and specific information about the taxonomy used for this study will be discussed in the

[12] The taxonomy in table 1 is somewhat simplified from the one Biber et al. used. There were some sub-sub-categories omitted regarding *personal/impersonal* stance bundles, which were not relevant to this study.

methodology chapter.

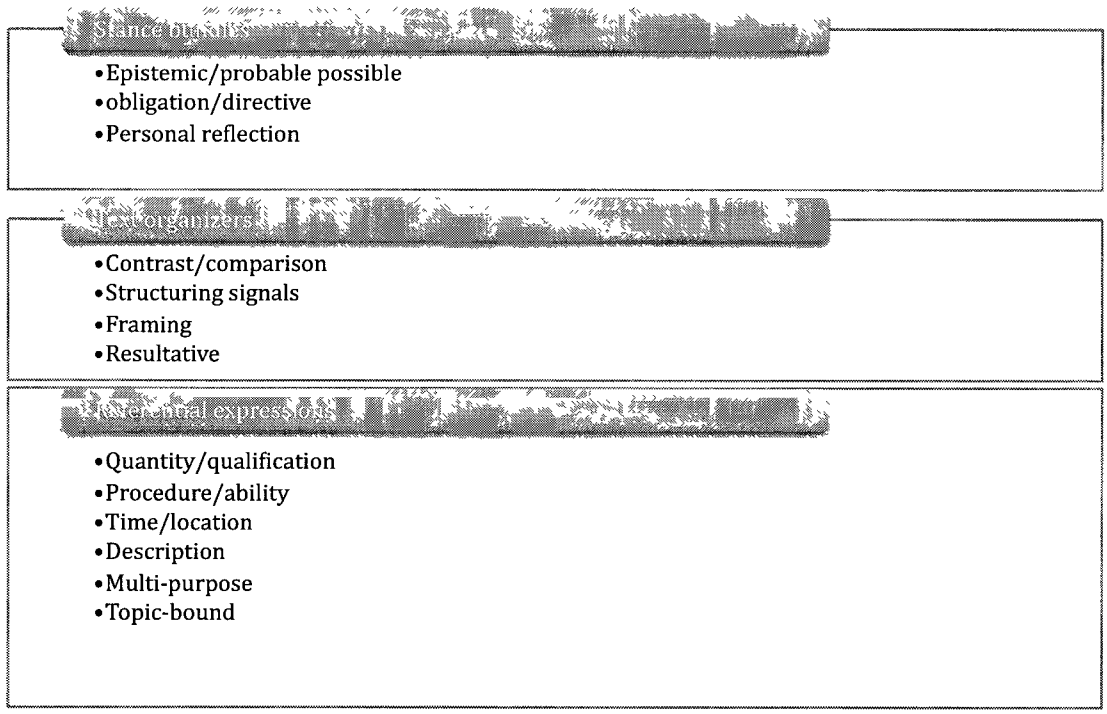

Figure 3.2. Functional taxonomy used for the ALDSN and ALDS-UL corpora

4.2 Lexical bundle research: spoken and written registers

Following the study on lexical bundles in university teaching and textbooks (2004), Biber (2006) wrote a book based on his finding from the TOEFL 2000 Spoken and Written Academic Language corpus (T2K-SWAL). T2K-SWAL is the same corpus used in the 2004 Biber et al. study, and is sampled from a range of academic registers including classroom teaching, office hours, study groups, on-campus service encounters, textbooks, course packs, and institutional written material (Biber, 2006, p. 24). It contains 2,737, 200 words from 423 texts. The corpus was analyzed for several types of patterns, including lexical bundles, which yielded surprising results. The analysis showed that there were almost twice as many lexical bundles in classroom teaching as there were in textbooks or academic prose. Biber

(2006) noted that, "it was surprising that textbook authors do not incorporate more lexical bundles in their writing, given the heavy reliance on bundles in classroom teaching" (p. 137). After applying the same functional divisions to the results, Biber (2006) had the following conclusions about the distribution of bundle functions across registers: (1) stance bundles are most common in classroom teaching and conversation (2) discourse organizing bundles most frequently occur in classroom teaching, and are somewhat less frequent in conversation (3) referential bundles are most frequent in classroom teaching and textbooks (p. 147).

Biber's work establishes that lexical bundles are far more common in registers that require quicker production of language and less opportunity to reformulate (as in spoken registers and some written registers, in contrast to textbook or academic prose); however there have also been several studies examining lexical bundles in student and expert writing that demonstrate the importance of bundles in those registers as well.

4.3 Lexical bundle research: University written registers

In Cortes' (2004) study on university history and biology writing, she found that there were significant differences in the writing between the two disciplines (far more stance bundles in biology writing), and "the use of target bundles at all levels still showed low frequencies in bundles that would be expected to occur more often in academic writing"(p. 414). While it is expected that students do not use the same bundles as academics since they are writing for a teacher instead of an audience of colleagues, it is important to note that the pragmatic functions of lexical bundles are not transparent or intuitive for learners: as Cortes (2004) concludes in her study,

"the problem arises when students face the functions that these expressions perform in academic writing, which sometimes are not as transparent as initially expected" (p. 421). This lack of transparency has important pedagogical implications because "the role of unconscious learning did not help students master the use of these expressions" (p. 421), indicating that explicit instruction on bundles is likely necessary for many learners.

In another more recent study, Hyland (2008) examines how bundles differ between the disciplines of engineering, biology, applied linguistics and business studies. The study found that there were extensive differences in structure, form, and function of bundles across the disciplines. Fewer than half of the top 50 bundles from each disciplines occur in all 4 registers, raising serious issues with the usefulness of general academic word lists.

4.4 Lexical bundle research: Second Language learners and EAP materials

There have been a number of studies comparing competence with lexical bundles of native and non-native English speaking students, as well as the bundles used in English for Academic Purposes (EAP) materials versus textbooks in the related disciplines. In Chen and Baker's (2010) study they compared the use of bundles in three types of writing: native-speaker (L1), non-native speaker (L2), and expert academic texts. They found that while the number of bundles used was similar between the two types of student writing, the "use of NP-based bundles differs the most amongst the three groups of writing" (p. 34). There were also notable differences between the student groups. Native speakers shared "a few features distinctive in academic writing such as the control of cautious language" (Chen and

Baker, 2010, p. 44) whereas L2 writing had a tendency towards "over generalizing and favoring certain idiomatic expressions and connectors" (Chen and Baker, 2010, p. 44).

Nekrasova (2009) and Levy's (2003) work lend further support to the notion that students with more lexical bundle competence will be more successful. Levy (2003) examined the link between levels of student writing and their use of bundles and found that "the level of writing class students ultimately placed in [...] is related to the register of lexical bundles used" (p. 83). Consistent with Biber's earlier (and later) findings, Levy (2003) found that advanced students were more competent with referential bundles, which have a higher proportion in academic prose and textbooks. Nekrasova (2009) also found that lower proficiency English learners were not able to accurately produce as many lexical bundles as did L1 speakers and higher proficiency learners (p. 674). Her study supported the notion that referential bundles are more difficult to acquire than discourse organizers, and like Cortes (2006), concluded that L2 learners need more direct instruction with lexical bundles to "notice the contexts in which these units are typically used, as well as the discourse functions they perform" and "become aware of how using lexical bundles can help them improve their writing" (p. 674).

In addition to studies on bundle use between L2 and L1 student writing, Chen (2008) found from her comparison of English for specific purposes (ESP) material and introductory Engineering textbooks that "there exists a gap in language use between the entry-level discipline textbooks and ESP textbooks" (p. 123). Chen's (2008) findings, along with those mentioned above, emphasize that less proficient L2 (and L1) learners are in need of more instruction on the use of lexical bundles,

and perhaps different materials than those currently used in ESP and EAP classrooms.

Along the same lines as Chen's study, Wood (2010) examined the occurrence of lexical bundles in EAP textbooks. From looking at sub-corpora of EAP textual material (mostly readings) and instructional material from EAP textbooks, Wood (2010) located 65 clusters that appeared at a frequency higher than 20 words per million. Although the analysis uncovered a surprisingly high number of bundles overall, there was a significant difference between the bundles found in the reading material from textbooks and those found in the instructional material; in short, there was no consistent effort made to present students with lexical bundles in their readings and text book material, and many of the bundles found "were not clearly relevant to academic reading and writing" (Wood, 2010, p. 103). Wood (2010) concluded that the lack of attention paid to lexical bundles in EAP materials does not adequately prepare students to read and write academically because the bundles in these materials are not authentic; the EAP textbook does not contain the most frequent sequences that are used in the target discipline, and these sequences are not effectively emphasized for instruction when they do occur.

In chapter 5 of the literature review, corpus research and lexical bundles are set aside for a broader look at theory and research in composition studies, writing development, and genre theory. Research in this area adopts a social view of discourse, focused less on text features and more on the social context of writing. This expanded view of writing can offer a better understanding of what makes one a successful novice writer, and how university students develop from novice writers into experienced members of their field; this chapter is used as a basis to

contextualize and triangulate some of the findings from lexical bundle analysis, which address the same issues of novice writing and transition within and across university registers.

Chapter 5 Writing development: A broader perspective

What it means to be a literate and successful writer, academic or otherwise, has been studied from as many angles as there are researchers (Bazerman, 2010). The focus of this study is on a relatively narrow part of writing, academic course work in a single discipline, and takes a micro-level approach to exploring how writers write the way they do at certain levels, and how one frequent clusters in their writing change as they transition to higher levels of academia. In order to better understand and explain results from the corpus phase of this study, research from genre studies, writing development, and composition will be considered. This chapter will broaden the scope of the study to consider what knowledge and approaches are generally ascribed to novice and experienced writers in university disciplines, as well as how genre knowledge has typically been defined and portrayed in pedagogical contexts. This last part of the literature review not only contextualizes the corpus and formulaic language research, but also provides a foundation for interpreting the results produced in the questionnaire and interview portion of the study.

5.1 Overview of approaches to the study of academic writing

Writing is not part of our innate acquisition of language. While humans are predisposed to learn to speak and understand oral language and gestures almost from the moment they are born, it is no certainty that a person will learn how to write or read any language, unless they are taught. The process of becoming literate is one that is highly valued in communities around the world, and championed as a key factor in achieving 'developed' status according to many nations. Literacy,

however, is not a simple matter of having it, or not having it. As Gee (1990) describes it, there are many 'master myths' in our society, and literacy is among them. Gee asserts that the ideology of Western countries such as Canada and the United States does not adequately define the term literacy when they include it on a government census, because there are many types of literacy with widely different societal values. This idea that literacy should not be defined in binary terms is an underlying assumption of the current study. In other words, it is possible to be a literate writer or reader in one context, but not have the skills to succeed in another. As Beaufort (2004) says, "numerous studies of writers crossing disciplinary boundaries, or moving from on social context to another or from on genre to another, support the view that writing is not a one-size-fits-all sort of skill" (p. 138). This process of crossing disciplinary boundaries or academic contexts (first year to Master's level) can be a significant learning curve even for students who have strong writing skills in other areas (creative writing, journalism etc.). The view of writing as a skill that one either has or doesn't have results in a preoccupation with the product of writing in academic and educational contexts.

Teachers often assign writing as means of evaluating students' grasp of course material and readings, implying that once all of one's thinking and understanding has taken place, writing is a final step where information is presented to the evaluator. This product view of writing perpetuates a type of teaching that assumes students will learn to think and develop complex cognitive strategies for understanding material in a stage separate from writing. Bizzell (1982) argued that, "the new demands on us as teachers can only be met, it seems, by a reconsideration of the relationship between thought and language. We are pretty much agreed, in

other words, that what we need to know about writing has to do with the thinking processes involved in it" (p. 388). This statement by Bizzell frames writing and the teaching of writing as a process that is tied in with thinking itself. Ong (1986) describes writing as a technology, like Internet, photography, or telecommunication, which allows humans to extend their natural capabilities of thinking and communicating. The technology of writing can be used not only to distribute information and communicate, but also as a means of developing new ideas and making connections that would otherwise not be discovered. This idea of writing being important as a thinking tool as well as a communication tool is important in understanding the broader scope of the current study. Past research has emphasized that in order for writers to be successful, they need more than the linguistic tools necessary to put ideas on paper; they require the more tacit knowledge associated with each writing genre about how writing and thinking are conventionally conveyed. The challenge of learning these conventions that pertain to each writing genre (and are often beyond the basic grammaticality of the language) is also a primary focus of Pawley and Syder's (1983) formulaic language research. They explain that although there are many options for accomplishing a rhetorical objective that are equally grammatical, only a small number (or only one) will sound native like. For example, when a man makes a traditional marriage proposal he does not say 'can I possibly marry you ', but rather 'will you marry me'. Any one who uses the former construction would be understood, but seem awkward because of the context and choice of words. The same principle can be applied to learning the genres of academic writing, but the issue is how to identify which forms

of writing are 'native like' to the discipline, and how teachers should go about helping students acquire them.

5.2 Genre and Academic Writing

Genre has been a key concept that has guided a great deal of research not only in writing studies, but in corpus linguistics as well. While I previously discussed the idea of genre in relation to academic disciplines and how corpus studies have shown differences in the language used across registers, the topic of genre in writing studies is covered from an angle that incorporates the interaction of social factors, context and form together.

In contrast to literary studies, where the term genre is often used to define a general category of texts, for the sake of organization, Giltrow (2002) characterizes genre as something that not only has a common set of formal features, but also is connected to situation. Activities such as essay writing, lab reports, thank you notes, and marriage certificates all constitute different genres with specific forms and situational characteristics. As Giltrow (2002) explains,

> Consider the thank-you note as a genre. People who know this genre not only know how to compose the note — what to mention, how much to say, how to begin, how to conclude, what kind of writing materials to use — but also *when* to do all this...the thank-you note genre is made up of not only a characteristic type of written expression but also of the situation in which it occurs" (p. 24)

Whether considering a thank-you note or more formal academic genre, genres can also be seen as socially situated, from Miller's (1984) groundbreaking work. Miller saw genres as "typified rhetorical responses to situations that are socially interpreted or constructed as recurrent or similar; genres are thus social actions" (as cited in Dias, Freedman, Medway and Paré, 1999, p. 21). If genres are dynamic

social actions that depend heavily on when and with whom they are being used, it is thus very difficult to present a learner with genre information if it is not contextualized within the activity or situation that is common to that genre. Learning how to participate in a written genre is a complicated process, both from the perspective of the researcher and the teacher. As with any socially situated action, producing a genre is a dialogic activity that is inevitably connected and influenced by many other genres; as Tardy and Swales (2008) say, "many academic disciplinary genres explicitly draw on previous texts, as they construct and represent an intertext of prior research" (p. 570). This interconnection of texts and their social actions was alternatively characterized as being similar to ecological systems in nature (Cooper, 1986). Cooper proposed that, "all the characteristics of any individual writer or piece of writing both determine and are determined by the characteristics of all the other writers and writings in the system" (p. 368). With many theorists agreeing that writing actions and the genres that associate with them are socially situated and dialogic, it is has been much less clear how individuals come to learn and master the genres of academia or the work place. The writing specialist Bazerman (2002) characterizes this complicated process of genre engagement in the following way:

> Individuals habituate places where particular genres are enacted; over time, they gain access to participation in those genres; eventually, they start writing in those genres, begin thinking in ways that result in the genre, and develop and commit to an identity within the genre's domain (as cited in Tardy and Swales, 2008, p. 571).

Piecing together what someone knows when they can successfully produce a genre, and how best to train a novice writer in acquiring a genre are vexing tasks that have

generated much debate and diversity of explanation. This debate in the field of genre studies can be seen quite starkly between the 'Sydney School' of genre researchers led by Halladay, Martin, Christie, Rotherie and others (Dias et al., 1999) and the North American genre theorists (also known as new rhetorical genre theory) including Medway, Freedman, Paré, Smart, Artemeva, and others. In the Sydney school of genre studies, genre is recognized as socially situated, but a special focus is placed upon the ideological and power relations implicated in the learning and dissemination of genres. In an effort to remedy some of the inequality of education and failures in the Australian curriculum, Sydney School proponents "advocated teaching the textual features of such genres to disadvantaged students in order to empower them" (Dias et al., 1999, p. 22). While this approach was perhaps well suited to students' needs in Australian schools, Freedman and Medway (1994) argued that teaching genre through a focus on textual form, and giving students a deconstructed template of the target product was a step back towards the decontextualization of genre. As Dias et al. (1999) argue,

> It is not simply the textual aspect of accepted genres, in the general form in which it can be imparted outside the specific site; one also needs knowledge of the culture and the circumstances, and one needs to understand and take on the local *purposes,* the social motives that prevail in that setting" (p. 22)

In an effort to better understand these social motives and the interaction between motivation and social action in genre production, some researchers have combined aspects of activity theory with the new rhetorical genre approach (Dias et al. 1999; Freedman and Pringle, 1980; Russell, 1997; Artemeva, 2005, 2008; Artemeva & Fox, 2010). Originating from the work of Leont'ev (1981), activity theory contains three levels of analysis used to explain what a person is doing when they engage in

discourse: *activity, action,* and *operation.* These levels explain how the basic *activities* we engage in such as learning or playing, are governed by socially constructed motives, and that our goals for these activities are realized by the *action* taken by an individual. This means that "the *action* of reading, depending on the goal, can realize the *activity* of play, or work, or learning" (Dias et al., 1999, p. 25). *Operations* are the final level of analysis that refers to the circumstances or conditions affecting the action and activities. Dias et al. (1999) suggest that activity theory is a very natural combination with genre theory, and can be used to make a convincing argument against the Sydney School 's view of genre:

> When they teach a genre we might argue that they are teaching an action, not an activity; in our terms that amounts to saying they are not teaching the genre. What people need to learn is to engage in the activity. It might be argued that familiarity with contributory actions is at least a great help in acquiring an activity, and in principle that is true. In the case of much situated writing however, the action has to be customized to suit the situation through a form of intelligent reconstruction in the light of the sort of knowledge...that genre teachers cannot impart (p. 28).

This debate on how to approach the teaching of genre has had important implications for researchers who have investigated the writing development and transition of students' writing at the university level. The questions of what genre knowledge students possess when they are novices in a particular discipline, as well as how they can go about acquiring the necessary genres to succeed were investigated by Artemeva and Fox (2010) in a recent study.

Focusing on an engineering department at Carleton University, Artemeva and Fox conducted a study that investigated how students' reported knowledge of the genre matched their performance by eliciting the genre knowledge of first year engineering students as they entered a class on writing in the discipline. Students'

reported knowledge of genre was then compared with the lab reports they produced. The study found that "overall, students' awareness of genre differences and their ability to identify and report genre features did not enable them to produce a text in the requested genre" (Artemeva and Fox, 2010, p. 497). Although most of the students were not able to produce an acceptable lab report, it was found that many of the students "were able to recognize and articulate rhetorical and textual features of different genres, including the genre of the technical report" (p. 502) without ever having been explicitly informed about the features of the genre. These findings support the work by Freedman (1994), Wardle (2007), and Artemeva (2005, 2008, 2009) and Dias (2000) that genre knowledge is often gained by participating in various work and academic activities, and is informed by knowledge of other genres that relate to those tasks. Artemeva and Fox's work lends further support to research by Freedman (1994) and Wardle (2007), which argued that, "in order to produce texts that are meaningful and recognized as appropriate by a relevant community of practice, a writer needs to be immersed in the practices of a particular community" (p. 482).

While participation and immersion in a community of practice has been established as an important part of learning to produce genres successfully, many other case studies and longitudinal investigations of novice university writers have looked at the other factors that contribute to learning the discourse conventions and genre knowledge specific to university registers.

In a longitudinal case study undertaken at a private American University, Haas (1994) investigated one biology student's (Eliza) development as she progressed from first year to graduation. Using data from 11 extended interviews

conducted over four years, Haas studied how Eliza's understanding of discourse and the rhetorical aspects of writing and reading developed over her time at university. At the beginning of her studies, Eliza saw texts as isolated, independent forms, and approached her reading and writing without much inter-textual awareness. As Haas (1994) said, "some readers seemed to rely overly on this strategy [re-reading], invoking it again and again. They attempted to understand the text not by moving out from it to a rhetorical context, but by moving in and focusing ever more closely on the text as an object" (p. 50). Over her years in school, Eliza's view of discourse shifted, as she was able to move away from her arhetorical view of texts, gaining more awareness not only about how texts were connected to each other, but what importance they had to her as a potential member of the discipline. Instrumental in this development was Eliza's work with other students from higher years of her program (mentors), and interaction with the genres she was tasked with reading and producing. Like the engineering students from Artemeva and Fox's (2010) study, participation and interaction played crucial roles in the development of genre knowledge, and Haas's (1994) research suggests "that in order to understand the development of student writing abilities researchers must investigate students' preexisting conceptions of texts" (Rogers, 2010, p. 367). In McCarthy's (1987) longitudinal case study (the earliest such study conducted), he followed a student, David, over a two-year period as he started university and eventually found a subject area to specialize in. Like Eliza in Haas' (1994) study, David also viewed each class and writing assignment as totally separate and disconnected from each other. McCarthy describes David as being a 'stranger in strange lands' because "he was unable to identity the connections among tasks...the student interpreted each

writing situation as being totally different from other writing tasks" (Rogers, 2010, p. 366). David did eventually start to experience some development in his understanding of academic environments and genres, aided greatly by the mentorship provided by his professors and other work with peers and their writing (Rogers, 2010).

In another study concerned with novice level university writing development, Beaufort (2004) followed a history student, Tim, over his first three years of university study. Beaufort says that developing writers like Tim, "must wrestle with writing –process, rhetorical/social-contexts, and genre demands, as well as vocabulary, sentence structure, and so on" (p. 144). Beaufort (2004) goes on to describe academic writing expertise as having "rich and specific domain knowledge that is well structured for easy access in routine situations" and "strategic knowledge, or the 'how' of problem solving in the domain" (p. 138). These characteristics Beaufort outlines suggest that assisting students with prefabricated or formulaic expressions to assist with 'domain knowledge' as well as vocabulary and sentence structure concerns could be important in helping novice writers progress. As Tim developed in his history program, he struggled to understand the rhetorical purpose behind many of his assignments; Beaufort (2004) concludes that, "there were small gains in discourse-community knowledge, in subject matter knowledge, and in efficiency in accomplishing writing tasks. Critical thinking skills, genre knowledge, and rhetorical knowledge showed the least growth (p. 173). Beaufort found that most of the improvement Tim showed as a history writer seemed to happen inferentially instead of through explicit teaching (Rogers, 2010). This finding again supports the research of Artemeva and Fox (2010), Wardle

(2007), Freeman (1994), and Haas (1994) who emphasized the importance of immersion in the practice of a discourse community and found that a great deal of genre expertise comes through experiencing learning, writing, and working in the community of practice.

The findings from these important case and longitudinal studies will be addressed again in relation to the findings from questionnaire and interview sessions with ALDS student writers (section 7.3). In the next section, methodology is discussed in two phases: the corpus study of lexical bundles, and questionnaire/interview methodology.

Chapter 6 Methodology

In order to address questions about how lexical bundles were used in different levels of ALDS writing as well as what importance formulaic language instruction and lexical bundle control might have on academic writing development, I used a quantitative dominant, partially integrated mixed method approach (Tashakorri and Teddlie, 2010). The subordinate qualitative phase of the study (questionnaire and interview) was classified as 'partially integrated' because analysis was conducted in a sequential non-iterative fashion (Nastasi, Hitchcock, and Brown, 2010). In other words, the quantitative corpus analysis was not altered or directly influenced by the qualitative questionnaire/interview findings; the questionnaire interview results were analyzed after the quantitative corpus phase, and used to triangulate analysis of lexical bundle usage and characterization of novice/upper-level ALDS writers. The methodologies for the quantitative corpus phase and qualitative questionnaire/interview phase are discussed in separate sections, each including the following sub-sections: participants/representativeness, materials, procedures, and analysis.

6.1 Methodology for the corpus phase

6.1.1 Participants and materials: Representative samples for three corpora

Three corpora were compiled in order to identify and compare lexical bundles in ALDS writing. All texts that were not published (public) material, were only collected and analyzed after ethics approval was granted by the Carleton University Research Office. The first corpus of upper-level ALDS writing (ALDS-UL) was composed of student essays from the working papers journal Carleton Papers in

Applied Linguistics (CPALS), course papers from ALDS graduate courses from 2010 and 2009 semesters, and seven fourth year ALDS term papers that achieved grades of A or higher. Although there was variation in the quality of the writing included in the upper-level corpus, all three sources of work were filtered on the basis of grade assigned to the paper and variety of content, ensuring that the corpus was as representative as possible of upper-level ALDS writing in the Carleton program. CPALS papers made up the majority of this corpus because they had been selected as exemplary writing from the discipline. CPALS writing contained some range of writing skill and language command, but each paper met a standard well within the scope of upper-level writing for the program. The CPALS papers were taken from three different years of publication: 1995-1996, 1999-2000, and 2004-2005. Graduate course papers were taken primarily from a Curriculum Studies class, but some were included from other courses including Language Testing, Systemic Functional Linguistics, and Directions in Applied Linguistics Studies. Included in the grad course papers were mini papers from the curriculum class of 2009 and 2010. Mini papers were four to five page examinations of topics in Applied Linguistics and Discourse studies. These papers were meant to include an introduction with thesis statement, literature review section, and critical analysis of the topic in relation to past research and/or personal experience. As the professor who assigned the graduate curriculum mini papers and ALDS 1001 mini papers was the same, a very similar rubric and assignment sheet were used. Therefore, these 11 graduate mini papers offered a very good comparison to the 1st year mini papers that make up the novice corpus.

In total, the ALDS-UL corpus contained 197 382 words from 35 texts (texts were an average of 5 558 words), written by 35 different authors. In previous studies on lexical bundles using corpus data, researchers have used much larger corpora, well over 1 million words (Biber et al., 2004; Biber, 2006; Cortes, 2006, 2008; Hyland, 2008; Chen and Baker, 2010). Cortes (2008), who has been a leading researcher in recent years on the use of lexical bundles, had this to say about corpus size: "It is, then, highly advisable to work with at least one million words to identify lexical bundles in a corpus and to draw reliable comparisons when using more than one corpus" (p 46). Despite the one million-word standard for specialized corpora, there have been many studies that used much smaller samples of text to investigate lexical bundles. Chen (2008) and Levy (2003) both used corpora that were well under one million words (247,000 and 127,000 words respectively) to compare lexical bundle usage in different contexts. The use of smaller specialized corpora can even be preferable in many cases; because smaller corpora allow researchers to carefully select texts that represent the target register, these specialized corpora are often better suited to look at specific genres than larger reference corpora. The relatively small size of the ALDS-UL corpus limits the conclusions drawn from the findings, but it is large enough to provide interesting insights into the types of sequences and bundles used by ALDS grad students.

6.1.2 Participants and materials: The ALDSN and ALDSI corpora

Built as a representation of novice ALDS writing to compare with the upper-level corpus, the ALDSN corpus consisted entirely of mini paper assignments (about five pages double spaced) from a single introductory ALDS course. These mini papers

were selected to represent novice ALDS writing because there was no other introductory course in the discipline during the time frame of data collection that offered the same amount of essay style work. While mini papers are a different genre from Master's level course papers or CPALS publications, they do share similarities specific to ALDS writing, such as citation format, expression of arguments, subject matter, and organization of introductions and transition elements. While upper-level papers were five times as long on average and contained more varied assignments (studies, literature reviews, mini papers) this variety and length is typical of upper-level writing in the discipline; novice mini papers could not be compared to short, mini paper style assignments alone because they are not produced in large numbers, or the most typical course papers at the upper-level of ALDS. Despite the considerable differences between the first year mini papers and upper-level writing assignments, the ALDS 1001 class was the best choice to represent novice writing because of the low experience level and lack of familiarity many students had with ALDS courses or writing. In total, the ALDSN corpus contained 105 papers (counted as 65 texts because of the multiple papers for some authors) and 112 244 words; the papers were an average length of 1045 words.

Of the participants who agreed to submit work for the ALDSN corpus, 30 were first years, 16 were second years, 10 third years, two fourth years, and six were students that did not submit any personal information. There were also three students from the Certificate in Teaching English as a Second Language (CTESL) program, which is a one-year diploma in ALDS that involves a practicum. Although the CTESL students were generally older and had completed more university

courses in the past, ALDS 1001 was likely still their introduction to the discipline. It is important to note that some of these 65 students had a great deal of experience writing for other purposes or academic disciplines, thus the 'novice' label only refers to their experience as ALDS writers. Students were classified as 'true novices' in ALDS if ALDS 1001 was their first ALDS class in university (this could only be confirmed for students who were not in the CTESL program).

In addition to year of study, language background information was gathered at the beginning of ALDS 1001. There were only 12 students that listed English as their second language, and of those 12, several achieved very high grades in the course; due to this small number of non-native speakers, the ALDSN texts were not tagged for language background. Studies carried out in 2011 by a class of ALDS Masters students support this decision to ignore the second language factor: using a database of demographic information of the same ALDS 1001 class paired with grade outcomes for each student, several group studies found that lingualism (the languages spoken by the students) had little to no bearing on the final grade or mini paper marks students received in the course.

To offer some measure of control to the novice and upper-level comparison, a third corpus consisting of 'intermediate' level writing in ALDS (ALDSI) was constructed. The label 'intermediate' was chosen because the writing came from two second-year ALDS classes[13] that included second and third year ALDS students. These papers were much closer in length, style, and topic-focus to the upper-level writing in the ALDS-UL corpus.

---

[13] The second year classes were "Oral language and society", and "Second language teaching methodology".

With assignments including term papers and take home essay question responses, the ALDSI corpus contained 19 papers, nine from one class and 10 from the other, written by 19 different students. The corpus contained 118 949 words with an average text length of

6 176 words.[14]

Despite its greater overall size and length of individual texts, the ALDSI was not chosen as the primary counterpart to the upper-level corpus because the writers in the second year classes came from such a wide mix of academic and ALDS writing backgrounds. Some students in the second year class were taking the course after having already completed many of the other required upper level ALDS courses and because full background information was not available for all students, they could not be presented as a uniform representation of any one writing level. There was still some variation of writing experience from the students in ALDS 1001, but with much more information about their backgrounds, and far more true 'novices' to choose from, the 1001 mini papers were the best option for the ALDSN corpus.

6.1.3 Procedures

Once a representative sample of texts for each corpus was decided upon, each text was edited to remove headings, reference lists, extra spacing, long sections of quotation, tables, lists and other superfluous bits of text. Short quotations as well as citation information were left intact. There were six cases in the ALDS-UL corpus where there were two mini papers authored by the same student (two mini papers

[14] The average text length is inflated because some texts included several separate pieces of writing from the same author

were assigned in the graduate curriculum course). In these cases mini papers were combined into a single file to eliminate confusion when applying distribution criteria. As there were multiple papers for each author in the ALDSI corpus, they were also merged into a single plain text file for each author.

In order to digitize the ALDS 1001 mini papers and hard copies of CPALS papers, each paper was scanned into a word document using optical image recognition technology. After being cleaned of headers, spaces and other erroneous text, each paper was transferred into a plain text document and given a case number with a letter to indicate if it was the first or second mini paper for that author, or which CPALS edition it had come from. Because there were some ALDS 1001 mini papers provided later in the data collection process that were not marked with case numbers, the corpus contained three papers from unknown authors. There were also eight papers that had to be removed due to TA or professor comments in the text, which disrupted scanning software.

6.1.4 Frequency and Distribution

Although the ALDS-UL and ALDSN corpora were smaller than the ideal of one million words, they were assigned frequency cut-offs of 40 words per million. Studies by Cortes (2004, 2008), Chen (2008) and Hyland (2008) use a 20 times per million cut-off, while Biber (2006) uses a 40 times per million cut-off, saying, "the actual frequency cut-off point to identify lexical bundles is somewhat arbitrary" (p. 376). Studies with smaller corpora, including Levy's (2003), set a cut-off of 10 times per million, which was the lower range originally suggested by Biber et al. (2004). Because even the larger ALDS-UL corpus was relatively small, the very high cut-off

of 40 times per million was chosen in order to assure a stronger relationship between bundles and formulaic processing, as well as to eliminate any red herring sequences that could come up as a result of low frequency parameters. An additional reason for this high cut-off was the unconventional comparison of long course papers with short first year papers. If any meaningful results were to be found, the frequency cut-off needed to assure that qualifying bundles were truly integral to the novice or upper-level registers, and not idiosyncratic uses. Because the ALDS-UL, ALDSN and ALDSI corpora were not at an even one million words, the frequency cut-offs were normalized. To normalize the frequency number, "number of tokens [were] divided into the number of words in the corpus and multiplied by, for example, one million, or any previously- established norming number" (Cortes, 2008, p. 46). Thus, a word string in the ALDS-UL corpus would be considered a lexical bundle (based on normalization from 40 times per million) if it occurs eight times or more in the text. The raw frequency count comes out to 7.89, which was rounded up. Chen (2008) used a cut-off of 32 times[15] per million in a corpus of similar size (247 346) to the ALDS-UL and justified her frequency cut-off because "the cut-off frequency for four- word lexical bundles was set at the high level of 20 times per million words [in Cortes, 2004], indicating that a narrower and more specific scope of target language tends to have more lexical bundles that reflect formulaic language use" (p. 65). For the ALDSN corpus which was 1.7 times smaller than the ALDS-UL, the same frequency cut-off of 40 times per million was used.

---

[15] There was an error in Chen's (2008) study for frequency normalization. She used a corpus of 247,000 and stated that her frequency cut-off was 20 times per million. However, she normalized that number to 8, which is actually a frequency rate of 32 times per million.

Although this number is very high for a small corpus, it was important to keep the same frequency cut-offs for both corpora in order to validate the comparisons.

Distribution parameters were set at a five-text minimum. In other words, a sequence needed to occur in at least five texts from different authors in order to be counted as a lexical bundle. In addition to the text minimum, it was required that each sequence occurred in at least two of the texts from different courses or assignment types. Therefore a bundle that occurred only in one edition of CPALS, only 4th year papers, or only in curriculum mini papers was not considered a valid bundle. This was done to ensure that bundle statistics were not being skewed by topic-specific language use related to the subject of the course. It is uncommon for a study to use a distribution of less than five texts, or 10% of all texts considered. Levy (2003) and Chen and Baker (2010) used a three-text minimum, however Chen (2008), Biber et al. (2004), Biber (2006), Hyland (2008), and Cortes (2004, 2008) all use five texts or 10% of all texts for their research. Again, because of the issues with comparability between the novice and upper-level corpora, the more conservative five-text minimum was chosen to avoid results that were overly specific to an individual writer or subject matter. Because the ALDSN, ALDS-UL (and ALDSI) contained texts of very different length, and had very different totals of text files, setting a common distribution cut-off was an issue. I decided to use the same five-text minimum for the ALDS-UL and ALDSN despite these differences. The distribution minimum could not be lowered for the ALDS-UL because the texts were relatively long, and a lower minimum could have resulted in far more idiosyncratic results. It seemed necessary to maintain the five-text minimum for the novice

corpus also, since a lower minimum would fall far below the standard 10% distribution limit normally deemed acceptable.

6.1.5 Analysis

The ALDS-UL, ALDSN and ALDSI corpora were analyzed using Wordsmith Tools 4.0. Hyland (2008) and Chen (2008) both used Wordsmith Tools 4.0 for studies involving analysis of lexical bundles. Wordsmith Tools was selected over other programs (such as Monoconc Pro, KfNgram, or N-gram Phrase Extractor) because of Wordsmith's ability to link concordance lines with the source text and its ability to create lists of frequent clusters. Ari (2006) compared Wordsmith Tools 3.0 to KfNgram and N-gram Phrase Extractor and also concluded that Wordsmith Tools had the best performance in terms of distribution and concordance information. Although a newer version of Wordsmith is available, it was not necessary to upgrade based on the small size of ALDS corpora and the limited range of Wordlist functions used for analysis.[16]

Wordsmith has three primary functions used to analyze text data: *wordlist*, *keywords*, and *concord*. Only *wordlist* and *concord* functions were used in this study. The first step for generating a list of frequent clusters from the text files was to create a *wordlist index*. Unlike a regular wordlist which only provides information on which words are the most frequent in relation to other words, an index saves two files for each word entry, *.token, and *.type. The *.token file saved for every word in the list provides information of how that word relates in position to every other

---

[16] The improvements included in Wordsmith Tools 5.0 were faster installation, easier file reading, corruption detector, text converter, and concordance organizer. The concordance organizing function would have been useful to me, but by no means necessary.

word in the corpus and the *.type file stores information about the individual word type (Scott, 2011). Using the information stored about each word's location in the text, an index can be used to generate a word list of clusters, specified for length, frequency and syntactic parameters (Figure 6.3). Figure 6.3 shows the cluster word list index with the option box for computing clusters.

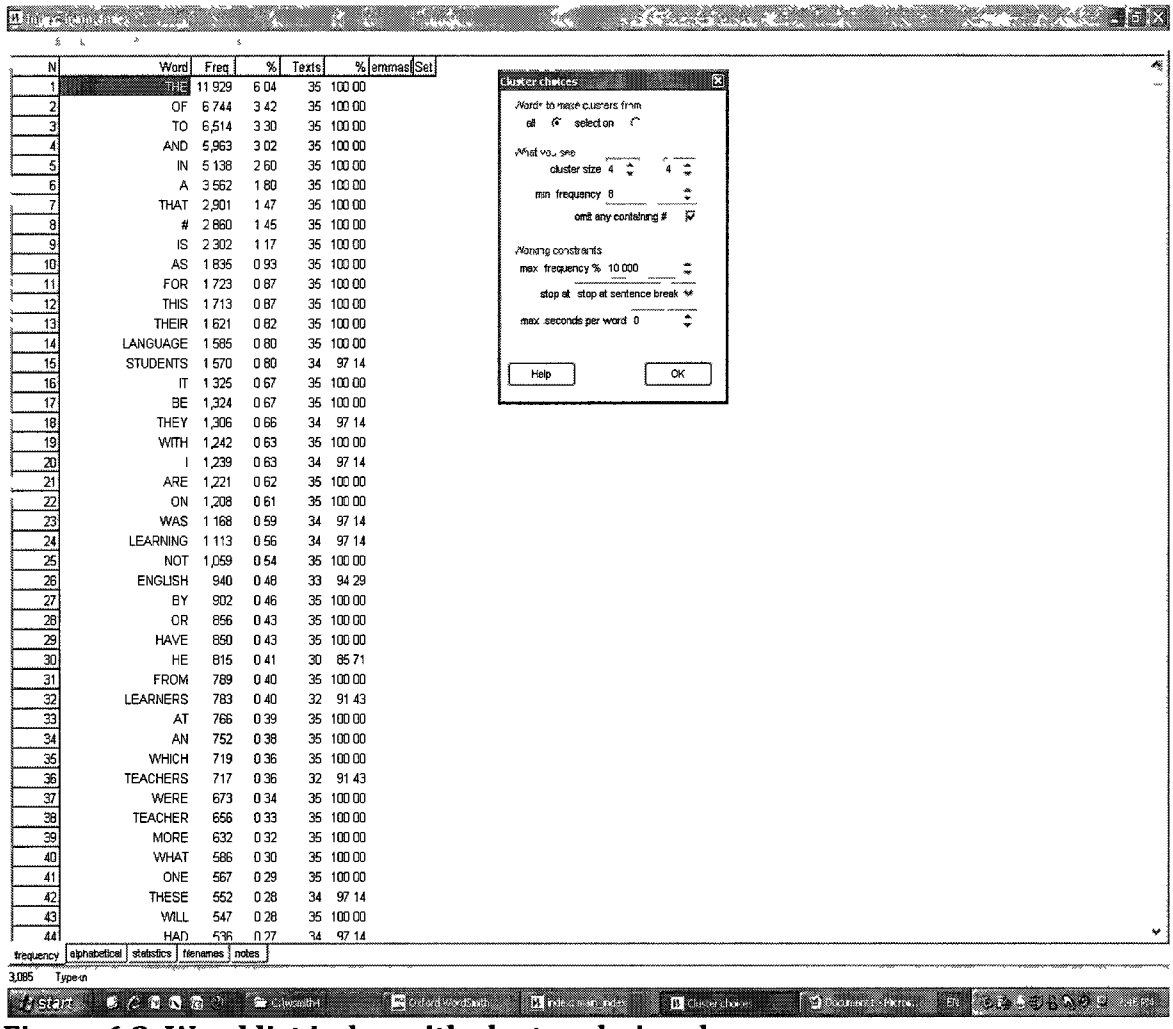

Figure 6.3. Word list index with cluster choices box

The initial word list index can be transformed into a cluster list by selecting compute: clusters (Figure 6.3). When this option is selected a box is displayed (shown in Figure 6.3) where parameters for frequency, sentence boundaries, and number inclusion can be adjusted. With this option box, one can choose to search for sequences that cross sentence boundaries, or those which do extend beyond periods. The analysis in this study searched for bundles that did not cross sentence boundaries in order to insure more pragmatically specialized bundles.

Once the parameters are set, a cluster index is created (Figure 6.4).

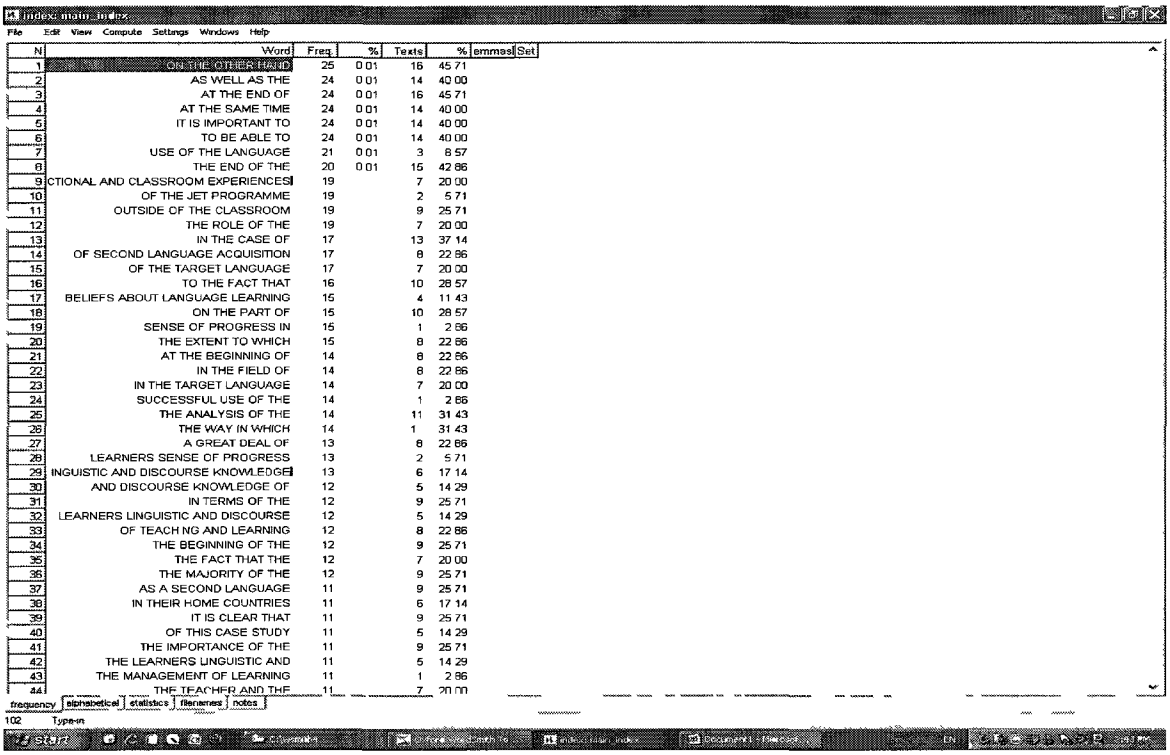

Figure 6.4. Cluster list from the ALDS-UL corpus with frequency and distribution information

The clusters are organized according to frequency, followed by the percentage of the total text, how many texts it occurred in, and what percentage of texts contain the cluster (Figure 6.4). After the cluster list is generated, each cluster can be investigated further with the concord function, displaying concordance lines with a hyper link to the source text. In addition to the main cluster list, it is possible to toggle between word list screens by selecting the tabs at the bottom of the page. Contained in the *statistics* section of the word list are aggregate numbers of the corpus including total words, words used to construct the word list, average length of texts, and specific information for each text used.

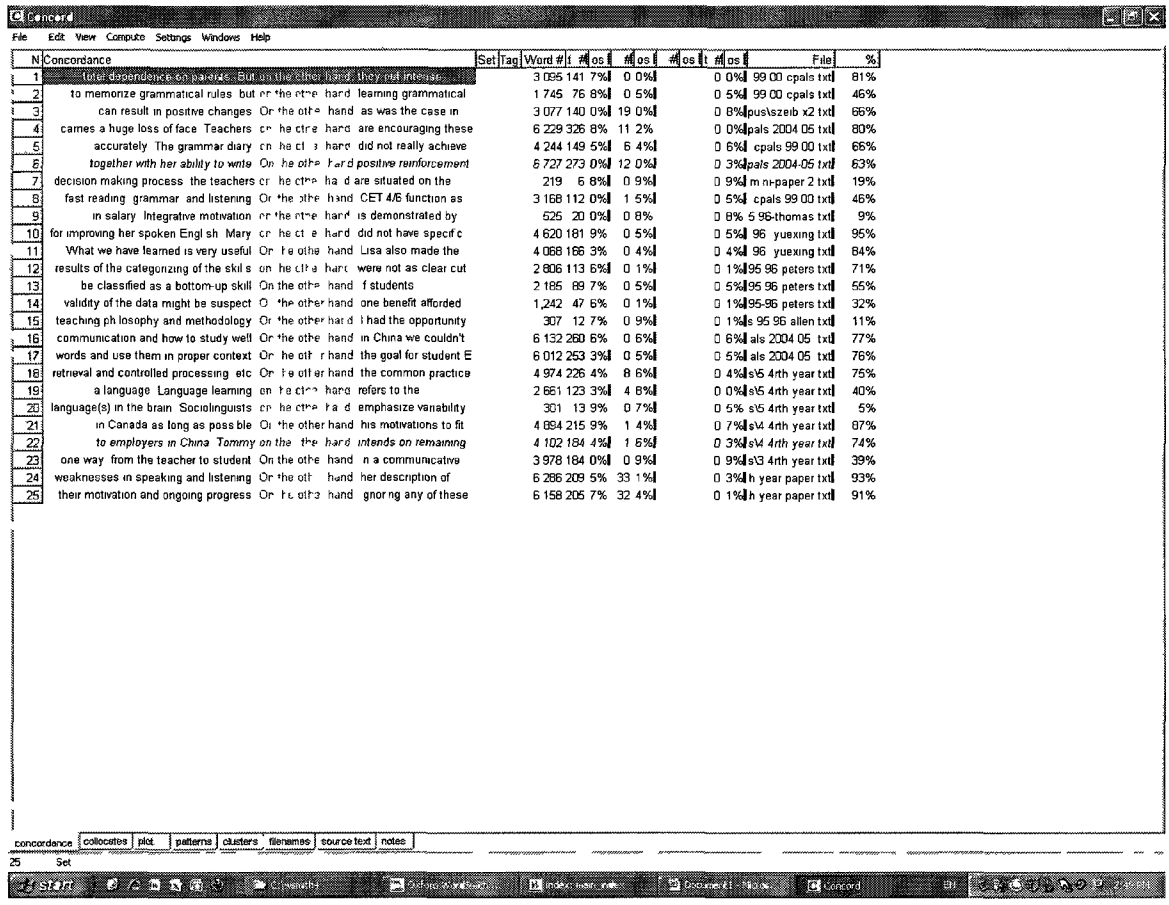

Figure 6.5. Concordance lines of the cluster *on the other hand*

Figure 6.5 shows the concordance lines for the cluster *on the other hand,* which can be displayed by selecting the desired cluster in the main index, and then going to compute: concordance. These concordance lines allow for contextualization of each bundle, each containing a hyper linked cluster that goes to the source text. At the far right of the concordance screen each source text is labeled, which allows for more careful checking of distribution parameters.

6.1.6 Functional taxonomy

Initially, bundles were organized based on Biber et al.'s (2004) taxonomy (shown in figure 3.1) that was used for the T2K SWALS corpus. Due to the speech registers, textbook writing and other text types found in the T2K SWALS corpus, the taxonomy

covered a much different set of functions than those encompassed in ALDS academic writing. For this reason, the taxonomy of Cortes' 2004 study, which classified bundles from History and Biology writing (much closer in mode and genre to the ALDS corpora) was also considered for bundle organization. Although Cortes (2004) used the same template as Biber (1999) for her taxonomy, she classified some bundles in a different manner. For example, the bundle *the fact that* was classified as a text organizing framing bundle, whereas Biber et al. (2004) classified the same bundle as a stance signifier. More importantly, Biber's taxonomy grouped the 'framing' function not under text organizing, but in the *reference* category. These inconsistencies demonstrate that the functional classification of bundles is a highly subjective process that must be done with care and attention to the particular characteristics of the corpus itself. In this analysis, a more liberal approach to the functional taxonomy was taken, one that incorporated elements of Biber et al.'s (2004), Cortes', and Hyland's (2008b) taxonomies. Because Hyland's (2008b) study was the closest in terms of text type to the ALDS corpora (he also looked at some applied linguistics Master's writing) his taxonomy was used more centrally. However, the labels chosen for each functional category and subcategory stayed as true as possible to Biber's (1999) initial labels; this was done in order to maximize the clarity of comparisons made between studies since most researchers use the labels Biber coined.

While it was not possible to find any single taxonomy that classified each bundle in exactly the same way, all of the classifications used in this study had been used previously in at least one of the taxonomies mentioned above (the full taxonomy for this study is shown in figure 3.2). The only new categories added to

the ALDS functional taxonomy were *personal reflection stance* bundles (only novice bundles) and *procedure/ability* reference bundles. The 'personal reflection' novice bundles (*who we are and, and how we relate, I was able to)* were not listed in any of the previous lexical bundle studies, and it was difficult to find an appropriate place for them within a preexisting category. 'Procedure' reference bundles were a category in the Hyland (2008b) study, but adding 'ability' to this sub-group was novel to the ALDS taxonomy. Combining 'ability' with 'procedure' seemed like the most natural fit for all the bundles concerned based on their uses in text. More explanation of bundle classification and specific concordance lines for ambiguous bundles will be discussed in the results chapter.

Because the classification process was somewhat subjective and presented a moderate learning curve for a novice corpus researcher, a second researcher evaluated all classifications presented in this study. The researcher who reviewed the ALDS functional taxonomy was also a student in the Carleton ALDS Masters program, familiar with Wordsmith Tools and lexical bundle classification.

6.2. Questionnaire and Interview methodology: Participants and materials

In the qualitative phase of the study, two novice ALDS writers, and two upper-level ALDS writers whose work was included in the corpora were asked to participate in the questionnaire and interview phase of the study. One native English speaker and one non-native English speaker from each level were included. Students from ALDS 1001 (novice writers) who were asked to participate (10 students were sent a request) were in the 'high' level of writing on their mini papers. Clara and Leo were chosen as pseudonyms for the novice participants, and upper-level participants

were called Helen and Victor. These students were selected because they would presumably have more meta knowledge about writing process and better articulation of their writing knowledge. Students whose mini papers fell into the low group were also contacted but no novice writers in the lower group agreed to participate.

The first novice student who completed the questionnaire and interview was a first year student who had returned to university after taking significant time off after high school to start a family and pursue a career. Leo spoke English as a first language and had only minimal instruction in other languages that took place once he arrived at Carleton. Leo was a very enthusiastic member of ALDS 1001, and his mini papers were strong. He managed to improve from an 8.5 out of 10 grade to a 10 out of 10 grade between the two papers, and used both TAs and professor extensively for help. This dedication to improving his writing and the high-level interpretation skills he demonstrated in the ALDS 1001 class made Leo an ideal candidate for the questionnaire and interview phase. The second novice participant, Clara, was also taking ALDS 1001 for the first time, but unlike Leo, she was in her 4th year of study in a Health Sciences program at the University of Ottawa. Clara spoke English and Arabic as first languages, but grew up in a house where almost no English was spoken. Like Leo, Clara showed large improvement from the first mini paper to the second, going from 6.5 out of 10 to 9.5 out of 10. She was also a keen student who regularly attended class, though she did not seek help with either professor or TAs during the course. The two Masters students who participated in the study were Victor and Helen. Helen was in the process of finishing her first year of the ALDS MA program, and had come from a mixed academic background of food

sciences and French. She was a native English speaker and her writing from the Masters level curriculum development course was included in the upper-level corpus. Victor was in the second year of the MA program, and was an international student. Speaking Spanish as a first language, Victor also spoke English and French at a fluent level. His writing from curriculum development was also included in the upper-level corpus.

6.2.1 Research instruments and procedure

The questionnaire was developed from questions and research approaches by Beaufort (2004), Haas (1994) and Sommers and Saltz (2004), who investigated different aspects of academic writing development (two by case study and one by longitudinal study of freshman year writing). The questions stemming from this past research focused on how writers transition between levels of education, how they conceive of the genres they are writing in, what aspects of education have helped them improve as writers, whether they are adequately prepared as ALDS writers, and how important writing is as a means of testing knowledge and evaluating students. In addition to these questions, there were several others that addressed the importance of linguistic command in ALDS writing, and what role language instruction might play in preparing ALDS writers (both at the undergraduate and graduate level). The questionnaire was piloted on Masters level ALDS students before being finalized. As a result of the piloting process, it became apparent the questionnaire alone was not going to be a sufficient source of data gathering, especially since the study already contained a large quantitative element. The questionnaire was thus recalibrated as a priming tool for interviews, which

followed immediately afterwards, allowing students to expand on their answers, and providing the researcher with a touch point for directing the interview.

Students from novice and upper-level ALDS courses were briefed about the questionnaire/interview in an email outlining the nature of the study with a consent form attached. Those who agreed to participate met with the researcher on Carleton campus outside of class time and completed questionnaire and interview individually. The questionnaire consisted of 13 questions (Appendix B.2) and took approximately 15 minutes to complete. Participants had chances to ask the researcher any clarifying questions while they completed the document.

Immediately following the completion of the questionnaire, participants were interviewed for between 22 and 30 minutes. With the consent of the participants, interviews were recorded with a portable microphone. Interviews took place (between March 10th and April 1st, 2011) near the end of the study because it was important to have most of the quantitative analysis done before crafting the questionnaire and interview. Interviews were conducted in a semi-structured manner, following a question script (Appendix C.3). The questions were intended to elicit further elaboration and reflection on several points from the preceding questionnaire including information about the student's language background, reflection about their transition between levels of academic writing, how they perceived their own writing process, what parts of ALDS writing they found most challenging and what types of support would be most helpful in achieving more proficiency as ALDS writers. Before the interview, each participant was briefed about the purpose of the study and what role their interview and questionnaire would play in the results and outcome. Any questions or concerns about the study

were addressed at this time. Only one of the novice writers had any questions before the interview took place, asking for elaboration on the meaning of 'idioms' and 'writing conventions' (jargon that was present in the questionnaire). Because the upper-level writers had some experience with ALDS conceptions of writing process and genre theory, they did not need much background information or context in interpreting the questionnaire or interview questions.

6.2.2 Analysis

Questionnaire and interview data were analyzed in a two-step process. First, Likert scale questions from all four questionnaires were entered into a spreadsheet, and compared to find any trends or patterns. These questions were analyzed with novice and upper level students grouped together, and then with L1 and L2 speakers together. Salient results from the Likert scale questions were used to better interpret and organize the recorded interview responses.

Interview recordings were then partially transcribed. Relevant sections of the recordings were transcribed and labeled with their corresponding question from the interview, or the associated trend highlighted from questionnaire responses. After transcription was completed, the results were analyzed again for salient patterns, both in regard to the questionnaire responses, and to findings from past research regarding writing development and novice/upper-level writing approaches.

Chapter 7. Results and Discussion

The results and discussion of findings from both corpus (quantitative) and questionnaire/interview (qualitative) phases will be presented in three sections. This first section presents results from aggregate data and discourse functions of lexical bundles in the upper-level and novice ALDS corpora respectively. This general picture of bundle use in both corpora will be followed in section 7.2 by a comparison of aggregate statistics and discourse functions of lexical bundles from the two corpora, as well as some examination of the intermediate ALDS corpus (ALDSI) as a control measure. In 7.3, the results from the qualitative portion of this study are discussed.

7.1 Lexical Bundles in the corpus of upper-level ALDS writing

To begin the investigation of lexical bundles in the upper-level ALDS corpus, general results from frequency, distribution, and aggregate bundle data will be presented first, followed by more specific analysis of the main discourse functions of 'stance', 'reference', and 'text organizing' bundles.

From a search of the 35 texts of upper-level writing, totaling 197,382 words, there were 102 four-word sequences that occurred at least eight times in the corpus. Of these 102 bundles, 17 occurred in just one or two texts, 11 bundles occurred in three or four texts, and 73 bundles occurred in five papers or more.

Table 7.1. Functional distribution of Lexical Bundles from the ALDS-UL corpus

	Bundles	Total and % of all bundles
Text Organizers		19 (29%)
Contrast/Comparison	*on the other hand*	
	in this section of	
	in this paper I	
Structuring signals	*I would like to*	
Framing	*in the form of*	

	in the case of	
	the extent to which	
	the ways in which	
	the way in which	
	in terms of the	
	the degree to which	
	with regard to the	
	on the part of	
	in accordance with the	
	in relation to the	
	as well as the	
	in order to get	
	a better understanding of	
	as part of the	
Referential		37 (57%)
	a great deal of	
	the majority of the	
	a wide range of	
	a high level of	
Quantity/qualification	*there are a number*	
	to be able to	
	the role of the	
	the analysis of the	
	the development of the	
Procedure/ability	*to the development of*	
Time/location	*at the beginning of*	
	in the field of	
	the meaning of the	
Description	*the content of the*	
	at the same time	
	the end of the	
	the beginning of the	
	one of the most	
Multi-purpose	*at the end of*	
	in their home countries	
	outside of the classroom	
	program at Carleton university	
	in the classroom and	
	of second language acquisition	
	of the target language	
	in the target language	
	the teacher and the	
	English as a second language	
	language teaching and learning	
	of teaching and learning	
Topic-bound/related to the	*as a second language*	
field of research	*of this case study*	

	to learn the language	
	of the English language	
	the role of the teacher	
	the culture of teaching	
	the process of learning	
Stance		9 (14%)
	to the fact that	
	due to the fact	
	the fact that the	
	it is clear that	
	the importance of the	
	it is possible that	
Epistemic/probable possible	*it is important to note*	
	it is important to	
Obligation/directive	*the best way to*	
Total		65

With this initial list of 73 bundles that qualified for frequency and distribution criteria, a check was done for each bundle to see that the distribution covered at least two of the academic contexts[17], and that the bundle used was not part of a direct quotation. After this second round of checking there were 65 bundles remaining. These 65 bundles (table 7.1) were divided into three main functional categories and 11 sub-categories. The taxonomy, taken primarily from Cortes (2004) Biber (1999), Biber et al. (2004), and Hyland (2008b) had considerably fewer categories than Biber et al.'s (2004) frequently cited taxonomy that considered conversation, textbooks, classroom speech, and academic prose. Not surprisingly, having only one mode of text (academic prose) from a single academic discipline (ALDS) greatly reduced the discourse function diversity and range of the bundles found. While the number of different bundles found was not surprising,

[17] There were four academic contexts for experienced papers: fourth year papers from a Methodology in Language teaching course (A- and higher), CPALS papers, mini-papers from a graduate curriculum course, and course papers from current ALDS graduate students from Language Testing and Critical Discourse Analysis.

their distribution across the taxonomy was noteworthy. Like Cortes' (2004) investigation of history and biology writing and Hyland's (2008b) look at bundles in academic registers, there were a large percentage of bundles occurring in the reference and text organizing categories. There were 19 text organizing bundles, grouped into 3 subcategories, which made up 29% of the overall bundles, 37 reference bundles in 6 categories (57% of the total), and 9 stance bundles in two subcategories, making up 14% of the total. The highest occurring bundles in the corpus occurred 25 times (45% of the total texts[18]) which normalizes to a frequency of 125 words per million.

Compared to the landmark study by Biber et al. (2004), the frequency and distribution of bundles in the upper-level writing seemed high at a first glance. In Biber et al.'s (2004) study using a subset of the Longman Spoken and Written English (LGSWE) corpus that contained 5.3 million words from research articles and academic books, they found only 19 lexical bundles that occurred at least 40 times per million over at least five different texts. It seems surprising that they would find only 19 lexical bundles compared to the 65 found in the ALDS-UL, considering their corpus was more than 26 times as large. Upon closer examination and considering several important factors. However, this stark difference can be accounted for.

It has been well established (Cortes, 2004, 2008; Hyland, 2008, 2008b; Biber et al. 2004) that as writers gain experience within a discipline and write papers at

---

[18] It should be noted that in cases where there was more than one text (mini-paper or course paper) from the same author, those papers were grouped into the same text file. This was done to ensure that idiosyncratic patterns in a single person's writing would not skew distribution statistics.

higher levels of academia, their use of frequently repeated sequences drops accordingly. Because the academic prose considered for the LGSWE corpus was all at a higher experience/academic level than the text from upper-level ALDS writers, one would expect there to be a drop off in the number of bundles found. In addition, the text from LGSWE came from a variety of academic registers and disciplines. Because I only considered ALDS writing from a single university program across a fairly restricted set of subject matter (over 60% of upper-level texts came from six different ALDS courses), it is not surprising that there were considerably more bundles found in my study, as writers tend to repeat more phrases when they are writing on similar topics, for the same type of assignment.

In comparison to other research with specialized corpora of academic writing, the number of bundles found in this ALDS M.A. work appears more normal. In Hyland's (2008b) study where he considered academic texts including research articles, Ph.D. dissertations and M.A. theses in four disciplines, search parameters of 20 times per million over 10 percent of texts yielded 130 lexical bundles. These findings are much more consistent with the total bundle counts in the current study, even though 130 bundles is still a much smaller proportion of a 3.5 million word corpus than 65 bundles out of a 197, 387 word corpus[19]. Chen and Baker (2010) who looked at a one million word corpus of academic prose found 108 lexical bundles using a three-text distribution and frequency of 25 times per million. While

[19] As corpora get smaller, the instances of lexical bundles do not continue to decline in a proportional fashion. Therefore a corpus that is 17 times larger than the ALDS-UL, such as Hyland's (2008b), would not contain 17 times more bundles, nor would the ALDS-UL be expected to contain 17 times fewer bundles.

their search parameters were more inclusive, these results again are much closer to the vicinity of the ALDS-UL bundle results.

In terms of distribution across functional categories, the ALDS-UL corpus is also relatively consistent with past research. Of the 19 bundles identified by Biber et al. from academic prose, 15 (78%) of these were referential, three (2%) were stance, and one (.06%) was a discourse organizer. In Hyland's (2008b) study he found for research articles, 25 (36%) bundles were 'research oriented' (this category roughly corresponds to 'reference' bundles), 37 (52%) were 'text oriented' (corresponds to text/discourse organizing), and 8 (11%) were 'participant' oriented (stance bundles). However, in the lowest level of writing (M.A. theses) Hyland found that 'research oriented' bundles were most prevalent, followed closely by 'text oriented' bundles, with 'participant bundles' in a distant third. This is consistent with the Masters and fourth year work examined in this research, even though no thesis or research essay material was considered[20].

The aggregate results from the ALDS-UL corpus will be considered again when they are compared with results from the ALDSN and ALDSI corpora, but before a full comparison can be made more specific functional analysis of the ALDS-UL is dealt with next.

7.1.2 Text-organizing lexical bundles in the ALDS-UL corpus

The text-organizing bundles in both the ALDS-UL and ALDSN corpora were the most difficult to classify because of the considerable variation between the taxonomies

---

[20] While several theses and research essays were initially considered for the ALDS-UL corpus, they were left out based on their length in comparison to all the other texts, and the significant differences between thesis writing and all other types of work at the graduate or undergraduate level.

and classifying criteria of different researchers. For the 19 bundles classified as text-organizing in the ALDS-UL, they had to be serving a function that related to the organization, framing, and connection of text elements to what preceded and followed in the text. These bundles fell into three subcategories: contrast/comparison signals, structuring signals, and framing signals. The majority of the bundles fell into the framing subcategory; these bundles provide some sort of limiting condition to an idea or argument so it can be situated within the text (Hyland, 2008, p. 49). Several concordance lines are presented to illustrate the framing function for the bundles *in the case of*, *in order to get,* and *the extent to which.*

1. However this does not happen overnight, it is a process, and *in the case of* beliefs they are very difficult to change.
2. *In order to get* as much relevant information as possible from the students special care was taken not to lead the students with the questions.
3. *The extent to which* a learner believed that she could influence outcomes came to be regarded as a critical factor in the degree of motivation she would experience.

Although the bundles in the framing category all seemed to perform this text organizing function, Biber et al. (2004) and Chen and Baker (2010) do not classify 'framing' as a text organizing function, but as a referential one. While it is true that many of these framing bundles "generally identify an entity or single out some particular attribute of an entity as especially important" (Biber et al, 2004, p. 393) the bundles simultaneously serve to situate those entities and their attributes within the text. While it is possible that Biber et al.'s concordance lines were very different for these framing bundles, it is more likely that this distinction is one of

perspective or subjective opinion rather than objective criteria applied to the taxonomy. It also seemed more prudent to classify these framing bundles as text organizers since the research by Cortes (2004, 2008) and Hyland (2008, 2008b) labeled them as such. Cortes and Hyland examined modes and registers of text (primarily expert academic writing) that were closer approximations of the academic texts in the ALDS-UL corpus.

After framing bundles in the text-organizing category, structuring signals, with three bundles, comprise the next largest subgroup. The other two groups, contrast/comparison signals, contained one bundle. The most frequent bundle in the ALDS-UL corpus, *on the other hand,* was the only contrast/comparison signal, occurring 25 times over 16 texts (45% of the total texts). This bundle was classified in the contrast/comparison group since it was used to compare information, or contrast differing positions on a subject. The last text organizing subcategory of structure signaling bundles contained three sequences that were found over seven and six texts respectively with a maximum of ten occurrences. These bundles served to structure discourse of a text, meaning they directly refer to parts of the writing itself, as opposed to internal concepts or subject matter. For example, the bundle *in this section of* is used to refer to a specific section of the essay:

1) In this section, I selected vocabulary with caution when using words such as improve, empower, or grow to avoid any unintended implication or absence of these abilities in advance.

In the concordance line above, *in this section* informs the reader about a part of the text itself, letting them know what will be contained in the following paragraph.

Overall, these text-organizing bundles make up the second largest proportion of the total bundles in ALDS-UL and account for many of the high frequency bundles with wide text distribution (see table 7.2).

Table 7.2. Average frequencies of text organizing bundle subcategories in the ALDS-UL

	Average Frequency	Average Distribution	Number of bundles in the category
Contrast/comparison	25	16	1
Structural signals	8.7	6.3	3
Framing	11.7	8.5	15
Total	11.9	8.6	19

7.1.3 Referential Bundles in the ALDS-UL corpus

As mentioned earlier, the referential bundles in ALDS-UL made up the largest proportion of bundles by function, (57%) something typical of non-expert academic writing (Biber et al., 2004). These bundles were divided into six subcategories, with most of the bundles occurring in categories of quantification (five bundles), procedure/ability (five bundles), multi-purpose (five bundles) and topic-bound (18 bundles). As can be seen in Table 7.3, the average frequency of bundles was highest in the multi-purpose bundle category, followed by procedure/ability and then description. With just under half of the total bundles in the referential section made up of topic-bound bundles, clearly there are some core phrases concerning language, teaching, and second language that upper-level writers cannot avoid using. The two categories of reference bundles that posed the most difficulty were procedure/ability and description bundles. In the procedure/ability category the bundles *the role of the* and *the description of the* were included even though they

also served direct description function like the bundles in the 'description' category, and a text organizing function similar to framing bundles such as *a wide range of*. The choice to put these bundles in the procedure ability group was based on their classification as such in Hyland's (2008b) work, and the examination of concordance lines.

1) There will undoubtedly be differences, due to, for instance, *the role of the* policymakers, and *the role of the* syllabi.

2) Thus, *the development of the* English language program for this huge group of students has significant influence on language teaching practice.

In line 1, the bundle is being used to describe the policymakers and syllabi in terms of their use or function; this bundle is termed 'procedural' because it allows the writer to "structure their activities and experiences of the real world" (Hyland, 2008b, p. 49) in reference to how something works or is affected by some sort of process. In concordance line 2, the bundle *the development of the* is procedural in its reference to the noun phrase *the English language program for this huge group of students,* but it also serves a framing function in the text by referencing the noun phrase, and framing it with information about its *impact on teaching practice.* This framing function was deemed secondary to the function of procedure reference, therefore this bundle (and others in the group) were not placed in the text-organizing or multi-purpose categories.

Table 7.3. Average frequencies of referential bundle subcategories in the ALDS-UL

	Average Frequency	Average Distribution	Number of bundles in the category
Quantification	10.2	7.5	5
Procedure/ability	15	9	5
Time/location	16	8	1
Description	11.5	6.8	3
Multi-purpose	17.6	12.2	5
Topic-bound	10	6.7	18
Total	13.4	8.4	37

7.1.4 Stance Bundles in the ALDS-UL corpus

By far the smallest functional group of bundles in the ALDS-UL corpus, the nine stance bundles make up only 14% of all bundles. The small proportion of stance bundles makes intuitive sense considering the emphasis on more 'formal language' in academic registers, and the extensive amount of hedging that takes place in most disciplines (O'Keefe et al., 2007). There are also corroborative findings of stance bundles in academic writing from many past studies; O'Keefe et al. (2007) Cortes (2004, 2008), Chen and Baker (2010), Biber (2006), who all found that academic writing contained far fewer stance bundles than other modes of discourse. The findings from ALDS-UL also show very little variety in stance functions, with only two subcategories being required to classify the seven bundles: epistemic/probably-possible, and obligation-directive features. The seven bundles in the epistemic/probably-possible category function in a way that allows the writer to

take an epistemic position on an element in the text (1) or to express ideas with a degree of skepticism or uncertainty (2).

1. *It is possible that* Lisa's goal was not 'hard' and 'specific' enough to function as an efficient motivator for raising her English to a higher level.

2. This is perhaps in part *due to the fact* that specific learning goals in China were usually defined and developed by the teacher.

The remaining two stance bundles, *it is important to* and *the best way to* function in a way that directly speaks to the reader of the text and instructs them in some way. This engagement with the reader must be made directly for a bundle to assume this function, as demonstrated in the concordance line below:

1. However, this interpretation remains only speculative, and *it is important to* consider the results as they stand.

Although there are relatively few bundles with a stance/participant oriented function, the few that in this group occur quite frequently. As seen in table 7.4, the average frequency of bundles with a stance function is 13.1, only .3 less than for reference bundles which occurred at 13.4 frequency on average.

Table 7.4. Average frequencies of text organizing bundle subcategories in the ALDS-UL corpus

	Average Frequency	Average Distribution	Number of bundles in the category
Topic elaboration/probable - possible	12	7.6	7
Engagement	15	9.5	2
Total	13.1	8.7	9

In Biber et al.'s (2004) work they have divided stance bundles into four subcategories each divided again (eight total) into 'personal' and 'impersonal'. The much more limited stance taxonomy in the current study is not a reflection of

subjective opinion in classification, but rather a different selection of modes and registers. Biber et al. examined several speech genres for their study, which greatly impacted the total number of bundles found, as well as their functions in the text.

7.1.5 Aggregate results from the beginner and intermediate ALDS corpora (ADLSN and ALDSI)

Moving on to the analysis of the ALDSN corpus, the aggregate data on bundle frequency, distribution and total numbers is presented, followed by closer examination of bundle functions looking at text organizing, reference, and stance bundles respectively. Unlike the discussion of the ALDS-UL corpus, there will not be the same comparison of aggregate bundle data to past studies, mainly because there are very few studies doing corpus comparisons with data bases similar to the ALDSN.

In order to keep consistent search parameters with the ALDS-UL corpus, a different frequency minimum was used to search the ALDSN. The ALDSN consisted of 112, 244 words, so a frequency of 40 times per million normalizes to five times. The raw frequency is 4.49, but this number was rounded up since a frequency count of five is already a very low threshold upon which to consider lexical bundles; a lower number would not have met the five text distribution criteria. The distribution is kept consistent with the cutoff for the ALDS-UL corpus (as discussed in the methodology) because although the texts considered were five times shorter than those of the upper-level ALDS writers, there were nearly double (65) the number of texts from novice ALDS writers. Using the aforementioned frequency (five times) a total of 169 bundles were found. With a distribution of at least five

texts, the number of bundles was reduced to 110. At this point, the concordance lines for each bundle were checked to see if the bundle was a part of a direct quotation, or was used repeatedly by the same author over both mini-papers[21]. After doing additional checks for 'authenticity' and distribution, the number of bundles was drastically reduced from 110 to 59.

Considering the small size the ALDSN reference corpus, 59 bundles is quite a high count, especially considering that 40 times per million is a very conservative cut-off limit for bundles. Many studies of academic writing registers use cut-offs of 20-30 times per million (Cortes, 2004; Chen and Baker, 2010; Hyland, 2008b), but a conservative limit seemed more appropriate in this case so as to increase the likelihood that bundles would be meaningful and potentially formulaic.

In terms of functional distribution, the novice ALDS bundles were used as reference markers in 69% of cases, text organizers in 20% of cases, and stance markers 10% of the time (see table 7.5 below for complete list of bundles organized by function).

Table 7.5. Functional distribution of lexical bundles from the novice ALDSN corpus

Function	Lexical bundle	Total and % of all bundles
Text Organizers		12 (20%)
Contrast/comparison	*On the other hand*	
	In this mini paper	
	In this paper I	
	Paper, I will be	
Structuring	*This paper I will*	
Resultative	*As a result of*	

[21] The ALDSN corpus, unlike the curriculum papers from the ALDS-UL corpus, did not have each author's mini-paper combined into the same text file. Instead, each author was given two case numbers, #a and #b for papers one and two. If a bundle occurred in both #a and #b papers, this was only counted as single text.

	That there is a	
	In the case of	
	Not be able to	
	In the form of	
	When it comes to	
Framing	*As well as the*	
Reference		41 (69%)
	The rest of the	
	There are many different	
Quantity/qualification	*One of the most*	
	To be able to	
	They are able to	
	The role of the	
	For the purpose of	
	In the use of	
	The use of the	
	Will be able to	
	The ability to learn	
	Was not able to	
Procedure/ability	*Was able to learn*	
	On a daily basis	
	By the age of	
	At the age of	
Time/location	*At a young age*	
Description	*His or her own*	
	Zone of proximal development	
	The grammar translation method	
	As a second language	
	Learning a new language	
	In the course pack	
	Through language we express	
	Applied linguistics and discourse	
	Speak the same language	
	Theory of language acquisition	
	Acquire and use language	
	French as a second language	
	Language acquisition is not	
	Learn a second language	
	To learn a language	
	In acquisition of language	
	In other words language	
	A second language in	
	And cultural contexts and	
	Are able to communicate	
	Be able to communicate	
	Communicate with each other	
Topic-bound	*To learn a second*	

Multi-purpose	*At the same time*	
Stance		6 (10%)
	Due to the fact that	
Epistemic/probably-possible	*To the fact that*	
	The fact that they	
	Who are we and	
	And how we relate	
Personal reflection	*I was able to*	
TOTAL BUNDLES		59

The relatively high number of reference bundles is not overly remarkable, especially considering the trends found in other studies that less experienced writing tend towards greater use of reference and text organizing functions. However, none of the past studies specifically examined first year ALDS academic writing, or any academic writing that approximated the length, purpose[22], and topic variety included in the ALDS 1001 mini-papers. One thing that stood out very strongly from the functional breakdown of bundles was the high percentage (38%) of topic-bound bundles. In no other study of academic writing that I reviewed did anyone find such a high count of topic-bound bundles. The potential significance of topic-bound bundles, and what role they might play in the development of ALDS writers will be discussed further in section 2 and 3 when the ALDS-UL and ALDSN corpora are compared and analyzed by grade level (ALDSN only). Next will be a closer look at how lexical bundles were used within each of the function categories and subcategories.

---

[22] Purpose refers to the specific conditions and criteria in place for the 1001 mini-papers. These papers were not written with the purpose of summarizing and analyzing large amounts of course/ALDS material as the graduate and fourth year papers were. They had much less rigorous criteria concerning topic selection, content, and depth of analysis

7.1.6 Text organizing bundles in the ALDSN corpus

In the ALDSN corpus, 12 bundles were classified as having a text organizing function, divided into four sub-categories: Contrast/comparison, structuring, resultative, and framing signals. Framing bundles make up the highest proportion of these with seven, while the others contain one, two, and three bundles (see table 7.6). In the contrast comparison subcategory, the bundle *on the other hand* is once again the only member. This bundle occurred 20 times across 16 texts; only one other bundle in the novice corpus occurred 20 times (*zone of proximal development),* but that bundle was only distributed across six texts. The structuring signals subcategory contains three bundles in the ALDSN, each one referring to the mini paper itself: *in this mini paper, in this paper I,* and *paper, I will be.* These bundles were tied with framing for the lowest average frequency of the text organizing subcategories (see table 7.6), although considering how closely they resemble each other the *in this paper* sequence is one that many (30 total occurrences) novice ALDS writers employ[23]. These bundles were almost all used in the introductions, and served to establish the paper's topic and provide some sort of road map for the structure and format. The following concordance lines illustrate this function:

1. *In this paper, I will* be discussing how L2 learners learn a new language, how classroom feedback or the so called "recasts" can cause one to fail, and how age plays a significant role in learning a second language.

2. *In this paper, I* look at three cases of feral children and question the validity of the innateness hypothesis and critical period in respect to them, and identify alternate factors that could have affected their ability to acquire language.

---

[23] The sample sizes of 1, 2, 3, and 7 occurrences for each of these categories are so low that it is not worth reading much into the average frequencies. They are presented in this section mainly to provide transparent reporting of results, not to infer conclusions about how important each subcategory might be.

From examining the 30 concordance lines for the three structuring bundles only two occurrences were found outside of introductions. The final two bundles found in conclusions were used to summarize the purpose of the paper.

The ALDSN also had two bundles which were used to mark results or outcome from another element in the text. Resultative bundles were not found at the 40 times per million threshold in the upper-level corpus, with the bundle *as a result of* occurring 7 times in the ALDS-UL, just one under the 8 frequency cut-off. *As a result of* occurred 18 times in the ALDSN over nine different texts, ranking as the third most frequent bundle used in the corpus. The resultative bundle *that there is a* also seemed to be used in textual framing ways, but overall had a more resultative function. Some examples of can be seen in concordance line 1, line 2 showing this resultative function:

1. In this mini-paper, I will be proving *that there is a* distinction between men and women's use of language by demonstrating a few studies made by some linguists and psychologists.

2. The benefits of acquiring a language with the right hemisphere of the brain is *that there is a* higher chance of risk taking and not being afraid of making mistakes within the language.

3. Some of believe *that there is a* higher source out their, which each and everyone a power, it may have come in different ways, but surely some of posses it, it is Language.

As can be seen in lines 2 and 3 above, there were some quite disjointed and weak sounding constructions using *that there is a*; in line 3, the mistakes were significant enough to heavily disrupt the meaning and grammaticality of the sentence. This misuse of the bundle in lines 2 and 3 could also suggest that framing textual elements and expressing results of processes is a difficult task for some novice

writers In concordance line 1, the writer also uses the phrase *I will be proving,* preceding *that there is a,* which was directly discouraged by the professor and TAs in the course. Combining this bundle with another phrase that was expressly discouraged indicates that instruction from the professor in class and guidelines on the assignment handout were not transparent or helpful to the student. Of the nine writers that used this bundle, most of them fell either into the very low level of writing (three of them) or into the medium high or high levels (two high, five medium high). It is possible that using "specifying bundles [which] identify abstract characteristics" between textual elements (Biber et al. 2004, p. 395) is an area that causes problems for weaker ALDS novices, and rewards those who do it proficiently. Further discussion of how weaker and stronger novice ALDS writers used bundles will be addresses in section 6.2.

Table 7.6. Average frequencies of text organizing bundle subcategories in the ALDSN corpus

Functional category	Average Frequency	Average Distribution	Number of bundles in the category
Contrast/comparison	20	16	1
Structuring	10	8	4
Resultative	18	9	2
Framing	9.8	8	5
Total	11.4	8.8	12

7.1.7 Reference bundles in the ALDSN corpus

Reference bundles were by far the most prevalent function of lexical bundles in the novice corpus, accounting for 69% of all bundles. Split into the same six subcategories as the ALDS-UL taxonomy, there were some notable differences in the

distribution of bundles across those categories. Firstly, there were a large number of bundles referring to 'ability' attributes of elements in the text. Of the 10 procedural/ability subcategory six of those bundles referred to some sort of ability, either in the affirmative or negative; bundles like *to be able to, they are able to,* and *was not able to* are somewhat flexible in the subjects or agents they refer to, while the bundles *the ability to learn* and *was able to learn* refer specifically to learning both first and second languages. Surprisingly, there was only one instance of an 'ability' reference bundle in the ALDS-UL corpus. The pairing of procedure and ability functions into one subcategory was a subjective decision, based on the common concept of how or with what means something is happening or working. Concordance lines 1, 2, and 3 give an example of ability and procedural function.

1. Signs of this male-conditioned society can easily be seen through *the use of the* English language

2. Although I was able to use French correctly in an academic and largely written context where I had a lot of support and tools to aid me, I *was not able to use* it as effectively in applied settings outside of the classroom.

3. Genie was not able to speak

In example 2 *was not able to* serves a textual organizing function, with the frame *although I was able to....I was not able to* allowing the writer to express connections between two elements of the text. This framing pattern was present in two of the five occurrences of the bundle *was not able to*, therefore it was classified along with the majority of its uses which were referential. The classification of ability reference bundles is a good example of how difficult and subjective the process of functional classification can become, especially when there is no relevant precedent from past research.

The category of topic-bound bundles stood out the most in the ALDSN corpus, in that there was no group in either the ALDSN or ALDS-UL that came close to its numbers, although the overall frequency was quite low (see table 7.7). All the way through from very high frequency bundles to the lowest frequency sequences, novice ALDS writers used a great number of topic-bound bundles referring to concepts, theories, sources, and subjects common to the course and essay topic. Altogether, the bundles with a topic-bound function accounted for 38% of all classified bundles (see table 7.7 for a breakdown of frequency and distribution in the reference function subcategories). Many of these bundles were used just enough times (five) to make the frequency cut-off, and many more were used extensively but only over one or two texts, violating the distribution criteria.

This indicates that novice writers are using a great deal of repetition in their writing, especially in referring to the topics they are exploring. Perhaps the newness of these topics for the writers, and their lack of reading on related subjects, has not given them the linguistic tools to effectively control topic-centered language and provide much sentence variety. It is possible that such topic-bound sequences are learned and memorized in a very formulaic way at the start of one's initiation into the writing for a specific discipline. As writers gain experience, perhaps they gain better understanding of how fixed or flexible these sequences are, and they either gain a greater variety of prefabricated strings to reference the same concepts, or they become better equipped at using their 'open-choice' (Sinclair, 1991) language device to generate less repetitive sounding prose.

Table 7.7. Average frequencies of referential bundle subcategories in the ALDSN

	Average Frequency	Average Distribution	Number of bundles in the category
Quantification	7.7	5.7	3
Procedure/ability	8.9	6.6	10
Time/location	7.7	6.7	4
Description	7	6	1
Multi-purpose	6	5	1
Topic-bound	8	5.7	22
Total	8	5.9	41

7.1.8 Stance bundles in the ALDSN corpus

Similar to the ALDS-UL corpus, stance bundles again make up a relatively small proportion of bundles in the novice ALDS mini-papers (10% overall). They are divided again into two categories: topic elaboration/probably-possible and personal analysis bundles. These bundles function to express some sort of epistemic stance of the author in relation to textual elements. This topic elaboration/probably-possible function is consistently identified in almost all research on lexical bundles, and as with the ALDS-UL corpus, classifying bundles with this function was relatively straightforward. Unlike the more experienced ALDS-UL writing, there were no instances of probable-possible functions only topic elaborating expressions, which are shown in the concordance lines below:

1. Personally e-mail has served me a great purpose *due to the fact that* I can keep in contact with my friends, while juggling work and school.

2. I have attended the school here in Ontario, through my years
there *I was able to* analyze and learn how to cope with my vision loss.

Finally, there were two bundles, which functioned as 'personal analysis/engagement' features, as Hyland (2008b) called them. Hyland (2008b)

based his taxonomy on concepts from Michael Halliday's (1994) work, and as a result does not find as much common ground with functional taxonomies of Biber, Cortes, Chen and Baker and others using the classic Biber (1999), Biber et al. (2004) taxonomy. Hyland's (2008b) designation of 'engagement features' was chosen to categorize the bundles *who we are and* and *and how we relate* because there was very little else in the past research that seemed relatable to their uses. The label of 'personal analysis' was created for this taxonomy because it seemed to fit best with how bundles were being used:

1. Our identities are *who we are and* without an identity we are lost, our identity places us in social categories in which we belong to.

The concordance line above shows how the bundle functions in a way that allows the writer to engage directly with their reader, speaking to them and implicating them in the prose, while also offering the writer's personal feelings about the topic. While Hyland classified bundles like *it should be noted that* and *it is clear that* as engagement features and did not specifically refer to the bundles being considered here, it seems as though this category is still the best fit considering the 12 concordance lines of evidence.

Table 7.8. Average frequencies of stance bundle subcategories in ALDSN

	Average Frequency	Average Distribution	Number of bundles in the category
Topic elaboration/probable – possible	12.2	7.3	4
Personal reflection	6	5.5	2
Total	10.2	6.7	6

7.2 Comparison of ALDSN and ALDS-UL aggregate data

The comparison of the ALDN and ALDS-UL corpora will be conducted in three sections: aggregate data comparison, comparison of top bundles, and levels of novice writing. These sections will be followed by discussion of how the results presented so far address the research questions concerning patterning of lexical bundles between the novice and upper-level ALDS students, and how lexical bundles might be used to characterize the development of ALDS writers.

To begin, the overall bundle counts will be looked at from each corpus to see which experience group uses more bundles. In the ALDS-UL corpus, writers used 65 bundles as compared with 59 in the ALDSN corpus. Although the ALDSN has just slightly fewer bundles using the same frequency and distribution parameters (40 times per million or more over at least five texts), it contains a higher proportion of bundles considering the corpus is only 57% the size of the ALDS-UL. In order to get a better gauge on how the total bundle counts compared with each other, two equivalent cross-sections of the ALDS-UL corpus, containing between 112,000-115,000 words, were searched for bundles (see results in table 7.9). These cross-sections were randomly sampled from texts in the upper-level corpus, with an effort to balance the amount of writing from each academic category (CPALS, course paper, mini paper, 4th year paper). The first upper-level writing sample, containing 114,097 words, produced 137 four word bundles total with a frequency count of five (40 times per million). Because there were only 19 texts in the cross section, doing a full break down of the distribution would not be meaningful.[24] The other

[24] There were 78 bundles that occurred over three texts or more, 50 over four texts or more, and 21 over five texts or more.

sample of upper-level writing (containing five of the same papers as the first sample), which contained 113,721 words, yielded 131 four-word bundles when searched with the same parameters as the first sample. Compared to the 169 bundles found in the ALDSN corpus (not considering any distribution criteria), the sample texts still had 19% and 23% fewer bundles respectively. Considering these sample texts, and the large proportional difference in bundles between the ALDSN and ALDS-UL corpora, it appears that novice ALDS writers use considerably more lexical bundles in their academic writing.

Table 7.9. Cross-section comparison of upper-level and novice bundles

ALDS-UL sample 1 (114, 097 words)	ALDS-UL sample 2 (113, 721 words)	ALDSN (112, 244 words)
137 four word bundles	131 four word bundles	169 four word bundles

The decreasing use of lexical bundles in higher-level academic genres (PHD dissertations, research articles, books) has been confirmed by several studies, but none of which looked at first year to Masters level writing in a single discipline. Hyland (2008b), Cortes (2004, 2008), Biber et al. (2004) and Biber (2006) all found that when student writing (usually thesis or PhD dissertations) was compared with academic journal articles, the student work contained considerably more bundles. In one example, Hyland (2008b) found 149 bundles in Master's theses, 95 bundles in PhD dissertations, and 71 bundles in academic papers. There are a number of possibilities why higher-level genres of writing might contain fewer bundles. Hyland (2008b) postulated that, "student genres are more phrasal than the published articles, and that apprentice writers are more dependent on prefabricated clusters in developing their arguments, with the Ph.D. students closer to the expert writers"

(p. 50). It is interesting that the findings from the ALDSN and ALDS-UL corpora are consistent with novice writers using more frequently occurring bundles, since the difference between ALDS 1001 mini-paper and Master's course paper is much greater than the transition from Masters to PhD, or student to professional academic. There has been no research I am aware of that has established a connection between first year writing and Masters level writing in terms of the reliance on prefabricated sequences. The current findings suggest that this relationship between experience and bundle use could extend beyond the student to professional transition into one that extends the length of one's university and possibly high school academic development.

Perhaps the most salient pattern in the comparison of novice to upper-level ALDS writing, was the difference in distribution and heavy use of lexical bundles in an idiosyncratic manner by novice writers. In the ALDS-UL corpus, there were a mere 29 bundles which occurred at a rate of 40 times per million but were used in less than five texts. In the ALDSN corpus there were 59 bundles used in less than five texts. In addition, there is a sizeable gap in the number of bundles disqualified from the initially qualifying 73 and 110 bundles from each corpus. In the ALDSN corpus 51 bundles were disqualified based on concordance line evidence showing they occurred in quotation or failed to meet distribution criteria since they were used in multiple texts by the same author. These same secondary criteria were applied to the ALDS-UL list of 73 bundles, but only 8 bundles were disqualified. These findings offer strong evidence that the first year genre of ALDS writing has a much higher incidence of repeated clusters, and reliance on prefabricated material. Answering why this might be the case is a much harder task. It is possible that this

reliance on prefabricated material is not a mark of less proficient writing, but is simply a convention of the genre. It is the case in many novice academic genres that teachers expect students to more clearly establish the foundations and justifications for their epistemic knowledge, which could require much greater repetition of key terms, ideas and references (Oakey and Hunston, 2010).

In order to look at this issue of genre convention vs. writing proficiency, papers were grouped according to the mark they received on the ALDS 1001 mini paper and tested as groups. Papers were divided into the following four levels: Low, containing 29 papers that received a grade of 6.5 out of 10 or lower, Medium-low, containing 25 papers with a grade of 7 or 7.5 out of 10, Medium-high, containing 21 papers that received a grade of 8 or 8.5 out of 10, and High, containing 25 papers that received 9, 9.5, or 10 out of 10. A frequency cut-off of three was chosen, which normalizes to 100 times per million. Since the only thing being compared with this search is total number of bundles, the frequency cut-off was not overly important. It was only important that there were enough bundles counted to see some sort of comparison between the levels.

The results of total bundles from the four levels of novice ALDS writing produce a picture of declining bundle use from less successful to more successful writers. The low group of papers, containing 27,904 words, had 137 four word bundles. This was easily the highest number, far surpassing the 77 bundles of the high group, whose papers totaled more words at 29,997. The medium-low group used 86 bundles in their papers (25,921 words), and the medium-high papers used 75 bundles (23, 465 words). These results seem to support the notion that there is a link between the grade a paper receives (the 'success' of the paper) and its

proportion of lexical bundles. The question of why this might be, and what impact it could have on pedagogy for novice university students will be addressed further in the analysis of questionnaire and interview data. In order to provide some measure of control for the aggregate results presented above, results from the comparison corpus of intermediate ALDS writing will be presented next.

The intermediate corpus is composed of 21 texts totaling 127,989 words. As mentioned in the methodology, the writing in this corpus, although technically at the 2nd year level, was done by a group of students with a much greater variety of experience with ALDS writing. Students in the course range from 2nd year to 4th year, some of them taking the 2nd year class after completing many others in ALDS. Because of this mix, papers are again grouped by success level (A-, A, A+ papers in one group, B+ and lower in the other). Looking at bundle use in these groups thus addresses questions about 'high-level' 2nd year writing, or low-level 2nd year writing, as opposed to questions about the experience level of the writers specifically. The first group of less successful intermediate papers had a total of 63,049 words, and had a total of 135 bundles occurring at least four times over any number of texts (distribution will not be considered for this control). The more successful papers, which received grades of A- or higher, totaled 64, 937 words and had 172 four-word bundles overall.

Interestingly, these results do not show the same pattern of less successful writing using a higher proportion of bundles. Proportionally, the lowest level of novice ALDS writing still used more lexical bundles than either group of intermediate writers, but the results still show that the use of prefabricated language does not have a simple correlation with experience and success in writing.

When all of the intermediate papers were analyzed together, they contained 228 lexical bundles at a frequency of 40 times per million (normalized to five). This is vastly more than either the upper-level or novice writers. Because there were essay questions on the take home exam being answered directly by the intermediates, it is quite possible that the inflated bundle numbers have more to do with topic-bound writing and narrower subject matter than anything to do with experience and competency. The results from this control show that more investigation needs to undertaken on questions surrounding prefabricated language and student writing development, but the overall trends observed by Hyland, Biber and others are certainly not refuted by any of these findings.

7.2.1 Co-occurrence of bundles between the ALDS-UL and ALDN corpora

The comparison of bundle occurrence across ALDSN and ALDS-UL will begin with the top 30 most frequent bundles, excluding topic-bound bundles, from both corpora. Topic-bound bundles were left out of this comparison because many of them were specific to the essay topic and course material and would not provide a meaningful indication of importance across the two corpora. In assessing how important lexical bundles might be, it is worth getting an idea of whether novice writers use the most frequent upper-level bundles and vice a versa. Figure 7.6 shows how many times the novice writers (red) use the top 30 most frequent lexical bundles in the ALDS-UL.

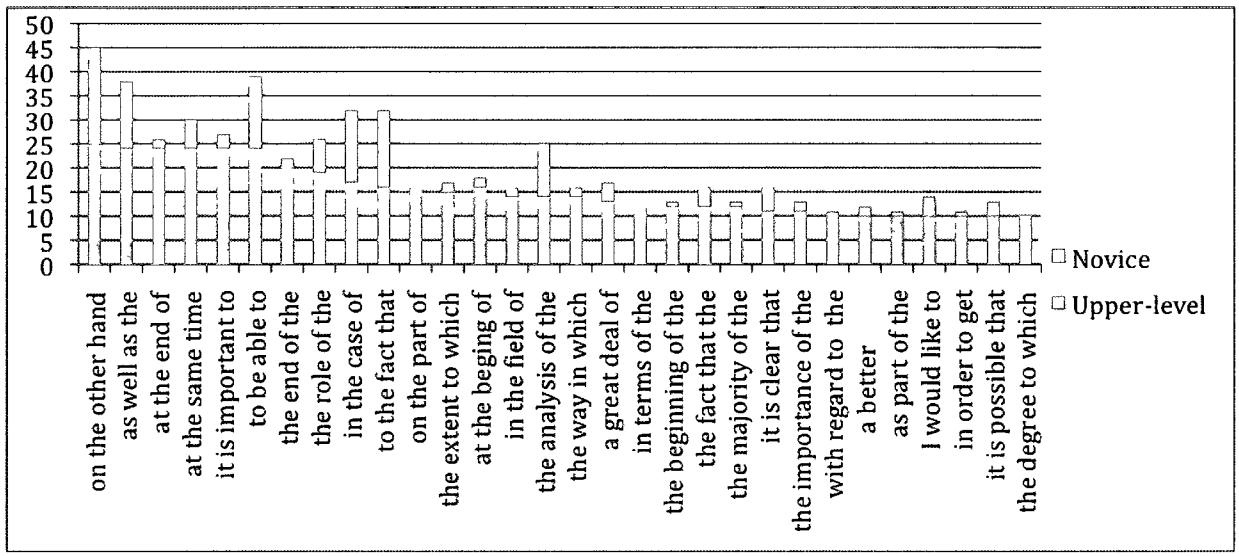

Figure 7.6. ALDS novice writers' use of frequently occurring upper-level bundles

Figure 7.6 shows that the 10 most frequent upper-level ALDS bundles are used relatively frequently by lower level ALDS writers. It must be noted that frequency counts from the ALDSN corpus are expected to be lower than the ALDS-UL corpus based on the smaller size, therefore bundles that have only a few less occurrences in the ALDSN are proportionally as frequent if not more so in some cases. In the top 10 upper-level bundles, novice ALDS writers used five bundles infrequently, and six bundles were used at an equal or greater proportion. Of the five bundles that occurred at a lower proportion in the ALDSN, three were multi-purpose reference bundles, one was a procedural reference bundle, and one was an obligation-stance bundle. The other six bundles in the top 10 that were used at the same or greater proportion were procedural reference, epistemic stance, and text-organizers. It thus appears from this small sample that text organizers, especially framing bundles and the contrastive *on the other hand* are relied upon by both novices and upper-level writers in ALDS.

After the top 10 bundles in the ALDS-UL, there is a stark drop off in how frequently the novice writers use the remaining 20 bundles. Only one of those 20, procedural reference bundle *the analysis of the*, was used at an equal or greater proportion in the ALDSN. Three of the remaining bundles were not used at all, and most of the others were used between one and three times. From figure 7.6, it seems that upper-level ALDS writers use a far greater proportion and variety of framing text-organizing bundles. Two of the 10 framing bundles in the top 30 had no occurrences in the ALDSN corpus, and eight others were used well under the equivalent proportion for the novice corpus. While these results are not sufficiently broad or representative enough to draw any strong conclusions, they do suggest important differences in the patterns of use and functional distribution of bundles between upper-level and novice ALDS students.

In figure 7.7, the inverse of the data from figure 7.6 is presented, with the top 30 novice corpus bundles showing their corresponding usage in the upper-level corpus (again excluding topic-bound bundles).

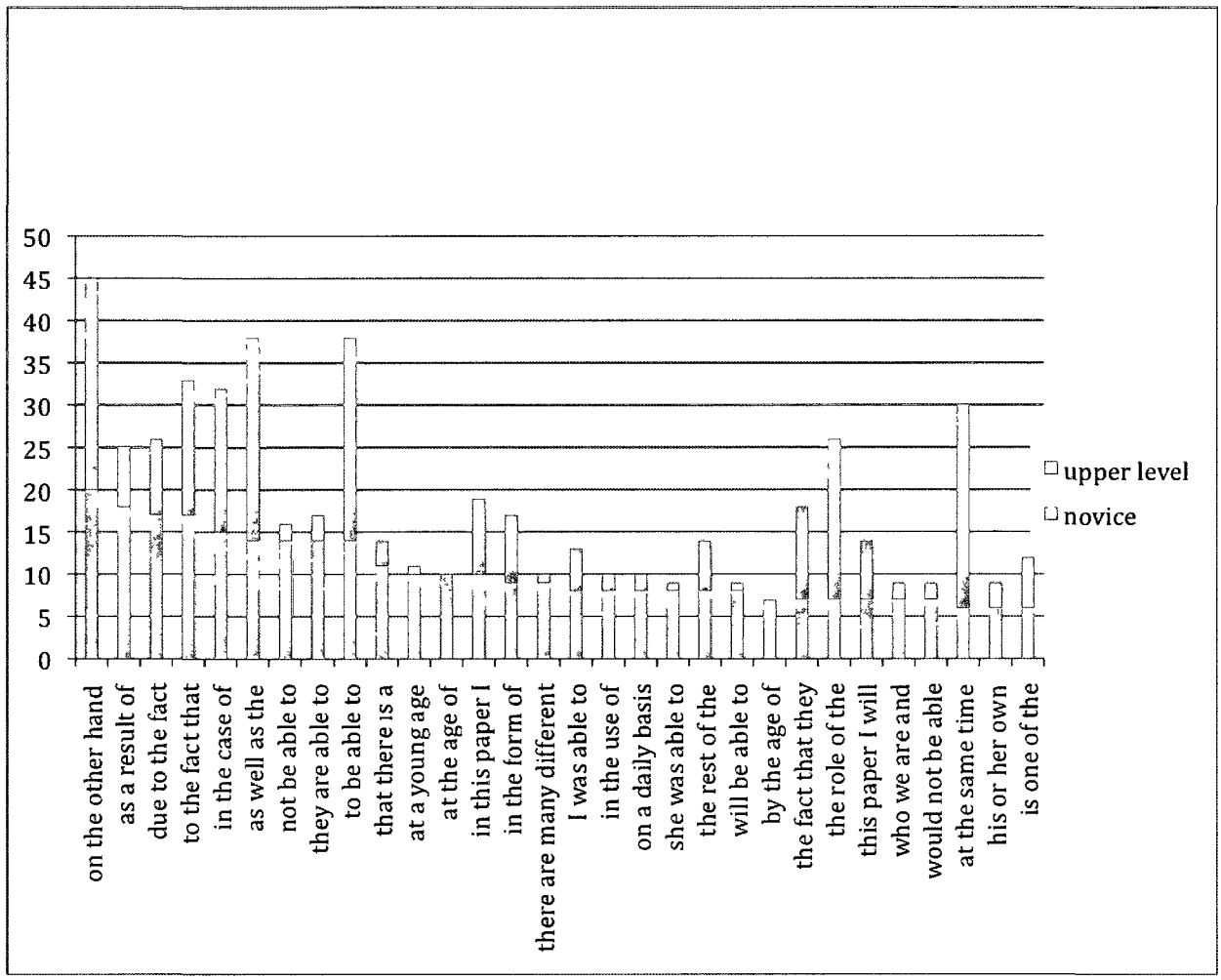

Figure 7.7. ALDS upper-level writers' use of frequently used novice ALDS lexical bundles

Figure 7.7 shows that once again the 10 most frequent bundles are used quite heavily in the opposite corpus, however there are uneven results for the remaining 20 bundles. Unlike in figure 7.6 where novice writers barely used any of bundles outside of the 10 most frequent ALDS-UL bundles, figure 7.7 shows a much less even decline. While there were two bundles that were not used at all, they were both time specific reference bundles that were highly linked with subject matter concerning first language acquisition, a topic that had far less coverage in the ALDS-UL. The most obvious pattern emerges in the use of procedural/ability reference bundles.

Eight of these ability/reference bundles were hardly used at all in the ALDS-UL, suggesting that such description is more diverse in language in upper-level writing and is likely influenced strongly by the ALDS 1001 subject matter.

Combined with figure 7.6, the results shown in the figure 7.7 support the notion that bundle usage between upper-level and novice writers varies considerably. In order to see how these differences might impact the grade level of novice writing and potentially hold importance for writers as they develop in the discipline, results from the ALDSN corpus with texts grouped by level will be presented next.

7.2.2 Results of upper-level bundle use in four levels of novice ALDS writing

Although the use of upper level ALDS bundles was quite infrequent for most of the 65 bundles identified from the ALDS-UL, there were numerous cases of upper level bundles having one to four uses by novice writers (below the frequency cut-off of five used in the ALDSN). To get a better sense of how novices were using these upper-level bundles, all 65 ALDS-UL bundles were searched in the novice corpus and entered into two separate graphs. The first graph (figure 7.8) shows the number of times the low, medium low, medium high, and high groups of novice writers used each major functional category of ALDS-UL bundles (text organizing, reference, and stance) and figure 7.10 shows how each level of novice paper used upper-level bundles in the ALDSN corpus. Because there are bundles in each group that have far more occurrences than others (i.e. *on the other hand,* which has 25), and with some novice writers using far more of the bundles in question than others, the results have some issues of imbalance.

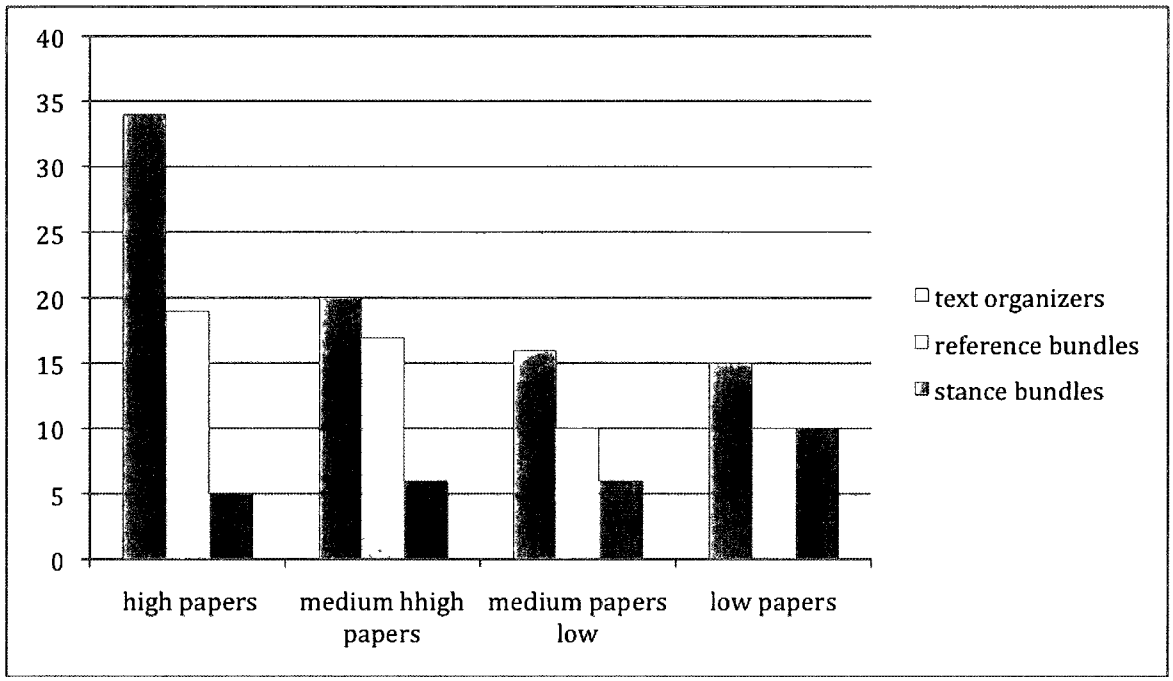

Figure 7.8. Use of upper-level lexical bundles in four levels of novice ALDS mini papers

With topic-bound bundles removed from consideration[25], there were 19 text organizing bundles, 20 reference bundles, and nine stance bundles considered in each major functional category, although not every bundle had occurrences in the ALDSN corpus. Taking this into account, there were 14 text organizers, 16 reference bundles, and eight stance bundles from the ALDS-UL that had at least one occurrence in the ALDSN (displayed in the chart above). In terms of overall occurrences in novice papers considered for each major functional category, there was variation in part due to the large number of reference and text organizing bundles compared with stance bundles. In total, there were 85 occurrences of text organizers, 57 occurrences of reference bundles, and 27 occurrences of stance

[25] Neither figure 7.8 or 7.10 includes results from topic-bound bundles for the same reasons they were omitted in the previous two tables.

bundles considered for figure 7.8. The results from the chart show two noticeable trends: (1) high level novice mini papers (especially those at the highest level) use a great deal more text organizing and reference bundles than lower level papers (2) stance bundles are used more [twice as often] in lower level mini-papers.

Based on these results, it seems as though there is a connection between the functional type of bundle novice writers' use, and the grade they receive on their papers; however, the more frequently occurring bundles occurred in both high and low graded papers. For example, *on the other hand* was used five times in the high papers, five times in the medium high, two times in the medium low, and five times in the low. This indicates, not surprisingly, that many bundles are not indicative of either strong or weak writing by their presence alone, but it is the way they are used and the surrounding content that defines the success of the writing. To get a better sense of how some of the less frequent bundles in the ALDSN were being used and whether there might be a relationship between level of paper and less frequently used bundles, figure 7.9 shows the same statistics on novice use of upper-level bundles, but only for those bundles which occurred less than five times in the ALDSN corpus (the minimum frequency cut-off.)

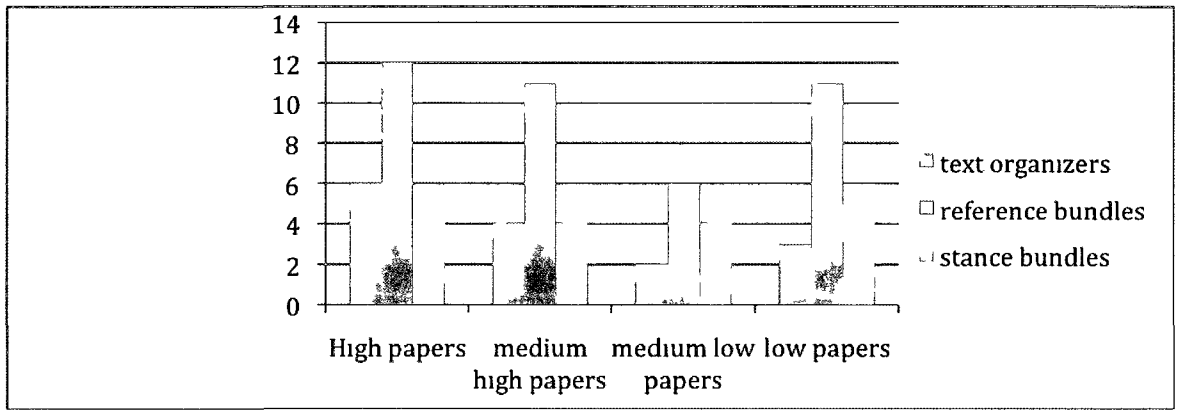

Figure 7.9. Relationship between low frequency novice lexical bundles and upper-level bundles

The results in figure 7.9 do not provide a clear answer to the question of how low-frequency novice bundle use might correlate with the given grade. For reference and stance bundles, it appears that there is a relatively even distribution of upper-level bundle use across all four levels of novice papers. For text organizers there does seem to be a trend towards higher-level writers using upper-level bundles more. High and medium-high papers had a combined 10 uses of text organizing bundles, whereas medium-low and low-level papers had only five occurrences combined. This is further evidence to support the notion that using the appropriate text-organizing bundles and using them with enough variety is correlated with stronger ALDS writing at the novice level. These results could also support greater emphasis being placed on certain bundles in pedagogical situations in an attempt to better prepare students for ALDS writing and raise their awareness of such sequences when reading.

After seeing that higher level novice papers contained more bundles from the ALDS-UL corpus, the 28 remaining non-topic-bound bundles in the ALDSN, which were not used in the upper-level papers were examined to see how they

corresponded with the mark assigned to the paper. This comparison was done to see whether higher-level novice papers contained more bundles in general (both novice and upper-level) or whether there might be a different pattern when those bundles specific to the ALDSN corpus were considered.

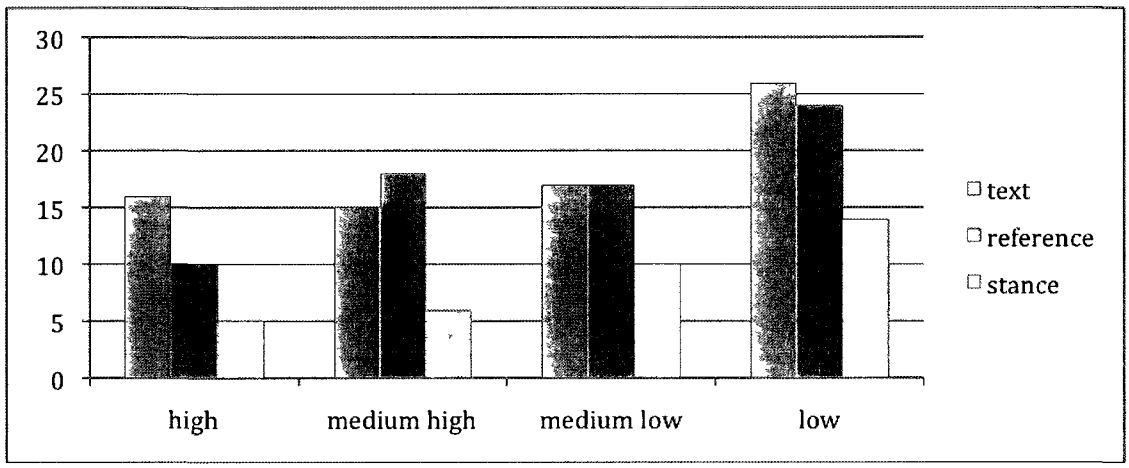

Figure 7.10. Use of lexical bundles occurring only in the ALDSN in four levels of novice papers

From figure 7.10, it is clear that when looking at bundles that only occurred in the ALDSN corpus a different pattern emerges across the four levels of novice writing. For each of the three main functional categories the lowest level of papers contained the most lexical bundle usage. In the low category of papers, all three functional categories showed higher use than for any other level, especially for stance bundles, which occurred 14 times in the low level papers, and only five times in high papers.

It is important to note that there were different proportions of occurrences in each functional category than those calculated for the upper-level bundles: 74 total occurrences of text organizing bundles, 69 occurrences of reference bundles, and 35 occurrences of stance bundles in comparison to 87 text-organizers, 57

reference, and 27 stance. These differences in proportion of bundles do not dispel the notion that novice writers seem to use many bundles that are specific to the novice genre and subject matter, and the use of such novice-specific bundles correlates with lower marks on the mini papers. Considered alongside results from figure 7.8, these findings provide the strongest evidence in this research that certain lexical bundles help ALDS writers, in both the novice and more advanced genres of the discipline, write papers successfully. In addition, it appears that the use of text-organizing bundles hold special importance in ALDS writing at the novice level. Writers from ALDS 1001 who used any of the 14 text organizing bundles that occurred in both corpora were more than twice as likely to be in the high mark range than the low one. Further discussion of how these results and those from previous sections impact the research questions regarding lexical bundles will be considered next.

7.2.4 Discussion of lexical bundle results

The first question posed about lexical bundles in this study was the following: How can corpus research concerning lexical bundles be used to identify novice and upper-level writers in a single academic discipline? More specifically, this question was meant to uncover which level of ALDS writing used the most bundles, and whether these bundles were used in a similar way in different levels of novice papers.

The results presented so far strongly indicate that novice ALDS writers use far more bundles than their upper-level counterparts. In terms of overall bundle use, the pattern that stands out most is the heavily idiosyncratic way that novice

ALDS writers use lexical bundles. Indicated in part by the disproportionately large number of topic-bound bundles, this idiosyncratic use is made most obvious by the disparity in bundle distribution between the novice and upper-level corpus. With 110 bundles disqualified for either distribution or direct quotation criteria, compared to only 37 bundles disqualified in the ALDS-UL, novice writers are engaging with the subject matter with a heavy reliance on formulaic and most likely prefabricated expressions.

In addition to differences between the ALDSN and ALDS-UL, there was also evidence that the number of bundles used is associated with grade level in the novice writing. In order to demonstrate the relationship between overall bundle usage (excluding distribution criteria) and grade level, a database of total bundle numbers corresponding to four grade levels across the novice and intermediate corpus were analyzed in SPSS 19. Data from upper-level bundle use was omitted from this SPSS analysis because there was no variation in grade for the CPALS and graduate/4rth year course papers.[26] I used eta to test for the degree of association between total bundles and grade level. According to Argyrous (2011),

> Eta... [is] a measure of correlation where the dependent variable is measured on an interval scale and the independent variable is categorical. Eta is therefore extremely useful in situations where we want to compare groups defined by a nominal scale in terms of some interval/ratio scale (p. 250).

Like other correlations, eta ranges between zero and one. The eta analysis showed that the association between bundle and grade level was .83, indicating a high degree of association between the two variables. Eta squared indicated how much of

[26] There was no way to run further correlations with SPSS such as Kendall's Tau-*b* or Spearman's Rho tests because of the small number of cases.

the variance in bundles can be explained by grade level. In other words, approximately 69% of the variance in bundle use can be explained by grades assigned to novice and intermediate papers.

Although this study does not examine the psycholinguistic reality of formulaic language or prefabrication in ALDS student language, it seems likely that students are learning these set phrases either from quotation, readings, or lectures and repeating them in writing with very little to no variation. Typical of the formulaic processing described by Wray (2008), Pawley and Syder (1983) and Weiner (1995), students initial engagement with large numbers of new concepts and subject areas seems to begin by processing chunks of language that represent such concepts (*zone of proximal development, second language acquisition* etc.) perhaps without fully understanding the semantic compositionality of the phrases. While students use quotations in their papers partly because they are required to in the marking rubric, there could also be another reason why so many novice writers repeat the same quotations and paraphrases. Dealing with complicated concepts, it is much easier from a cognitive and creative standpoint to rely on the wording provided by the professor or other writers in the field. Relieving some of this cognitive burden associated with writing in a new discipline, in many cases for the first time in university, was undoubtedly important for many students in ALDS 1001.

The second part of the question posed about patterns of lexical bundles asked whether bundles were used in the same way between the two corpora. In general, it does seem that the bundles that are shared between the two corpora display very similar uses in context. In many cases, bundles have a highly restricted

grammatical and functional range based on their structural and semantic properties. For example, the bundle *on the other hand*, which occurred at a very high frequency in both corpora, was used in almost exactly the same fashion by both novice and upper-level writers. Because this bundle falls well towards the idiomatic end of the spectrum in regards to its figurative meaning and relatively fixed structure, a writer has limited choices in terms of how or when the bundle can be applied in text.

Other bundles are far less idiomatic: as Cortes (2004) says, "their meaning is transparent, fully retrievable from the meaning of the individual words that make up the bundle" (p. 400). While the bundles may be transparent in terms of their literal and figurative meanings, the functionality of these bundles, and the position they occupy in the conventional discourse of applied linguistics is far less transparent. Writing successfully in any discipline requires students to know not only what words mean and what combinations are grammatically correct, but which specific combinations are preferred in their target genre, be it first year mini paper or graduate term paper. The results from the comparison between the ALDSN and ALDS-UL corpora suggest that some of this genre knowledge is not known or transparent to novice writers. While the heavy reliance on quotation and set phrases may be a more accepted convention in the novice genre of ALDS writing, novices had noticeable difficulty using text organizing bundles in an appropriate manner. For example, the bundles *that there is a, as well as the*, and *not be able to* all displayed a number of awkward sounding or grammatically incorrect usages. In addition to misusing some of the bundles, novice writers also might be missing variety of text-organizing and reference bundles that help to execute the rhetorical tasks required in the mini paper. Results from figure 7.8 support this notion,

showing that high level novice papers had more than double the number of upper-level text organizing and reference bundles. Although far more data is needed to answer these questions more conclusively, results do support the findings of Cortes (2004) that reading and lectures alone are not adequate in themselves as means for students to acquire and master the lexical bundles commonly found in the writing of a target genre.

The next research question regarding lexical bundles was whether or not there are key bundles in ALDS that have significant occurrences with either high or low level novice ALDS writing. In looking at the questions related to level of writing (the grade it achieved) and the use of lexical bundles, the four levels of novice papers were tested in the following ways: a list of clusters at each level of ALDSN papers was generated, each bundle from the ALDS-UL corpus was matched with its usage in the four novice levels, and each bundle used exclusively in the ALDSN was matched with its usage in the four novice levels. This matching of bundle use with level of writing was not undertaken with the ALDS-UL corpus because that writing was all considered to be representative of upper-level ALDS genres. While past research had provided some expectations for findings in this study, there had not been anything done that specifically looked at how novice writing at different levels compared to upper-level writing in terms of lexical bundle usage. Thus, the most striking results from this section came from the graphs which showed how upper-level bundles tended to occur more than twice as often in high and medium high level mini papers. This correlation was most apparent in the use of text organizing bundles, with reference bundles also showing greater frequency in high level mini-

papers. Stance bundles, even those used in the upper-level corpus, did not correlate with higher grades on the mini papers.

The potential importance of text-organizing and stance bundles was further supported by results shown in figure 7.10 that the novice writers who used text organizers and reference bundles unique to the ALDSN corpus tended to fall into the low or medium low mini paper levels. These results suggest that there are not any 'key' bundles, which can be isolated as crucial in the novice ALDS mini-paper genre. This is because even the upper-level text-organizers and reference bundles were used a few times by low level writers; in many cases, a bundle occurred in high-level and low-level papers.

Since many upper-level bundles only occurred a handful of times in the ALDSN, it is not possible to label a bundle as 'key' if it has only two or three occurrences in high-level papers. A more reasonable conclusion is that there are 'key' functional groups of bundles, which writers must be able to use in the right contexts and in the right forms to achieve the rhetorical exigencies of mini paper writing. Further inquiry into the four levels of mini-papers, including cluster lists for each category, did not provide any further support for key bundles. In order to test these findings further and gain more understanding of how bundles influence the grade a paper achieves, a larger database of text is needed which can be separated into large sub-corpora of papers by level.

To triangulate some of these findings about bundle usage between novice and upper-level writers, and to assess how important the learning and focus on lexical bundles might be for writing development, results and discussion from questionnaire and interview sessions with four ALDS students are presented next.

7.3 Results and discussion of questionnaire and interview phase

The final phase of this study involved gathering some personal accounts of students currently studying in the ALDS program at Carleton. These students from the ALDS M.A. (Helen and Victor) and undergraduate levels (Clara and Leo) were interviewed to see how they viewed the writing process in university, what challenges they faced in writing successful ALDS papers, and what types of support would help them improve as writers. These questionnaire and interview sessions were meant to investigate how novice and upper-level writers were similar or different in these regards, as well as how lexical bundles might fit into future pedagogy based on their evaluation of writing needs and concerns.

7.3.1 Results and Discussion of Questionnaire

The first question on the questionnaire regarding writing (number 4) asked students to rate out of 10 how challenging the component of writing was in university courses. All four of the participants felt that writing was as difficult as any other part of the course (10 being most difficult), with the two non native, and bilingual English speaker answering 9, and the two native English speakers answering 8. It was interesting that the native English speakers, regardless of ALDS experience or level, reported writing as slightly less demanding than the other participants.

In terms of writing approach, all of the participants answered that their approach since either high school or first-year university had changed. The MA students listed components relating to writing process (drafts, editing, time management) as the biggest changes, whereas the novice ALDS writers listed knowledge of vocabulary and use of academic sources. For questions 7 to 10

relating to language and content aspects of essay writing, there was not a great deal of variation in any participant's answers. The content related aspects were valued more highly in general, but except for idiom use, all language aspects were ranked as relatively important. These questions proved useful only in verifying that language and content aspects both play a significant role for novice and upper level ALDS students. With the limited amount of context given, it was not possible to tell how the students felt about language and content aspects in relation to each other, especially since the answers on the Likert scale were so similar.

For question 11, which asked where students learn (have learned) most about writing, 'working with students' and 'readings' were ranked as the most important. Helen and Leo both ranked professors as an important source of writing support, whereas Victor listed tutorial services and Clara listed high school classes as most important. For question 12, all participants said that they do try to use expressions they hear in class as well those from readings in their writing. This is not surprising though as it is difficult to know how much success they have in this regard. Finally, three of the participants ranked ideas and content as being a primary factor in receiving high marks on writing assignments in ALDS. Only Clara did not follow in this case, ranking 'content and ideas' as a 3 out of 10 in importance. These answers will be discussed further in the next section.

7.3.2 Results and discussion of interview component

The results from the interview component will be split into six main questions that were addressed with each student, and have a basis in the literature on writing development.

The first question put to each student was how or if their approach to writing had changed either since high school or university depending on their most recent transition. While Leo, in his first year of undergraduate studies, could not pinpoint any change because he felt he had no tangible approach to writing before university, Clara was forthcoming in this regard. As a fourth year student, Clara said that she had undergone a significant shift in her approach from high school to university, but had not shifted in her approach as much going from first year to fourth year university. This was reflected in the questionnaire as she answered that 'high school classes' were where she learned the most about writing. Although she felt that high school had taught her a great deal about writing, she felt that the transition to university as a writer was a difficult one: "I think it is good, we have to grow up and get to that point [writing more research based papers]...the problem is we don't have that transition, you can't just become a great writer, for years in high school you have been writing a different way" (personal communication, March 16, 2011).

Both MA participants also reported this lack of support in writing during the transition to grad studies. Victor said that while his approach had certainly changed, he was not sure in what way he needed to change his approach coming to the MA program. Helen also said that "I could have used more support" (personal communication, March 24) in coming to the Masters program, but she could not pin point exactly when her approach to writing shifted or changed: "in general I noticed through undergrad that the earlier I started the better I did...my writing from then till now has definitely improved dramatically because now I start really early", but for her it was a more gradual shift as she gained experience in other disciplines, and even when she was away from school before the MA program began. The difficulty

Victor and Helen had in recalling exactly how or when their approaches to writing changed might partly be due to the lack of transitional support and meta-knowledge that is given to students as they pass from discipline to discipline or move up in level. Giltrow (2002) says that the most important difference between high school writing and university writing is that university writing is situated in a research institution (p. 26). This awareness of how university and high school writing are socially and academically situated was not expressed by any of the participants.

The next part of the interview delved into what parts of writing were most difficult for the students, and what value they placed on writing as a part of ALDS courses. All of the students agreed, as they expressed on their questionnaires, that writing is one of the most difficult parts of university. However, they had very different ways of evaluating the importance of writing in their ALDS courses. Victor and Leo, although at different levels of the discipline, saw the value of writing in a similar ways. They valued writing primarily as a tool for evaluating students, and in Victor's case, in preparing students for future academic jobs. Clara and Helen had a different view that was more in line with the freshman students at Harvard that Sommers and Saltz (2003) interviewed. Clara and Helen felt that writing served a purpose not just for evaluation, but also in helping students develop their ideas and do meaningful intellectual work. As Clara said, "I really think writing is important...to actually be able to share the knowledge that you have you need to write it or present it in some way" (personal communication, March 16, 2011). Helen echoed a similar sentiment saying, "you don't know if you understand something until you explain it to somebody" (personal communication, March 24, 2011). These answers also connect back to Ong's (1986) view of writing as a

technology, allowing us to think and develop our ideas in a way that is different from what speech or thought allows us to accomplish.

It is interesting that despite Helen and Clara's awareness that writing is important in formulating thinking and communicating with others, they both indicated significant struggle in understanding how to get the most out of their writing and produce 'expert' level work. Clara was especially frustrated in this regard, saying that "you are not given the skills of how to write rich research papers...I'm fourth year and I still haven't mastered that skill, I feel really sad that I haven't progressed that much...I feel that I have only progressed in maybe using key words to make the TA or professor, like impress them, but it's not really that I understand them" (personal communication, March 16th). This sadness in failing to produce 'rich papers' could be due to several sources, but the problem of using key words to impress without fully grasping their meaning and impact on academic style seems like an issue of genre and language and content knowledge.

The novice and upper-level writers interviewed also said that knowing how to write and use language in a discipline-appropriate way was a struggle. Helen said that, "we all sort of struggle with... even at our level, even considering how much reading we're doing, with what a paper in our program really looks like, and what it sounds like". In her assessment of what is most important in ALDS papers, Helen added that, "it's the style, how they are expressing things, how they are linking things together" (personal communication, March 24, 2011). Giltrow (2002) and Oakey and Hunston (2010) also point out the importance of linking and connecting ideas in applied linguistics. In their book on applied linguistics writings, Oakey and Hunston (2010) include a chapter specifically for "identifying and reporting other

people's point of view" (p. 199). This guide to applied linguistics writing uses data from academic writing corpora to illustrate authentic examples and let students work through these components of essay construction.

Unfortunately, instruction like this is not offered at the first year or MA level. All the participants agreed that more support in writing was needed, both from professors in their feedback, but also in structured classroom activities. Clara expressed her frustration with the guidance on writing from professors, saying that "you just get feedback and it's useless feedback...I'm not going to look at it, I'm going to look at my mark and that's it" (personal communication, March 16, 2011).

In addition to questions about the importance of writing, participants were also asked to elaborate on what makes a piece of writing in ALDS successful. In their elaboration, all of the participants initially emphasized the importance of content and clear development of ideas in their papers. However, when they reflected more, there were a variety of answers about the importance of language and vocabulary. Leo asserted that, "if you have the concept it won't get you the full marks, vocabulary helps, as well as the understanding of how to use the vocabulary" (personal communication, March 10, 2011). In response to the question, what makes a good ALDS Master's paper, Helen said that, "I'm always impressed when I read someone's papers when they are using the language fully, when they have made each word count" (personal communication, March 24, 2011).

These answers from Helen and Leo indicate that there is something beyond the rote learning of vocabulary and memorizing of phrases that goes into good, genre appropriate writing. From their answers, it seems that there is not enough done either by professors, or outside support to help students understand the

specific ways that ALDS writers use vocabulary and express their ideas. As Cortes (2004) found in her research, reading articles and listening to a professor speak does not do enough to give students an adequate understanding of how to implement lexical bundles in their writing. This elusive ability to "use language fully" perhaps could be understood better with corpus data like those provided in this study; however, presenting students with data about expert usage of lexical bundles or lists of key bundles may not be a helpful approach. Even if students feel they need such information about expert writing, they may not be at the developmental learning stage to internalize information about bundles that is presented to them. For teachers and diagnostic testers, information about lexical bundles in novice and upper-level papers of different levels (grades) could more useful. If teachers were able to see how often students were repeating the same structures, perhaps they could identify problem spots in essays or weak points in writing more generally, and provide students with more opportunities for practice and directed support.

These questions about what makes a piece of ALDS writing successful, led to further elaboration about what kind of support might best equip students to improve their writing. While Leo did not have a clear answer on this question, Victor, Helen, and Clara all answered in a similar fashion. They all made clear that what would be most helpful in achieving successful ALDS papers would be seeing more examples of work produced by their peers. Both upper-level participants said that although they were instructed with criteria and guidelines for writing papers, they had very little opportunity to actually see or read writing on similar assignments produced by other students. Helen stated, "I think to improve in your

writing you need to read other people's writing of the same paper. It's rather intimidating to go to somebody, so, can I read your paper?" (personal communication, March 24, 2011). Helen, Victor and Clara all expressed feeling somewhat isolated in regard to what other students were producing, and what high level writing actually looked like.

These points about having models of other student writing and being able to read not only high level papers, but also those "below your level and at the same level" (personal communication, March 21, 2011) indicate a strong value for the type of corpus research being conducted in this study. Despite their value, models of writing, and extracted lists of key bundles must still be approached with skepticism as pedagogical tools. As mentioned previously, even if students have a desire to see models of other writing, there is no guarantee such models will help them, even with analytical information about frequently used bundles and prefabricated sequences. Until more research has been done that shows the value of a modeling approach to writing instruction, lexical bundle data is probably most useful to teachers and curriculum designers.

7.3.3 Comparison of qualitative results to past research

To better situate some of the findings from the interview and questionnaire, it is important to look at the way past research has dealt with these issues of writing development and transition. The first issue addressed is what elements of support and learning help writers develop throughout their university careers.

In Haas' (1994) study, she found that the biology student she followed (Eliza) was helped immensely by the mentorship of upper-level students in her discipline.

In her interview session, Helen also strongly expressed that mentorship and help from her peers played a significant role in the turning point of undergraduate studies. When she transferred to a French program from her initial focus of nutritional science, Helen said that "two of my professors really helped me make that transition [in the French program]" and "at the MA level it's working with other students, but sometimes it's the blind leading the blind" (personal communication, March 24, 2011). Clara also expressed strong feelings about the importance of teacher support, saying that lack of development in her writing was in part due to lack of support and availability from her professors. She was dismayed at not only how difficult it was to meet with professors and get personalized instruction, but also how difficult it was to gain any kind of useful feedback (personal communication, March 16, 2011).

In much of the research on writing development (Artemeva and Fox, 2010; Freeman, 1994; Beaufort, 2004; Haas, 1994) it was found that students acquire genre knowledge and develop as writers mostly through implicit learning that takes place from participation and experience in the discourse community. This view of genre learning that de-emphasizes the explicit teaching and de-construction of genres was shared by several of the interview participants. When asked about how he developed knowledge about the writing genres in ALDS, and whether more explicit language or content instruction would have helped him, Victor had the following response: "to be honest, I don't think it's as important... I have acquired implicitly all that...I don't think it's as important, the structural part of writing, you acquire those just from immersion and contact" (personal communication, April 1, 2011). Victor conceded that some extra support in writing would have been helpful,

especially in raising awareness about writing process and genre conventions, but he felt strongly that direct teaching of these forms would not have been helpful for him. He attributed his success to "many factors", including his strong motivation and extensive reading of ALDS research that he undertook upon arriving in the MA program at Carleton. Although he acknowledged that his ESL background made his transition into the ALDS MA writing more difficult, Victor felt that, "it is unlikely that someone could outperform me in ALDS writing if they were not a part of discipline, even an L1 speaker" (personal communication, April 1, 2011). None of the participants directly alluded to difficulties caused by shifts in writing expectations and context between courses, but all of them expressed difficulty with writing partly caused by lack of support and access to models of other students' writing.

Although they were asked in several ways to discern which factors contributed most to their development as writers, none of the four participants spoke about work experience or practice in the discourse community outside of the classroom. This extra-classroom experience was found to be a key indicator of writing performance and genre expertise in students studied by Artemeva (2005, 2008, 2009) and Artemeva and Fox (2010). It is possible that the lack of work opportunities to practice ALDS writing (compared to Engineering opportunities) affected these answers, but it appears that novice as well upper-level ALDS writers tend to undervalue the importance of work experience or other avenues for immersion in the target discourse community.

7.3.4 Summary of qualitative phase

The results from the questionnaire and interview portion of the study shed light on aspects of ALDS writer knowledge and experience. All participants agreed that more writing support for their ALDS assignments would be beneficial to their overall learning experience and grade averages on written assignments. However, the means to achieve better writing skills and results were not clear from the participant's responses. The participants did not feel that explicit focus on language forms and textual features would be the most useful approach. The upper-level participants especially felt that the lexico-grammatical knowledge necessary to write successfully in ALDS (or any discipline) could be attained through reading, writing and other forms of unconscious learning (contrary to Cortes' (2004) findings). In light of these responses from ALDS students, it seems that lexical bundles should not be the focus of writing instruction in university classrooms. On the other hand, all of the participants emphasized that writing expert level ALDS papers was a daunting task, in part because expert language usage is not transparent to many students (supporting Cortes' findings). If participants are going to examine models of student or expert papers in their discipline, as they expressed a desire to do, it could be very valuable to have corpus data about how many lexical bundles are occurring in the writing models, as well as how those bundles are functioning in the discourse.

At the conclusion of the qualitative phase of this research, there does not appear to be a one-size-fits all approach to lexical bundle pedagogy. It seems that

the value of lexical bundle research will be one of awareness raising for teachers, tutors, TAs and curriculum designers. Explicit bundle teaching could prove effective when teachers or tutors are working closely with individual students and have a better sense of their needs and language competency.

General conclusions from the first and second phases of this study will be discussed in chapter 8, as well as further discussion of how this study can inform pedagogy and future research.

Chapter 8 Conclusions

At the beginning of the study, one primary question was posed: Are there any noticeable patterns in the way novice or upper-level ALDS student writers use lexical bundles and can these patterns be used to characterize novice or upper-level ALDS writing in any way? By analyzing aggregate data from Wordsmith tools for three corpora (novice, intermediate, and upper-level), and classifying bundles according to their primary discourse functions, there were several substantive findings.

The total bundle counts from the corpora, as well as the number of bundles removed after distribution criteria were applied, provided the most salient distinguishing feature between novice and upper-level writing in ALDS. The finding that novice writers tend to use more repeating lexical sequences than upper-level writers corroborated previous research (Hyland, 2008b; Biber et al. 2004; Biber, 2006), but was surprising in many regards since no studies had specifically compared bundles in first year writing with Masters level work, and because longer course papers (those in the ALDS-UL) were expected to contain higher levels of repetition.[27]

It was also found that there was a connection between the discourse functions of bundles used, and the level of writing (grade assigned). The comparison of the ALDSN and ALDS-UL showed that there was a considerable difference in the most frequent bundles used in the two corpora and high level novice writers used

---

[27] Longer papers can be expected to contain more repetition of sequences because they have more transitions and other typically formulaic organizing features than short course papers. The task of writing about a single topic or study for many pages can also entail a larger amount of repeated phrases and references.

more of the text organizing and reference bundles found in the ALDS-UL than their lower-level novice counterparts. Results from searching each level of novice paper for occurrences of ALDSN and ALDS-UL bundles suggest that there is a considerable range in the command of lexical bundles between writing levels, and that developing an adequate mini paper for ALDS 1001 is dependent to some extent on lexical bundle usage. These findings, combined with students' interview responses that highlighted a need for more writing support, strongly suggest that lexical bundle research can be a powerful tool in supporting writers and diagnosing deficits in language and rhetorical expression. The current study has benefit to the specific ALDS program investigated (ALDS), and it demonstrates the value in building specialized corpora that focus on a single discipline; as past research has shown (Hyland, 2008; Cortes, 2004; Biber et al. 2004; Biber, 2006) lexical bundles vary in type, number and function depending on the discipline in question.

8.1 Pedagogical implications

The results of this research have a variety of pedagogical implications. The potential for this research to support writing tutors and TAs is addressed followed by implications for teacher education, and the development of pedagogical tools and curriculum design.

8.1.1 Writing tutor and TA support

From the interview portion of this study, as well as past research by RGS scholars (Artemeva and Fox, 2010; Freedman, 1994; Giltrow, 1994, 2002; Dias et al. 1999; Artemeva, 2005, 2008, 2009; Cortes, 2004) it does not appear that providing students with explicit textual deconstructions of academic genres for writing

support, even those informed by lexical bundle research, would prove effective. However, specialized corpus research (and lexical bundle analysis) combined with the qualitative analysis of students' writing knowledge could be helpful in supporting TAs and writing tutors when they work with students individually.

Deconstructing texts and modeling writing for students in a large class is problematic partly because the teacher has no way of knowing whether students are ready to internalize such instruction; it is inevitable that not all students will be learning at the same pace. When a TA or writing tutor (like those at Carleton's Writing Tutorial Centre) work with individual students, they have the opportunity to gauge progress and present models or lexical bundle information about assignments when students are ready (in developmental terms) to receive them. This 'just in time' teaching could be supported with lexical bundle data that reflects other student writing relating to the particular assignment. For example, if a writing tutor noticed that a student was struggling with cohesion in their text, he/she could introduce several examples of text-organizing bundles (in context), or analyze the students' writing to see which common bundles were absent, and which were overused.

As mentioned in the discussion, it could also be very valuable for writing tutors and TAs to see examples of how lexical bundles are used by student writers in their discipline. Many tutors and TAs are aware that writing varies considerably across disciplines, but they are not equipped with explicit knowledge about where these differences lie, and which sequences of words are most frequently used to accomplish rhetorical tasks. Building specialized learner corpora, similar to the ALDSN and ALDS-UL, could take place in a variety of disciplines, and could help TAs

and tutors give students direct instructions about their writing informed by more than intuition alone.

8.1.2 Teacher education

Although high school teachers and university professors may not be advised to present lexical bundle data to learners in the same manner as TAs and writing tutors, there are still many ways this research could support their awareness and development of teaching material.

For professors, the most practical use of lexical bundle research may be to raise awareness about the structure of student writing and what differences might exist between upper-level and novice students, and low and high-level papers. This research has found that there are meaningful differences that can be observed between these groups of student writing, and that there are specific discourse functions (text-organizing in particular) that novice and low-level students struggle to implement effectively. If teachers have more knowledge about lexical bundles specific to their discipline and thorough enough to examine different groups of students, they could provide more useful feedback to students about their writing, and recognize when a student needs extra help with general or discipline specific language.

While this research has value for teachers dealing with native English speaking students (or second language learners who meet native-like proficiency levels), there is also great value for ESL instructors. Granger and Meunier (2008) assert that there is an "urgent need for more empirical evidence of the actual impact of a phraseological approach to teaching and learning" (p. 249). This lack of

evidence about how to use lexical bundles in an ESL classroom poses an obstacle for teachers and curriculum designers alike, but it should not discourage research on the subject. Byrd and Coxhead (2010) emphasize that, "as with other reported corpus data, the problem for teachers is getting access to such data about related lexical bundles"(p. 53). It is important to provide ESL teachers with this corpus data and information about lexical bundles, but it will not hold significant value if the corpus is not properly constructed to represent the target genre. This study of ALDS writing could serve as a model for providing corpus data that is sufficiently representative of student writing at a variety of levels, and could be used to produce ESL teaching materials that assist learners in their transition into academic writing.

8.1.3 Development of diagnostic tools and curriculum

Beyond the benefit that this research could provide those involved with the support and teaching of writing, there is also potential to inform the development of diagnostic language tests and curriculum for writing intensive courses.

In recent years, there have been considerable advancements in the corpus and text analysis tools available to students. While some of these corpora are being used in classrooms (O'Keefe et al. 2007), other programs for analyzing text have been made available for no charge. Programs like Lextutor (Cobb, 2008), and the online tests using data from the academic word list (Coxhead, 2000; Nation and Laufer, 2011) allow students to analyze their own use of vocabulary and repeating phrases (Lextutor), and take tests online to evaluate their knowledge of key academic vocabulary. Although there is some possibility for students to analyze frequent clusters in their writing with these free online tools, there is not enough

data to inform students about how lexical bundles vary across disciplines and experience/proficiency levels. The community of learning that is fostered by online tools such as Lextutor, could be expanded with specific information about lexical bundle usage. Allowing students and teachers to access these tools would raise awareness about formulaic language and lexical bundles generally, and allow students to see how they compare to upper-level, novice, or high/low-level writers in any specific discipline.

These tools have been geared at second language students primarily; however, the notion of diagnostic language testing for all students at the first year level has been used to benchmark students and provide them with the appropriate level of support. Tests like diagnostic English language needs assessment (DELNA) have been administered to first year students and used to benchmark the writing and reading level of first and second language English speakers (Fox and Hartwick, 2011; Fox, 2009). Results on lexical bundle use for novice and upper-level writers from a variety of disciplines (other than ALDS) could assist in the design and interpretation of diagnostic assessments like the DELNA, as well as the language support ultimately given to each student.

In addition to diagnostic tests and online pedagogical tools, corpus research can help support the design of curriculum for ESL and other academic courses, as well as writing centres. While corpus generated findings on lexical bundles can be useful in the design of curriculum, they must be treated with proper caution and skepticism. Hyland (2008b) says, "frequency should never, by itself, determine classroom decisions, [but] learner corpus data can play an important role in the selection, sequencing and structuring of teaching content" (p. 61). To avoid possible

misinterpretations or over-reliance on corpus results, it is important for experienced corpus researchers to work with curriculum designers and teachers; having a reciprocal relationship between those involved with teaching and designing curriculum and those producing corpus research will allow for research that is more pedagogically applicable and more likely to be used effectively by others. The construction of the ALDSN, ALDS-UL and ALDSI were undertaken with this reciprocity in mind. If specialized learner corpora could be constructed and maintained in many disciplines by other researchers and discipline insiders, they could help design teaching content that was more directly relevant and useful for particular assignments and courses. Specialized learner corpora could also be developed for the purpose of ongoing research at writing centres. A writing centre with specialized corpora for a variety of disciplines could increase the understanding of disciplinary commonalities and differences, and encourage additional research among scholars who have an interest in academic writing development.

The findings from this research and the pedagogical implications listed in this section must be considered in the context of several important limitations to the quantitative and qualitative phases of the study, which will be discussed next.

8.2 Limitations and direction for future study

The most significant limitation in this study arises from the comparison of the novice mini papers with upper-level ALDS writing. Although there were meaningful findings from the comparison, the differences in length, subject matter, and purpose limit the strength of any conclusions drawn from the comparison. The differences in

text length and number of texts between the ALDSN and ALDS-UL made it difficult to use distribution criteria in an effective way for many parts of the analysis. It is also fair to assume that the upper-level work was much more thoroughly edited before submission. Editing differences between the upper-level and novice papers were not considered a serious impediment to the analysis because edited or otherwise, ALDS-UL papers were still representative of upper-level writing in the discipline and thus provided a valuable contrast to the novice ALDS work.

Another aspect that limited some of the findings regarding levels of novice and intermediate writers was the small and unbalanced number of papers from each grade level. The sample size of 20-30,000 words of text in each of the four levels of novice writing was too small to conduct more in depth analysis of bundle usage, and grouping papers into four levels instead of according to their exact percentage grade further limited the analysis.

In future studies, it will be important to increase the overall size of the novice and upper-level corpora, finding papers of greater variety and length in the lower and upper-level categories. It would have also been useful to collect a range of levels (assigned grades) in the upper-level writing, allowing for deeper analysis and comparison within the upper-level genres. For example, the current study did not have any means of investigating how lower-level Masters writers compared to novices or other Masters level writers in terms of lexical bundle usage.

Another consideration for future discipline-specific corpus building is the collection of demographic information about all the writers included. In this study, demographic information about writers of the ALDS-UL corpus was not available. Although papers had to meet a grade level standard to be included in the ALDS-UL

corpus (A or higher) there was undoubtedly considerable variety in the academic and writing backgrounds of the upper-level students. Having a more controlled sample and richer information about each writer included would allow for many other questions concerning genre knowledge, writing success, and lexical bundle command to be addressed.

This issue of representativeness was also a limitation with the novice and intermediate writers, although to a lesser extent. While there was demographic information for most participants indicating year of study, major, and language background, there were still six novice students that did not provide any personal information, and six students from the ALDSI that did not provide any information. While it is difficult to ensure that each student at the novice, intermediate or upper-level fits all the criteria established for such a classification, a larger corpus and more thorough background checking of each participant would improve the study in this regard.

As mentioned in chapter 4 and 6, the taxonomy used for this study was adapted from the work of Biber et al. (2004), Cortes (2004) and Hyland (2008b). Because there were many inconsistencies between these studies in how bundles were classified, it is likely that some bundles would be reclassified in the view of a more experienced lexical bundle analyst. In an effort to counter this limitation, a Masters student in ALDS checked the discourse functions and taxonomy for the ALDS-UL and ALDSN corpora after they had been initially classified. Structural classification and analysis of lexical bundles was also omitted from the study. In future research, structural classification of sequences could offer further understanding of how novice and upper-level ALDS writers use lexical bundles by

comparing the most common structural bundle forms between the corpora, and analyzing the differences in structural bundle use for the different levels of novice (and potentially upper-level) writing.

With very little existing research that examines the outcome of teaching and informing students with lexical bundle data, it is important to examine how students who receive instruction with lexical bundles fare in their development throughout university. The pedagogical implications of lexical bundle research must be tested and examined in learning environments across a variety of disciplines in order to better assess their value as a pedagogical tool (Granger and Meunier, 2008; Byrd and Coxhead, 2010).

In addition to investigating how bundles are taught and learned, more extensive tagging of the sections and surrounding context in which bundles occurred could expand lexical bundle analysis and maximize the results gained from specialized corpora like the ALDSN or ALDS-UL. For example, if each mini paper in the ALDSN had been tagged for introductions, paragraph breaks, topic sentences, thesis statements, conclusions, etc., it would have been possible to learn much more about how lexical bundles are used and how valuable they are in certain parts of the text. Learning more about which bundles are used in a certain section, and seeing whether papers that use more bundles in a section achieve higher grades, would allow the researcher to better understand which bundles should be emphasized in pedagogy and which parts of an assignment they correspond with. More textual and rhetorical context provided along with bundle findings, in addition to their discourse functions, would improve the usability of corpus results for students and teachers.

Finally, future studies should consider increasing the amount of qualitative investigation paired with lexical bundle research. Having more interviews with student writers over the course of their university careers (longitudinal approaches) and interviewing professors and TAs involved with the evaluation of writing and teaching could be highly beneficial. With a better understanding of how teachers evaluate writing and what they expect from students, as well as what the students perceive as important in producing successful assignments could improve the usability and relevance of lexical bundle findings.

In the final section, my personal experience as a graduate student and writing tutor are reviewed in order to contextualize the rationale for the study and offer some examples of how this work can be applicable to authentic learning and teaching situations.

8.3.1 Final Thoughts

As a member of the ALDS Masters program at Carleton University, and a writing tutor for first year, graduate, and special students from a variety of programs, questions about why some students succeed at writing, and how they learn to be successful writers were of particular interest both personally and professionally. In my first year of study as a Masters student, I was placed as a tutor in the Writing Tutorial Service (WTS). At the WTS, undergraduate and some graduate students from a variety of programs, mostly in the humanities and social sciences, came for feedback and guidance with writing. Without a significant amount of experience writing in any discipline other than applied linguistics it was my first chance to see

the myriad differences in the expectations and conventions between different fields of study.

It quickly became apparent that what many students needed most was someone who had knowledge of not only general writing strategies, but of the particular criteria (format, language, organization, argument style) that corresponded to assignments in their discipline. As many WTS tutors lacked personal experience outside of their area of study, it was common that students and tutors were limited in their sessions by this lack of discipline knowledge. My interest in pedagogical approaches to writing grew as I became an 'academic coach' for special students[28] at Carleton's Centre for Initiatives in Education (CIE). Unlike at the WTS, my position as an academic coach allowed me to integrate much more into the learning experience of my tutees. From observing them in class, speaking with their professors, and reading their assigned texts and assignment sheets, I was able to provide much more in-depth feedback and guidance as a writing coach.

Working with a variety of students, my curiosity steadily grew about why some students were able to succeed with their writing, while others continually struggled. As I proceeded with my studies in formulaic language and academic writing, I began to offer some students help with parts of introductions, conclusions, and topic sentences by providing them with several formulaic strings (often lexical bundles I would later find in the corpus study) to accomplish specific rhetorical goals, like writing a thesis statement, or citing a research article. These attempts to

---

[28] Special students at Carleton were not enrolled in a degree program, but instead were placed in one mandatory first year ALDS seminar and chose two electives from a program of their choice. If these students achieved a minimum grade point average after two terms they are permitted to join a regular degree program at the university.

blend some formulaic language instruction into tutoring sessions were not conducted as a precisely controlled experiment, but they seemed to show promising results with many of the students (in terms of their writing quality and ability to work independently) and ultimately fueled the rationale and direction of the present study.

A common thread among almost all of the students I interacted with, whether it was students at the WTS, special students at the CIE, first year students in ALDS 1001, or MA students in ALDS, was they felt that more writing support was necessary. The results from both corpus and questionnaire/interview phases could help support writing help services and tutors in understanding what challenges novice writers face as they come from high school or other areas of life. If the findings on which types of lexical bundles are normally problematic for novice students, and which discourse functions are most difficult to master are reproduced with corpus research in other disciplines, they can better equip teachers, teaching assistants and writing tutors to help novice students as they struggle to write and acculturate in a new discipline or level of academia.

One concrete example of how this study could be used as a pedagogical tool comes from my own thesis writing experience. After finishing a draft of the first seven chapters of this thesis, the document was cleaned of tables, page numbers, and other superfluous text (the same as the other papers used for the corpus study) and converted to a plain text file. An additional research paper and thesis completed in the ALDS Masters program were also cleaned and converted into plain text to

offer an appropriate comparison in terms of text length and genre[29]. Comparing the total number of bundles used in my thesis to the other thesis and research paper (ALDS-TR), as well examining the top 20 bundles (excluding topic-bound bundles) from the ALDSN, ALDS-UL and ALDS-TR corpora, I gained several useful insights. For one, I was able to see that my thesis used a roughly equivalent number of bundles to the research paper (155 and 157 bundles respectively), whereas the other thesis used 258 bundles. If lower bundle counts are typical of higher-level writing (as suggested by this study), it could indicate that my writing is meeting at least one standard of Master's thesis work. Another helpful insight came from examining the top 20 frequent sequences from the other two papers. In my writing I had been struggling to find an appropriate way of referring to my own study, using awkward phrases such as *in the study here...* or *the study I've done has.* I noticed from the analysis that the papers from the ALDS-TR used many structure signal text-organizing bundles, and most frequently referred to their work with the bundle *in the current study.* This finding not only improved a small part of my writing expression, but also alerted me that I might need more explicit transitions between chapters; examining the concordance lines for the structure signaling bundles in the ALDS-TR provided clear examples of how often, and in what words other transitions between sections and chapters had been made.

Although the analysis of lexical bundles in my own writing using Wordsmith Tools was done at a small scale to provide an example of pedagogical application, there were still helpful insights gained by comparing my writing to other similar

---

[29] My thesis was 31 469 words when it was converted to plain text, the other thesis was 22 869 words, and the research paper was 21 827 words.

student work, and from my awareness of what normally constitutes upper-level bundle usage (gleaned from previous findings reported in the current study). This self-analysis using Wordsmith Tools provides a glimpse of how corpus tools and lexical bundle knowledge can be used to assist students in evaluating the level of their writing, as well as to identity sequences that can assist them in constructing and linking content or textual information.

My experience as a tutor and thesis writer combined with the findings reported from the two phases of the current research, have provided some meaningful insight into the writing of novice and upper-level students in ALDS. This study has attempted to address a small gap in the literature on lexical bundles, and do so with a more diverse methodology than is commonly found in similar corpus studies. The findings and design of this research can benefit those interested in researching formulaic language and lexical bundles, as well as those studying text and writing from other perspectives.

Appendix A.1. Examples of structural classifications of lexical bundles given by Biber et al. (2004)

Examples of lexical bundles from Biber et al.'s (2004) study (p. 381).

Verb phrase fragments

Discourse marker + VP fragment: *I mean you know, you know it was, I mean I don't*

Yes-no question fragments: *are you going to, do you want to, does that make sense*

Dependent clause fragments

WH-clause fragments: *what I want to, what's going to happen, when we get to*

If-clause fragments: *if you want to, if you have a, if we look at*

Noun phrase and prepositional phrase fragments

Noun phrase with other post modifier fragment: *a little bit about, those of you who*

Prepositional phrase expressions: *of the things that, at the end of, at the same time as*

Appendix B.2. Questionnaire given to both novice and upper level ALDS writers

Questionnaire: Language and Writing

Year of study _____

Major _____

1) What is your first language(s)

2) Do you speak any additional languages? (please name the language(s) and circle the appropriate number)

_____ (beginner) 1 2 3 4 5 6 7 8 9 10 (fluent)
_____ 1 2 3 4 5 6 7 8 9 10
_____ 1 2 3 4 5 6 7 8 9 10
_____ 1 2 3 4 5 6 7 8 9 10
_____ 1 2 3 4 5 6 7 8 9 10

3) If you do not speak English as your first language, have you ever taken any of the following classes?

English as Second Language Yes ☐ No ☐
English for Specific Purposes Yes ☐ No ☐
English for Academic Purposes Yes ☐ No ☐

4) How challenging for you is the writing component of university courses?

(very easy) 1 2 3 4 5 6 7 8 9 10 (the most difficult part of the course)

5) Has your approach to academic writing changed since your first year of university/high school?
Yes ☐ No ☐
5b) If yes, how has it changed?

6) Do you think that writing essays in your classes is a worthwhile exercise?
Yes ☐ No ☐
6b) If yes, why?

7) How important are the following aspects of essay writing in university?

Vocabulary	1 2 3 4 5 6 7 8 9 10
Sentence structure	1 2 3 4 5 6 7 8 9 10
Paragraph organization	1 2 3 4 5 6 7 8 9 10
Idioms	1 2 3 4 5 6 7 8 9 10
Grammar	1 2 3 4 5 6 7 8 9 10
Formatting	1 2 3 4 5 6 7 8 9 10

8) Do you feel you need more instruction on how to use the following components of academic language for your writing assignments?

Vocabulary	(No) 1 2 3 4 5 6 7 8 9 10 (Yes)
Sentence structure	1 2 3 4 5 6 7 8 9 10
Paragraph organization	1 2 3 4 5 6 7 8 9 10
Idioms	1 2 3 4 5 6 7 8 9 10
Grammar	1 2 3 4 5 6 7 8 9 10
Formatting	1 2 3 4 5 6 7 8 9 10

9) How important are the following areas for writing successful essays?

Key terminology	1 2 3 4 5 6 7 8 9 10
Proper Citation format	1 2 3 4 5 6 7 8 9 10
Ideas and concepts covered in the course	1 2 3 4 5 6 7 8 9 10
Additional concepts related to the course	1 2 3 4 5 6 7 8 9 10
Conventions used by other writers in the discipline	1 2 3 4 5 6 7 8 9 10

10) Do you feel you need more instruction on the following components of academic essay writing?

Key terminology	(No) 1 2 3 4 5 6 7 8 9 10 (Yes)
Proper Citation format	1 2 3 4 5 6 7 8 9 10
Ideas and concepts covered in the course	1 2 3 4 5 6 7 8 9 10

Additional concepts related to the course	1 2 3 4 5 6 7 8 9 10
Conventions used by other writers in the discipline	1 2 3 4 5 6 7 8 9 10

11) Where do you think university students learn (have learned) the most about writing? (you can choose more than one)

Professors	(learned little) 1 2 3 4 5 6 7 8 9 10	(learned a lot)
High school classes	1 2 3 4 5 6 7 8 9 10	
Tutorial services	1 2 3 4 5 6 7 8 9 10	
Working with other students	1 2 3 4 5 6 7 8 9 10	
Readings	1 2 3 4 5 6 7 8 9 10	
Other (please specify) _____	1 2 3 4 5 6 7 8 9 10	

12) Do you try to use expressions or phrases in your writing that...

you hear in class	Yes ☐	No ☐
Read in articles/textbooks	Yes ☐	No ☐

13) How important to you think ideas/content are to receiving high marks?
(Not important) 1 2 3 4 5 6 7 8 9 10 (extremely important)

Appendix C.3. Semi-structured interview script

1.) Have you ever taken any ESL classes or classes that focused on improving command of English?

2.) If yes, do you think these classes gave you enough preparation to be a successful University student?

3.) Do you find the writing process difficult?

4.) What kind of support (other than practice) could make this process easier?

5.) In what ways do you think the demands of University writing requirements are addressed in ESL learning?

6.) Do you feel like you are at a disadvantage in a writing intensive class (like those you have taken in ALDS) as a non-native speaker? Why or why not?

7.) Can you tell me a little bit about your writing process? What do you do yourself when you write essays? Has this changed at all since you first came to University?

6) Are there any other areas that are important to writing well, or learning how to write well in ALDS?

7) In achieving a good mark on a mini-paper or essay in ALDS, what do you think that writing should have?

8) Do you think command of language is essential to getting a high grade on applied linguistics writing?

9) Would it be helpful to have a list of frequently used language from experienced writers?

References

Argyrous, G. (2011). *Statistics for research.* London, UK: Sage.

Ari, O. (2006). Review of three software programs designed to identify lexical bundles. *Language Learning & Technology, 10*(1), 30-37.

Artemeva, N. (2005). A time to speak, a time to act: A rhetorical genre analysis of a novice engineer's calculated risk taking. *Journal of Business and Technical Communication, 19*(4), 389-421.

Artemeva, N. (2008). Toward a unified social theory of genre learning. *Journal of Business and Technical Communication, 22*(2), 160-185

Artemeva, N. (2009). Stories of becoming: A study of novice engineers learning genres of their profession. In C. Bazerman, Bonini, & D. Figueiredo (Eds.), *Genre in a Changing World.* Fort Collins, CO: WAC Clearinghouse and Parlor Press.

Artemeva, N. & Fox, J. (2010). Awareness versus production: Probing students' antecedent genre knowledge. *Journal of Business and Technology, 24*(4), 476-515.

Bazerman, C. et al. (Eds.), (2010). *Traditions of writing research.* NY: Routledge.

Beaufort, A. (2004). Developmental gains of a history major: A case for building a theory of disciplinary writing expertise. *Research in the Teaching of English, 39*(4), 136-185.

Becker, J. (1975). The Phrasal Lexicon. Bolt Beranek and Newman Report No. 3081, AI Report No. 28.

Biber, D. & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of Corpora: Studies in Honor of Stig Johansson.* Amsterdam: Rodopi, pp. 181-189.

Biber, D. & Conrad, S. (2001). Quantitative corpus-based research: Much more than bean counting. *TESOL quarterly, 35*(2), 331-336.

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

Bizzell, P. (1982). Cognition, convention and certainty: We need to know about writing. *Pre/Text, 3*(3), 213-243.

Bolanger, M. (1989). Prefabs, patterns and rules in interaction? Formulaic speech in adult learners' L2 Swedish. In K. Hyltenstam & L.K. Obler (Eds.), *Bilingualism Across The Lifespan* (pp. 73-86). Cambridge: Cambridge University Press.

Butler, C.S. (1997). Repeated word combinations in spoken and written text: some implications for functional grammar. In C.S. Butler, J.H. Connolly, R.A. Gatward, R.M. Vismans (Eds.), *A Fund of Ideas: Recent Developments In Functional Grammar* (pp. 60-77). Amsterdam: University of Amsterdam.

Chen, L. (2008). An Investigation of Lexical Bundles in Electrical Engineering Introductory Textbooks and ESP textbooks (Master's thesis). University of Carleton: Ottawa, Ontario.

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

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge: MIT Press

Cobb, T. (2008). Compleat lexical tutor v6.2. Retrieved from http://www.lextutor.ca/

Cooper, M. M. (1986). The ecology of writing. *College English, 48*(4), 364-375.

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

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education,* 17, 391-406.

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

Coxhead, A. (2000). The new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Coxhead, A., Byrd, P. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL,* 5, 31-64.

Dias, P., Freedman, A., Medway, P., & Pare´, A. (1999). *Worlds apart: Acting and writing in academic and workplace contexts.* Mahwah, NJ: Lawrence Erlbaum.

Dias, P. (2000). Writing classrooms as activity systems. In P. Dias & A. Paré (Eds.), *Transitions: Writing in Academic and Workplace Settings,* (pp. 11-31). Cresskill, New Jersey: Hampton Press Inc.

Erman, B. (2001). Pragmatic markers revisited with a focus on you know in adult and adolescent talk. Journal of Pragmatics, 33, 1337–1359.

Flowerdew, L. (1998). Corpus linguistic techniques applied to text linguistics. *System, 26*, 541-552

Fox, J. (2009). Moderating top-down policy impact and supporting curricular renewal: Exploring the potential of diagnostic assessment. *Journal of English for Academic Purposes, 8*, 26-42.

Fox, J. & Hartwick, P. (2011). Taking a diagnostic turn: Reinventing the portfolio in EAP classrooms. In D. Tsagari & I. Csepes (Eds.), *Classroom-Based Language Assessment*: Frankfurt: Peter Lang

Freedman, A. & Pringle, I. (1980). Writing in the college years: Some indices of growth. *College Composition and Communication, 31*(3), 311-324.

Freedman, A. & Medway, P. (1994). *Learning and Teaching Genre.* Portsmouth, NH: Boynton/Cook.

Gatbonton, E., & Segalowitz, N. (1988). Creative automatization: Principles for promoting fluency within a communicative framework. *TESOL Quarterly, 22* (3), 473-492.

Gee, J. P. (1990). *Social linguistics and literacies: Ideology in discourses.* London: The Falmer Press.

Giltrow, J., Valinquette, M. (1994). Genres and Knowledge: Students writing in the disciplines. A. Freedman and P. Medway (Eds), *Learning and Teaching Genre* (pp. 47-63). Portsmouth, NH: Boynton/Cook.

Giltrow, J. (2002). *Academic writing: Writing and reading in the disciplines.* Mississauga, Canada: Broadview

Granger, S. & Meunier, F. (2008). Phraseology in language learning and teaching: Where to from here? In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 247-252). Amsterdam: John Benjamins.

Haas, Christina. (1994). Learning to read biology: one student's rhetorical development in college. *Written Communication,* 11(43), 43-79.

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

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

Hunston, S. & Oakey, D. (2010). *Introducing applied linguistics: Concepts and skills.* London; New York: Routledge.

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

Hyland, K. (2008b). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics 18*(1), p. 41-62.

Hyland, K. (2009). Corpus informed discourse analysis: The case of academic engagement. In M. Charles, D. Pecorari & S. Hunston (Eds.), *Academic Writing: At the Interface Between Corpus and Discourse,* (pp. 110-129). New York: Continuum International Publishing Group.

Kuiper, K. (2004). Formulaic performance in conventionalized varieties of speech. In N. Schmitt (ed.), *Formulaic Sequences: Acquisition, Processing and Use* (pp. 37-55). Amsterdam: John Benjamins.

Langacker, R W. (1987). *Foundations of Cognitive Grammar, Volume 1, Theoretical Prerequisites.* Stanford: Stanford University Press.

Levy (2003). Lexical Bundles in Professional and Student Writing. (Doctoral dissertation). University of the Pacific. Stockton, California

Lin, P. (2010). The phonology of formulaic sequences: A review. In D. Wood, (ed.), *Perspectives on Formulaic Language: Acquisition and Communication* (pp. 174-194). London: Continuum.

Nastasi, B. K., Hitchcock, J. H. & Brown, L. M. (2010). An Inclusive framework for conceptualizing mixed methods design typologies. In A. Tashakkori & C. Teddlie (Eds.), *Mixed methods in social & behavioral research: Second edition,* (pp. 306-330). Thousand Oaks, California: Sage.

Nation, P. & Laufer, B. (2011). Levels tests on-line. Retrieved from http://www.er.uqam.ca/nobel/r21270/levels/

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

Nekrasova, T. (2009). English L1 and L2 speaker's knowledge of lexical bundles. *Language Learning,* 59 (3), pp. 647-686

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

McCarthy, L. P. (1987). A stranger in strange lands: A college student writing across the curriculum. *Research in the Teaching of English, 21*, 233-265.

Miller, C. (1984). Genre as social action. *Quarterly Journal of Speech, 70*(2), 151-167.

Oakey, D. (2002). Formulaic language in English academic writing: A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines. In R. Reppen., S. M. Fitzmaurice & D. Biber (Eds.), *Using corpora to explore linguistic variation.* Philadelphia: John Benjamins Publishing Company.

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

Ong, W. (1986). Writing is technology that restructures thought. In G. Baumann (Ed.), *The Written Word: Literacy in Transition,* (pp. 23-50). Oxford: Clarendon.

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

Rogers, P. (2010). The contributions of North American longitudinal studies of writing in higher education to our understanding of writing development. In C. Bazerman et al (Eds.). *Traditions of Writing Research,* pp. 365-377/ New York: Routledge.

Russell, D. (1997). Rethinking genre in school and society: An activity theory analysis. *Written Communication, 14*(4), 504-554.

Smagorinski, P. (Ed.). *Research on composition: Multiple perspectives on two decades of change.* NY: Teachers College Press.

Schmitt, N., Grandage, S., and Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid. In N. Schmitt (ed.), *Formulaic Sequences: Acquisition, Processing and Use* (pp. 127–151). Amsterdam: John Benjamins.

Scott, M., and Tribble, C. (2006) *Textual patterns: Key words and corpus analysis in language education.* Amersterdam/Philadelphia: John Benjamins Publishing Co.

Scott, M. (2004). *Wordsmith Tools (Version 4).* Oxford: Oxford University Press.

Scott, M. (2011) WordSmith Tools Help. Liverpool: Lexical Analysis Software. http://www.lexically.net/downloads/version5/HTML/proc_tag_handling.htm (last accessed 04/03/2011)

Sidtis: Van Lancker Sidtis, D. (2009). Formulaic and novel language in a "dual process" model of language competence: Evidence from surveys, speech samples, and schemata. In R. Corrigan, E. A. Moravcsik, H. Ouali, & K. M. Wheatley (Eds.), *Formulaic language: acquisition, loss, psychological reality, function, Volume 2: Typological studies in language* (pp. 445-472). Amsterdam: John Benjamins.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. (2005). Corpus Design. In M. Wynne, (Ed.). *Developing linguistic corpora: A guide to good practice.* Oakville: Oxbow Books.

Sommers, N & Saltz, L. (2004). The novice as expert: Writing the freshman year. *College composition and communication, 56*(1), 49-124.

Tardy, C. & Swales, J. (2008). Text Organization, genre, coherence, and cohesion. Bazerman (Ed.) *Handbook of research on writing: History, society, school, individual, text.* New York: Taylor and Francis Group.

Tashakkori, A. and Teddlie, C. (Eds.). (2010). *Mixed methods in social & behavioral research: Second edition.* Thousand Oaks, California: Sage.

Wardle, E. (2007). Understanding "transfer" from FYC: Preliminary results of a longitudinal study. *Writing Program Administration, 31*(1/2), 65-85.

Weinert, R. (1995). The role of formulaic language in second language acquisition. *Applied Linguistics*, 16, 180-205.

Willis, D. & Willis, J. (2007). *Doing task-based teaching*. Oxford: Oxford University Press.

Wood, D. (2001). In search of fluency: What is it and how can we teach it? Canadian Modern Language Review, 57, 573–589.

Wood, D. (2002). Formulaic language in acquisition and production: implications for teaching. *TESL Canada Journal, 20* (1), 1-15.

Wood, D. (2010). Lexical bundles in an EAP textbook corpus. In D. Wood, (ed.), *Perspectives on Formulaic Language: Acquisition and Communication* (pp. 174-194). London: Continuum.

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

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

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. NY: Oxford University Press.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3,* 215–229.

Yorio, C. (1989). Idiomaticity as an indicator of second language proficiency. In K. Hyltenstam and L. Obler (eds), *Bilingualism Across the Lifespan Aspects of Acquisition, Maturity and Loss* (pp. 55-74). Cambridge: Cambridge University Press.