UNDERGRADUATE VS. GRADUATE ACADEMIC ENGLISH:
A CORPUS-BASED ANALYSIS

A Dissertation in
Applied Linguistics
by
Tracy S. Davis

© 2012 Tracy S. Davis

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2012
The dissertation of Tracy S. Davis was reviewed and approved* by the following:

Xiaofei Lu  
Gil Wats Early Career Professor in Language and Linguistics  
Dissertation Advisor  
Chair of Committee

Suresh Canagarajah  
Edwin Erle Sparks Professor in Applied Linguistics and English

Xiaoye You  
Associate Professor of English and Asian Studies

Robert Schrauf  
Professor of Applied Linguistics

Karen Johnson  
Professor of Applied Linguistics  
Director of Graduate Studies

*Signatures are on file in the Graduate School
ABSTRACT

While English as an Academic Language (EAP) has been the focus of much research, the differences between academic English at the undergraduate level and that at the graduate level have largely been ignored. The present study serves to fill that gap in the literature by a) creating two corpora of academic English at the two levels of study and b) comparing the vocabulary, lexical bundles, and grammar related to parts of speech (POS) found in each by emulating a series of corpus studies done on academic English in other contexts. Undergraduate academic English was found to have a more diverse vocabulary, a vocabulary that did not support teaching the Academic Word List (AWL) at the undergraduate level, less diversity in lexical bundles, a greater reliance on lexical bundles with different functions than those used at the graduate level, significantly different use of semantic types of nouns and verbs, significantly different use of personal pronouns, and significantly different use of voice and tense between disciplines at the two levels. While the results varied from nominal to dramatic variation between the two types of academic English, the study taken as a whole clearly demonstrates that academic English cannot be assumed to be the same regardless of level, and that these differences need to be addressed in EAP courses and potentially in undergraduate writing courses as well.
# TABLE OF CONTENTS

List of Figures ........................................................................................................................................ vii

List of Tables .......................................................................................................................................... viii

Acknowledgements .................................................................................................................................. ix

**Chapter 1 Introduction** ...................................................................................................................... 1

1.1 Comparison of Undergraduate and Graduate Programs ......................................................... 1
1.2 Research Questions .......................................................................................................................... 5
1.3 Structure of the Dissertation .......................................................................................................... 7

**Chapter 2 The Corpus** ....................................................................................................................... 9

2.1 Defining Academic, Undergraduate, and Graduate English ....................................................... 9
  2.1.1 Academic English by Context ................................................................................................. 12
  2.1.2 Academic English by Approach ............................................................................................. 13
  2.1.3 Academic English in this Study ............................................................................................. 17
2.2 Corpus Design ............................................................................................................................... 18
  2.2.1 Text Inclusion ......................................................................................................................... 18
  2.2.2 Discipline Selection ............................................................................................................... 20
  2.2.3 Size ....................................................................................................................................... 23
  2.2.4 Transcription and Coding of Text .......................................................................................... 25
  2.2.5 Tagging and Manipulating Text ............................................................................................. 27
2.3 Summary ...................................................................................................................................... 29

**Chapter 3 Vocabulary** ..................................................................................................................... 30

3.1 What Types are Frequent and What is their Range? ................................................................. 39
  3.1.1 Methodology .......................................................................................................................... 39
  3.1.2 Results .................................................................................................................................... 43
3.2 How Well Does the AWL Cover the AUGER Corpus? ............................................................... 49
  3.2.1 Methodology .......................................................................................................................... 49
  3.2.2 Results .................................................................................................................................... 51
3.3 The America-centric Use of Language ......................................................................................... 55
3.4 Summary ...................................................................................................................................... 58

**Chapter 4 Lexical Bundles** ............................................................................................................... 59

4.1 Lexical Bundles in Academic English ......................................................................................... 59
4.2 Analysis of Lexical Bundles ........................................................................................................... 64
  4.2.1 What is the Frequency across Corpora? .................................................................................. 65
LIST OF FIGURES

Figure 1.1 Requirements for Undergraduate Degrees ........................................3
Figure 3.1 Frequency of Lemmatized Types ...................................................44
Figure 3.2 Types by Discipline and Level .......................................................46
Figure 3.3 Normalized Type Token Ratio .......................................................47
Figure 5.1 Comparison of Content Word Classes by Level ..............................87
Figure 5.2 Nouns by Group .......................................................................89
Figure 5.3 Personal Pronouns .....................................................................94
Figure 5.4 Pronouns per million Words by Discipline Group .......................94
Figure 5.5 Percent of Tokens for Semantic Classes of Verbs by Level ..........99
Figure 5.6 Past Tense and Present Tense in the Soft-Pure Fields ...............103
Figure 5.7 Past Tense and Present Tense in the Hard-Pure Fields ...............104
Figure 5.8 Percent of Passive Voice by Discipline and Level .......................106
Table 5.1 Types of Nouns........................................................................................................................................85
Table 5.2 Frequency of Nouns by Level.............................................................................................................88
Table 5.3 Most Frequent Abstract/Process Nouns by Level...............................................................................92
Table 5.4 Verbs per million Words by Semantic Class .....................................................................................98
Table 5.5 Semantic Classes of Frequent Verbs.................................................................................................100
ACKNOWLEDGEMENTS

There are a number of people I must thank for their direct or indirect roles in this dissertation. First and foremost I am extremely grateful for my advisor, Xiaofei Lu, who provided much feedback and direction throughout the thesis process and whose practical advice I will remember for years to come. I would also like to acknowledge my committee, Robert Schrauf, Suresh Canagarajah, and Xiaoye You. Their comments and questions helped me to strengthen my thesis and understand its limitations.

I would also like to thank a number of people whose assistance in completing this project were less obvious. Most importantly, I want to thank my parents, Elaine Davis and Michael and Joanne Davis, for their continuous love and support and for all the influences they have had in my life that lead me to this point. I must also thank Sunita Dillon, Priscilla Ortiz, Janice MacGregor, and Cathy Hicks-Kennard, all of whom provided exactly the kind of support I needed at the right times. Many thanks to everyone for making this a reality.
This dissertation is dedicated to my grandfather, Dr. William O. Davis.

I may never have had the privilege of meeting you, but you originally inspired me to become the next Dr. Davis.
Chapter 1

Introduction

English for Academic Purposes (EAP) is an area of burgeoning research. Part of the interest in EAP stems from a growth in the need for EAP instruction. In the academic year 1996-1997 there were 457,984 foreign students studying in the U.S. and 198,064 in the UK (Flowerdew & Peacock, 2001). In 2010 the number of international students studying at American Universities had increased to 723,277 (McMurtrie, 2011). At this institution there are over 5,000 international students, roughly half of whom are enrolled in graduate studies (Pennsylvania State University, 2012).

Furthermore, as the per capita income of households in developing countries rises, the ability of these countries to satisfy the tertiary educational needs becomes taxed, thus creating more and more students from these countries going abroad to study, primarily to the UK, US, and Australia (Graddol, 1997). Given these circumstances, students are leaving their home counties to study at both the undergraduate and graduate levels.

1.1 Comparison on undergraduate and graduate programs

It is intuitive that the English demands placed on a graduate student are not the same as those placed on an undergraduate (see Table 1.1 for a comparison). Indeed, entrance exams at the undergraduate level include the Scholastic Aptitude/Assessment Test (SAT) and the American College Test (ACT), while the graduate level requires the Graduate Record Examination (GRE) or Law School Admissions Test (LSAT). Yet, at both levels international students for whom English is not their native language must take
the same Test of English as a Foreign Language (TOEFL) and International English Language Testing System (IELTS) exams to prove English proficiency.

Table 1.1: Comparison of requirements at the undergraduate and graduate levels based on requirements at a large, public university in the eastern United States.

<table>
<thead>
<tr>
<th></th>
<th>Undergraduate</th>
<th>Graduate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entrance Exams</td>
<td>SAT, ACT</td>
<td>GRE, LSAT</td>
</tr>
<tr>
<td>Required English Test</td>
<td>TOEFL*, IELTS*</td>
<td>TOEFL*, IELTS*</td>
</tr>
<tr>
<td>Focus</td>
<td>One or more majors, optional minor, 42 credits of general education classes</td>
<td>One discipline, with some exceptions</td>
</tr>
<tr>
<td>Writing Expectations</td>
<td>Begin with short papers, around 1,000 words or less. References may or may not be required. Papers in the final semesters may be much longer.</td>
<td>Papers are expected to be researched, always include a set of references. Length depends on the discipline, but are usually 2,000 – 6,000 words. Often a thesis or dissertation is required to graduate</td>
</tr>
<tr>
<td>Required Reading^</td>
<td>Primarily textbooks with some lab books, course packets, and mass media.</td>
<td>A mixture of journal articles, monographs, reference books, and textbooks. Required reading varies based on discipline.</td>
</tr>
</tbody>
</table>

* Both the TOEFL and the IELTS were created for high school seniors preparing to attend undergraduate study.

^ Required reading determined based on a study of syllabuses (see below). Different disciplines sometimes require different types of reading, although this is more prevalent at the graduate level.

While it may be a logistical issue (e.g. class size, number of instructors, budget) that many institutions that offer EAP courses tend to combine both future graduate students and future undergraduate students into the same classes, certainly having the same exam and similar final expectations for these students contributes to the justification of these combined classes. Consider a search of English language institutes that offer instruction in academic English. In an informal look at the first thirty institutions that
appear in a Google search that I conducted in December 2009, only two had any differentiation between graduate and undergraduate academic English: one offered graduate specific courses in addition to general academic English courses, and the other noted that its program is not for graduate students. Several of the institutes offered TOEFL preparation courses in conjunction with the other classes.

I would argue that this mono-level approach to academic English is hardly ideal. Expectation alone helps to under-gird this idea. Graduate students are expected to have a certain level of knowledge within their field before they begin coursework in a program as a prerequisite for admission. The classes they will take will typically be well within the same discipline; for example, the required courses for both a masters degree and a doctorate in physics at one American university are all physics courses (Pennsylvania State University, 2012).

Figure 1.1: Requirements for undergraduate degrees in two disciplines
based on data from Pennsylvania State University (2012)

BA in Sociology Course Requirements

BS in Physics Course Requirements
In contrast, the average undergraduate at the same American university has to take approximately one third, or 45 credits, of his or her classes in areas outside the major (known as general education credits at some universities). These classes include nine credits each in English (speaking and writing) and the natural sciences, six each in quantification and mathematics, the arts, humanities, and social and behavioral sciences, and three in health and physical activity (Pennsylvania State University, 2012). Some of these courses may overlap with some of their major courses, but given the diversity it is rare to see more than a handful of credits do so. Indeed, it is normal for an undergraduate student to take fewer than half of his or her classes within a major field of study, as can be seen in Figure 1.1.

Graduate students must also have a certain level of writing expertise in order to achieve their goals within their field. That is to say, while graduate students are rarely expert writers, they typically are familiar with the basics of writing within in their chosen discipline. Moreover, with the exception of cross-disciplinary studies, these students are not expected to have to write for different disciplines; the focus is squarely on the field of study. On the other hand, undergraduate students enter the American university as novices and are required to take at least one and usually two courses on how to write an essay for class. The students, including native English speakers, are expected to learn the language of the academy as part of their college education. Given the diversity of the courses undergraduate students are taking, they may also be required to write essays in several different, unrelated disciplines.

A final important contrast between these two groups of students is that, while graduate students tend to be fixed in a course of study, undergraduate students can
actually begin university study with no declared major and once they have declared a major, they can and, often do, switch majors. While exact numbers are hard to find, it is estimated that 80% of college freshmen have not declared a major, and over half of the ones that do change their major do so one to three times (Ronan, 2005). Another study indicated that roughly half of the students at a university changed majors at least once before graduation (Micceri, 2001). Specific numbers could not be found for international students, but it should be noted that while student visas do require a course of study to be listed, changing that course of study is not grounds for revoking the visa and simply requires that the proper paperwork be filed.

1.2 Research Questions

Despite these very clear differences in student experience based on level, for some reason research in EAP, when considering variation within, has considered only disciplines, genres, and registers, including spoken and written, and has not compared the differences between graduate and undergraduate academic English. Previous research has tended to combine textbooks, student writing, articles, and other writing from both levels together (e.g. Hyland and Tse, 2007; Simpson et al., 2002; Coxhead, 2000). Being so inclusive can allow for a broader understanding of general academic English, but if the target is to inform EAP pedagogy then the corpora need to be more focused. Other research has looked exclusively at journal articles (e.g. Hyland, 2008; Harwood, 2005) as a source of academic English. Again, journal articles are an important part of the academy, but they do not address the needs of the average undergraduate student at an American university. Some research has divided the two levels (see Biber, 2006) but failed to compare and contrast. It is the purpose of this dissertation to fill this gap in
previous research by doing a corpus-based analysis of undergraduate and graduate level academic English.

The primary question is, what, if at all, is the difference between undergraduate academic English and graduate academic English? There are many ways in which to the levels could be differentiated, but this dissertation will address only five research questions in three overarching areas: vocabulary, lexical bundles, and general grammar.

Vocabulary is a building block of language. It also lends itself readily to corpus-based studies. The research in this dissertation will begin here with the following questions:

1. What types are frequent and what is their range?
2. How many words are high frequency and how many are low frequency?
3. What is the range of words used in the different disciplines and levels, and how do they compare?
4. What are the most frequent types in the two different levels, and do they overlap?
5. If the Academic Word List (AWL) worked well for a general corpus of academic English, how would it fare on a corpus specifically geared toward undergraduate input?

When three or more words co-occur repeatedly, we refer to them as n-grams or lexical bundles. Lexical bundles are the logical next step in analysis following vocabulary. The questions addressed in this section will include:

1. What is the frequency of lexical bundles in general across corpora?
2. Which lexical bundles are frequent at the graduate and undergraduate levels?
3. Which disciplines use lexical bundles frequently?
4. What are the most frequent lexical bundles in disciplines and how do they compare between levels?

5. What is the function of these lexical bundles?

Finally, we will focus on grammar in a broad sense as grammar itself is a fairly broad area. Specifically, we will examine the content word classes for both frequency and function. The questions that will be addressed are the following:

1. What are the most frequent word classes at the two levels?
2. What types of nouns are used frequently at the two levels?
3. What types of verbs are used at the two levels?
4. How are personal pronouns used at the graduate and undergraduate levels?
5. How might verb tense, aspect, and voice vary between graduate and undergraduate academic English?

1.3 Structure of the Dissertation

This dissertation strays from what has become the standard format for a thesis. As was laid out in section 1.2, there are three main areas of analysis. Including an overarching literature review in the second chapter would potentially position the relevant literature far from the methodology and results. As such, the dissertation has been organized to keep information more or less contained within the same chapters.

The next chapter details the corpus used to make these comparisons. The American Undergraduate and Graduate Expected Readings (AUGER) corpus contains 2.3 million words taken from texts that, according to course syllabuses, would be assigned to students at the graduate and undergraduate level. The chapter describes
previous corpora used to study academic English before explaining the process of building the AUGER corpus itself and providing a description.

The third chapter focuses specifically on the vocabulary portion of the analysis. I begin with a literature review that has informed my research questions and my methods of analysis. It is here that I will discuss type/token ratio and determine what types are frequent and if they are frequent across disciplines. I also reconsider previous research on the Academic Word List (AWL) in light of graduate and undergraduate expectations.

In the fourth chapter I analyze lexical bundles, or n-grams, words that frequently co-occur together. Again, I begin the chapter with a literature review before addressing the individual research questions. While I briefly look at three and five word lexical bundles because they do offer some insight into variation between the two levels, the majority of the analysis focuses on four word lexical bundles, following precedent set by previous research (e.g. Cortes, 2004). As with the analyses in the vocabulary chapter, I consider frequency and range and briefly discuss function.

The fifth and final analysis chapter considers grammar, with a specific attention on the parts of speech and verb phrases, including tense, aspect, and voice. Attention is paid to the function of nouns and verbs and the use of pronouns at the two levels. There are also some hypotheses offered for the dramatic differences that are found.

The last chapter synthesizes these various studies and discusses what, together, the results mean. It includes a section on pedagogical implications, limitations within the studies, and future directions for research.
Chapter 2
The Corpus

The underlying premise of this dissertation is a comparative study of undergraduate and graduate English, but how does one define these Englishes? And once these Englishes have been defined, how does one compare them? The purposes of this chapter are to:

1. Define undergraduate and graduate English, as they relate to academic English;
2. Use this definition to design two representative corpora;
3. And describe the corpus that is ultimately used to complete the research in this dissertation.

2.1 Defining Academic, Undergraduate, and Graduate English for the Corpus

While academic English has been the focus of corpus studies for over twenty years, little continuity can be found regarding what texts actually comprise “academic English.” While all of the corpora include texts found in a university setting, some include only peer-reviewed journal articles while others include monographs or textbooks, and still others include student writing. The Longman Grammar of Spoken and Written English (LGSWE) (Biber, Johansson, Leech, Conrad, & Finegan, 1999) consists of academic journal articles and extracts from academic books, of which only a small margin were written for "a lay audience, including student textbooks" (p. 33) in the portion of the corpus dedicated to academic prose. Coxhead (2000) includes journal articles, textbooks, lab manuals, web-based journal articles, monographs, and portions of
the "learned and scientific" sections of previously constructed corpora in her attempt to create a balanced corpus of academic written English across disciplines. Biber (2006) created the TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL) which includes textbooks, academic program brochures, university catalogs, student handbooks, course packs, and course management texts like syllabuses, as well as spoken texts from classroom teaching, office hours, classroom management, study groups, lab and in-class groups and service encounters. The Michigan Corpus of Academic Spoken English (Simpson, Briggs, Ovens, & Swales, 2002) contains the same spoken components but also includes meetings and symposiums.

However, some academic corpora include a learner or student component. The Hyland Corpus (Hyland & Tse, 2007) includes research articles and book reviews along with Ph.D. dissertations and Masters theses. The British Academic Written English Corpus (BAWE) (Nesi, Gardener, Forsyth, Hindle, Wickens, & Ebeling, 2005) is composed entirely of upper-level student writing, as is the MiCUSP corpus (Michigan Corpus of Upper-level Student Papers, 2009), and the Hong Kong University of Science and Technology Learner Corpus (HKUST) (Milton & Chowdhury, 1994) contains essays from university and upper-secondary school level non-native speakers.

So academic English, while a topic of discussion for several decades, remains a somewhat nebulous concept. Bhatia (2002) notes that principled investigation into what "academic discourse" has been fairly sparse. He writes:

"In principle, the use of the term 'academic English' presupposes the existence or at least an understanding of what might be called an 'academic core' underlying most of the discourse types used in the academy; however, in practice, the
existence of such an academic core is often assumed, rather than investigated or established." (p. 26)

Strevens (1988) sets out four defining characteristics of EAP, one of which being that it is in contrast to "general English." He also states that it needs to meet the learners’ individual needs, the content should be related to the specific occupations, disciplines, or activities of the learner, and that the focus is on the language form. While many of Strevens' points are quite valid, they follow the notion that language is a "transparent and autonomous system" (Lillis, 2001, p. 34). English is being treated as a skill rather than a social construct.

The continued study of EAP may equally be attributed to changes in theoretical thinking that shift away from this idea. The New London Group (1996) explains that "there [is] not a singular, canonical English that could or should be taught anymore" (p. 3), and that in fact, there are multiple literacies which need to be taught. That is to say, the English used in one situation is actually quite different from the English used in another situation, both spoken and written. To be multiliterate means that one is capable of using the Englishes, or discourses, of discourse communities or for different situations.

This view of English highlights two distinct issues in EAP. First, students attending a university need to be literate in a particular kind of English in order to be successful; in many ways the type of English is determined by the discipline. However, it also underscores that crucially undergraduate students need more than to be literate in the English of their chosen field when they attend an American university. Not only do they take coursework outside of that field, but there is an expectation that they will participate in non-academic activities as well (e.g. student government, student run
organizations, Greek life). These same non-academic activities are marginalized to an extent for graduate students, but regardless they have lessor need to interact outside of their area of interest.

2.1.1 Academic English by Context

There have been several contextual definitions provided for EAP, the unifying factor being that all insist that it is connected to language that students use in academic settings. Some studies limit EAP to study purposes (ETIC, 1975; Dudley-Evans & St. John, 1998), while some include other situations in an academic context, but still generally limiting it to tertiary settings (Flowerdew & Peacock, 2001; Hyland, 2006), and others widen the definition to include all aspects of school life, including pre-tertiary education (Baily & Butler, 2002).

Dudley-Evans and St. John (1998) set out four EAP contexts for students, the first being study in a country where English is an official language (e.g. The United States, Great Britain, or Australia) and another where English is the language of the university (e.g. Hong Kong). Previous corpus based studies of academic English have generally used context when building corpora of student work (Nesi et al., 2005; Milton & Chowdhury, 1994; Michigan Corpus of Upper-level Student Papers, 2009) at least in part because of selection; all of the students belonged to a single university setting. Coxhead (2000) organized her corpus based on traditional New Zealand university structure, but did not purposely include texts from those universities. Biber (2006) used the structure of his university to design his Y2KSWAL corpus, but he did not include texts exclusively from his university. However, context does play an important role when preparing students for university study and will be addressed in this research. The context falls into
Dudley-Evans and St. John’s (1998) first category, but even here, as was discussed in the previous chapter and will be revisited in the next section, we do not see continuity.

2.1.2 Academic English by Approach

There are two approaches to EAP: English for General Academic Purposes (EGAP) and English for Specific Academic Purposes (ESAP). English for General Academic Purposes focuses on the skills and language common across disciplines such as note taking, listening to lectures, writing essays and examination answers, and other common tasks in the school setting. For example, Dudley-Evans and St. John (1998) explain that reading a textbook requires the same skills across disciplines: one must understand the main idea presented, recognize the supporting details, distinguish the author’s point of view, take notes, and be able to skim and scan for general and specific information. Quite often EGAP is referred to as teaching a "common core" which goes beyond basic student and study skills to include giving oral presentations, expressing cause and effect relations, presenting results, and other tasks thought to be common across most disciplines.

On the other hand, English for Specific Academic Purposes focuses on the differences between the disciplines. Dudley-Evans and St. John see ESAP in a developmental role, a class students would take after completing an EGAP course that would integrate the general skills learned previously with the more precise requirements of their discipline. ESAP classes vary from discipline to discipline, but the focus would be the more exact language and writing that one would be expected to produce in the field of study. There are arguments for and against both approaches.
Spack (1988) argues that discipline based academic writing should be taught by teachers in those disciplines and that English teachers, L1 or L2, should teach general academic writing. She explains that while an English teacher may be quite competent in the instruction and assessment of form, without control of the content, the teacher is incapable of giving constructive feedback on, for example, ways to show full understanding of the knowledge supposedly being shown in the paper. Moreover, a discipline may have various subdisciplines which are vastly different from each other, as with anthropology wherein an article in physical anthropology may be quite similar to those of natural scientists and an article in cultural anthropology may be more like those of literary scholars (Faigley & Hansen, 1985, cited in Spack, 1988).

Hyland (2006) counters Spack's argument saying teachers entrenched in their disciplines lack the motivation and, more importantly, the skills to teach EAP courses. These teachers have been trained in the subject matter, but not in the language used to express it, and that is what the focus of an EAP class should be. To make an analogy, it is similar to having any native speaker teach a language class because clearly they know the language, even if they have had no training in language teaching. It is possible that some teachers will excel, or that some students will garner a great deal from the course, but it is equally, and perhaps more likely that both the teacher and the students are being set up for failure.

Agreeing with Hyland in principle, Raimes (1991) adds that moving EAP teachers to an auxiliary role within departments devalues the profession and gives universities leeway to marginalize ESOL instructors. However, if Hyland (2006) is correct, then
clearly these ESOL instructors are needed because discipline based instructors are not capable of the task.

Bhatia (2002) notes that universities (no specific country is given) have begun to adapt academic programs to include more interdisciplinary practices as a result of changing workplace demands including a need for multiple specialties. American undergraduate programs have also largely adopted multidisciplinary requirements in an attempt to give a more well-rounded education, requiring students to take courses in the humanities, social sciences, natural sciences and mathematics regardless of the students’ actual major. Students now face the need to not only be literate in their field of study, but in other fields as well.

Another argument for EGAP is that courses in this vein are more convenient for practitioners (Bhatia, 2002). Without the need for excessive collaboration with other ESOL instructors, instructors in the target disciplines, and administration, it takes less time to design a syllabus and plan courses, particularly if the course contains students from multiple disciplines.

However, according to Hyland (2000) EGAP courses can be detrimental to the students that take them. When "common core" elements are presented to students who lack the sufficient analytical and English skills to recognize disciplinary differences, it serves to "disguise" the variation and "misrepresents" academic discourses as "naturalised, self-evident and non-contestable ways of participating in academic communities" (p. 147). It should also be noted that some of the practices and linguistic structures thought to be in the "common core" are disputed. For instance, citation
practices in the pure sciences differ greatly from those in the social sciences (Hyland, 2002).

Widdowson (1983) makes the distinction between the two with the analogy that EGAP is more like education and ESAP is more like training, a lesser activity than education itself. Dudley-Evans and St. John (1998) do not agree with this perspective, believing instead that both approaches to EAP are education oriented. Hyland (2006) takes the perspective that, while there are some generalizable skills for EGAP, it is a waste of time.

Some of the variation in perspective on EGAP and ESAP may come from the differences in university structure. In Great Britain, where a large amount of EAP research has taken place, it is not commonplace to require undergraduate students to take courses outside of their respective disciplines. Undergraduate studies often take three years, with the courses for each year clearly mapped out. Moreover, the average student enrollment at British universities is around 15,000 at a single institution (Higher Education Statistics Agency, 2007). However, in the North America most universities have adopted the practice of requiring all undergraduate students to complete a number of credits in general education that range across the disciplines. At some universities these credits make up a third or more of the total credits required to graduate. Not surprisingly, given the difference in populations between Great Britain and the United States, universities tend to be larger, which makes coordination between departments and stricter structure of courses unreasonable. With this in mind, it makes sense that there would be a stronger drive for ESAP courses in Great Britain while EGAP courses have a more compelling argument on the other side of the ocean.
In any case, a clear definition of academic English is never outlined. Instead, it is understood that context plays a decided role in describing academic discourse. The overarching principle is that it is English used in an educational or scholarly context, but it is influenced by the level (primary schools, secondary schools, higher education), linguistic context (English as the native language, English as an auxiliary language, English as the language of the university), discipline (sciences, social sciences, humanities), and location (expectations of a university in a particular country).

2.1.3 Academic English in this Study

For the purpose of this study, the academic English at American universities, where English is the national language, will be the focus. The corpus will be comprised of published written English required in classrooms at a large, public university on the East coast of the United States and will cross a range of disciplines. The reasons for focusing only on required readings are two-fold. First, these are readings that students, in theory, must be able to read to succeed in a course. Indeed, some students (primarily native-speaking) believe that if they complete all the required reading they do not necessarily have to even attend classes to do well. While the veracity of this theory likely varies based on professors, topics, and the students themselves, professors do assign reading under the auspices that it will assist their students in learning the subject matter. Secondly, these texts are level specific. Institutional materials do not necessarily focus on graduate or undergraduate students, or even students at all, but course readings are specific to the course itself and the students who are enrolled in it.

Necessarily, the findings will be applicable to university settings in the United States, and less so to settings in other native English speaking countries. It has been
acknowledged that there is a limitation as universities in these other countries have different requirements for students, use different materials in their classes, and arguably use a different standard dialect. Referring to the English in this study as “published written American university English” is a mouthful and hereafter will simply be referred to as written academic English with the understanding that it does not encompass written modes and genres from all possible contexts at the micro-level (e.g. disciplines, classes) and the macro-level (e.g. international contexts).

### 2.2 Corpus Design

Because this is a comparative study, we will need two corpora with a parallel structure, one representing undergraduate academic English and one representing graduate academic English. In this section I will outline how these two corpora were designed. When designing a corpus, one must consider several factors including both representativeness and balance. The research conducted with a corpus is only as good as the corpus itself. Representativeness of a corpus means that the corpus fits the function for which it was designed. Factors that affect representativeness include what texts are selected and how and text size. On the other hand, balance alludes to the number of different types of texts (Sinclair, 2004; Kennedy, 1998).

#### 2.2.1 Text Inclusion

The focus of this study is to compare undergraduate and graduate academic English. As discussed earlier, academic English has been represented in corpora in different ways with the common theme being that the English used is found in an academic community. However, for the purpose of this study, some contexts for texts can be excluded.
Biber (2006) found that while there are words that are used only in oral classroom teaching, the number of words is very small and often includes everyday words (e.g., bagel, nail) or words textbooks would be unlikely to include (e.g., cop, dude). Textbooks, on the other hand, contain several words that were not found in the spoken discourse. Because this study primarily compares lexical features between graduate and undergraduate English, this corpus will include only written texts.

The question of which texts to include is solved by limiting the context. Professors require certain texts in their classes because they believe they are important to the course or to the education of the student. In theory, the content of the course (e.g., lectures, assignments, and exams) is largely connected to these required texts. It is also a factor for which one can easily control. While students may in fact use other resources in their studies (e.g., professional notes, websites, reference books), the required texts are the universal factor for both students and professors.

Hyland (1999, 2000, 2001) used discipline informants to nominate "leading journals" from which to extract the first three articles. For each of the eight disciplines, three articles from ten journals are included, for a total of 240 published articles. Harwood (2005) also used discipline informants. The “discipline informants” for this corpus are the syllabuses presented to students at the beginning of each semester and course websites, which include any required textbooks and often list additional required reading in the form of journal articles, websites, or manuals.

The syllabuses were primarily acquired from the course management software supported by the university. Syllabuses were collected from two levels: 0-200 and 500+.

Courses at the 0-200 level are beginning or introductory classes, undergraduate, and
generally taken in the first two years at the university. Students in these classes may or may not be taking them as part of their major but instead for general education credit, as required at many US universities. Syllabuses at this level significantly indicate that textbooks are the only required text, if any, thus the texts in this level exclusively include textbook chapters from the required textbooks.

Courses at the 500+ level are graduate classes either for masters or doctorate study. Students in these classes will probably only be enrolled in courses that assist in their area of study. Syllabuses at this level indicate that textbooks, monographs and external readings are equally required. Chapters from the textbooks and monographs as well as whole journal articles are included in the graduate corpus.

2.2.2 Discipline Selection

Corpus studies that have considered discipline variation in the past have taken a number of approaches for subdivision. For example, Coxhead (2000) chose four subdivisions (Arts, Commerce, Law, and Science) following “the traditional degree divisions in tertiary institutions in New Zealand, and internationally” (Coxhead’s unpublished master’s thesis, p. 35). The “internationally” portion of her justification is arguable. Law, for example, in American universities, is a single undergraduate major or a separate institution for post-undergraduate study, similar to medical school, a division not included in Coxhead’s corpus.

However, in many ways it does make sense to follow the natural division of the university. Biber (2006) and his colleagues use the natural division of an American university for their T2K-SWAL corpus, considering business, education, engineering, humanities, natural science, and social science. Hyland and Tse (2007) also used a
similar structure but included fewer areas limiting their study to engineering, sciences, and social sciences. The problem with such a division is that some disciplines straddle colleges; economics might fight into social sciences, the humanities, or business, applied linguistics could be seen in education, the humanities, or social science, while statistics might be engineering, natural sciences, or even social sciences. These fine lines between disciplines, particularly with a current trend towards interdisciplinary studies, will continue to be an issue.

Hyland (2000, 2001) used a different division for his Hyland Corpus (as it is referred to in Krishnamurthy & Ksoem, 2007). He included only eight specific disciplines determined by contrastive themes of pure science (physics and microbiology), applied science (mechanical engineering and electrical engineering), social science (marketing and philosophy) and applied social science (sociology and applied linguistics). While some might take exception to philosophy being a social science instead of a cornerstone of the humanities, his reasoning was for a comparative study of "pure" and "applied" fields.

Harwood (2005) and Hyland (2000, 2001) adapted a taxonomy of disciplines created by Becher (1989, 2001). Becher's taxonomy, on the cognitive dimension, placed disciplines on a continuum between “hard” and “soft” fields and “pure” and “applied” fields, terminology he borrowed from Biglan’s own taxonomy of disciplines (1973a, 1973b). Biglan based his taxonomy on faculty perception of their fields and divided disciplines along three contrasting dimensions: hard-soft, pure-applied, life-nonlife. Becher (1994) compared Biglan’s taxonomy with Kolb (1981), who looked at student learning strategies and applied the Kolb Learning Style Inventory (LSI) to determine how
disciplines should be divided along the abstract-concrete and the active-reflective
dimensions. When Becher ignored Biglan’s life-nonlife dimension, he found that the two
sets of quadrants are significantly similar. Indeed, Becher (1994) noted that “the
coincidence of their analyses is significant, given that Biglan’s initial concern was with
the nature of the subject-matter of research, while Kolb’s was with styles of intellectual
enquiry” (p. 152).

While Becher’s taxonomy is a continuum, Biglan-Kolb adopt a more concrete
quadrant taxonomy, which is ideal for determining which disciplines belong in which
category. It is this taxonomy which has been adopted for the corpus in this study.
Biglan’s terminology has also been adopted, creating two subcategories: hard-pure
(natural sciences and mathematics) and soft-pure (social science and the humanities),
further delineated in Table 2.1.

<table>
<thead>
<tr>
<th>Table 2.1 Disciplines in each subcategory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard-Pure</td>
</tr>
<tr>
<td>Biology</td>
</tr>
<tr>
<td>Chemistry</td>
</tr>
<tr>
<td>Earth Sciences</td>
</tr>
<tr>
<td>Mathematics</td>
</tr>
<tr>
<td>Physics</td>
</tr>
</tbody>
</table>

Not surprisingly, there are very few courses in the hard-applied (science based
professions) and soft-applied (social professions) categories offered in the first two years
of undergraduate study, which is easily explained by the fact that many of the applied
fields (e.g. engineering, medicine, education, social work, business) either require lower
level courses before one can take these more specific courses (e.g. physics and math for
engineering) or are major specific and do not carry general interest. Thus, the hard-
applied and soft-applied disciplines have been excluded as it would be very difficult to
include these disciplines in the undergraduate corpus and the two corpora (undergraduate and graduate) should be parallel.

2.2.3 Size

Coxhead (2000) was careful to include short, medium, and long length texts and to balance them across seven subject areas in each of the four disciplines. She generally used full texts, but also included portions of the learned and scientific sections of previously created corpora (the London/Oslo/Bergen [LOB], Brown, MicroConcord, and Wellington Corpus of Written English). These corpora use 2000 word text segments, so are not full texts. Hyland and Tse (2007) took exception to the use of the text segments from other corpora because a) they are dated (the Brown and LOB corpora were created from texts in 1961) and b) text segments do not show the variation in language that occurs in a full text. Evidence of intratext variation has been shown in research articles through IMRD (introduction, materials, results, discussion) studies (e.g. Swales, 1990; Biber & Finegan, 2001). Hyland and Tse's own corpus uses full texts and normalized the data as each discipline did not have the same number of words. For each of their three disciplines they collected an even number of textbook chapters, journal articles, academic book reviews, master’s theses, doctoral dissertations, final year project theses, and, in the case of their sciences subcorpus, scientific letters, thus including long and short full texts. Their justification is that the corpus should represent the “range of sources students are often asked to read at a university and so include the kinds of lexical items they will frequently encounter” (p. 238).

The corpus for this dissertation also uses full texts, or complete parts of texts (e.g. a chapter from a textbook), and normalizes for differences in length. An explanation of
“full text” needs to be included here. Titles, by-lines, references sections, footnotes, endnotes, charts, graphs, pictures, indices, appendices, formulae, side bars, and added sections are not considered the “main” text and were not included in the corpus as they can skew the results of an analysis, or in the case of non-ASCII elements, are simply not readable in plain text format, the format used for analysis. This is not to say that these added pieces of text are unimportant, and they may be included in a future analysis as there are .pdf files of each text for reference. However, this particular study is not a multimodal analysis and the point is to focus on the “meat” of the information and the language in which it is written.

The size of the corpus must be large enough to draw a reliable conclusion from the results. With respect to grammar and general structure, a relatively small corpus is adequate; however, a much larger corpus is required for the analysis of vocabulary. Unfortunately, there is the added restraint of time: a single researcher with a limited amount of time in which to construct a large enough corpus to make reliable claims. A corpus of one million words is fairly reliable with respect to grammar, structure, and even frequent vocabulary. However, information about “sense and association” (p. 30), or context, may be less so (Biber, Conrad, & Reppen, 1998). Because the focus is on words and lexical bundles that occur frequently, a corpus of one million words at both the undergraduate and graduate level should be sufficient to at the very least determine if this research is worth continued pursuit.

With that in mind, the corpus itself contains ~2.3 million tokens, 1.15 million undergraduate tokens, and 1.16 million graduate tokens. The goal was to have 500,000 tokens for each section (undergraduate hard-pure, undergraduate soft-pure, graduate
hard-pure, graduate soft-pure), and 100,000 tokens for each discipline, with roughly the same number of texts. In order to maintain balance as well as representativeness, it was important to include similar number of texts to account for subject and author variation.

To achieve these goals, each discipline had at least 13 texts, but no more than 15 texts, as this kept the variation in the number of texts small, while still allowing disciplines to come close but not excessively over exceed the 100,000 token goal. The smallest number of tokens in a discipline was 95,445 (undergraduate math, 15 texts), and the largest number of tokens in a discipline was 153,029 (undergraduate history, 13 texts).

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Undergraduate</th>
<th></th>
<th>Graduate</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># of texts</td>
<td># of tokens</td>
<td># of texts</td>
<td># of tokens</td>
</tr>
<tr>
<td>Hard-Pure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td>13</td>
<td>100,612</td>
<td>15</td>
<td>97,697</td>
</tr>
<tr>
<td>Chemistry</td>
<td>13</td>
<td>96,318</td>
<td>15</td>
<td>101,399</td>
</tr>
<tr>
<td>Earth Science</td>
<td>13</td>
<td>102,043</td>
<td>15</td>
<td>111,840</td>
</tr>
<tr>
<td>Mathematics</td>
<td>15</td>
<td>95,445</td>
<td>15</td>
<td>96,639</td>
</tr>
<tr>
<td>Physics</td>
<td>13</td>
<td>97,236</td>
<td>15</td>
<td>103,607</td>
</tr>
<tr>
<td>Total</td>
<td>67</td>
<td>491,654</td>
<td>75</td>
<td>511,182</td>
</tr>
<tr>
<td>Soft-Pure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geography</td>
<td>13</td>
<td>116,179</td>
<td>13</td>
<td>128,139</td>
</tr>
<tr>
<td>History</td>
<td>13</td>
<td>153,029</td>
<td>13</td>
<td>146,214</td>
</tr>
<tr>
<td>Philosophy</td>
<td>13</td>
<td>116,179</td>
<td>13</td>
<td>145,354</td>
</tr>
<tr>
<td>Political Science</td>
<td>13</td>
<td>134,201</td>
<td>13</td>
<td>114,625</td>
</tr>
<tr>
<td>Sociology</td>
<td>13</td>
<td>138,418</td>
<td>13</td>
<td>114,262</td>
</tr>
<tr>
<td>Total</td>
<td>65</td>
<td>658,006</td>
<td>65</td>
<td>648,594</td>
</tr>
<tr>
<td>Corpus Totals</td>
<td>132</td>
<td>1,149,660</td>
<td>140</td>
<td>1,159,776</td>
</tr>
</tbody>
</table>

2.2.4 Transcription and Coding of Text

Texts needed to be converted to .txt (i.e. plain text) format for various software tools to analyze them. This process was done in one of two ways, depending on the source of the text. If the text was in hardcopy (e.g. a textbook), the pages were scanned in and then converted to .txt using optical character recognition (OCR) software, in this
case both Adobe Pro and Omnipage Pro 16. Texts that were in electronic format but were not easily converted to text files (e.g. .pdf files that have not been converted from word processing files) were also scanned by the OCR software. The OCR software converts images to text, but it is not perfect and several factors play a role in its accuracy including font and the number of symbols imbedded in the text. Thus, after texts were scanned and converted, they were “cleaned” of obvious errors, misspellings, and imbedded numbers. This process was particularly labor and time intensive and was the primary reason for the smaller size of the corpus. Texts that were in electronic form and readable were simply converted to .txt files.

All files were further processed by removing headers, footers, titles, images, tables, graphs, charts and other visuals. Reference sections were not included in the .txt files. Equations, formulae, and logic statements were removed and replaced with \(<\text{statement}>\) as all three were found to be embedded in text but not easily read in .txt. Text must be ASCII printable characters to be read and manipulated by the programs used for analysis. Moreover, converting the statements to text would mean determining a standard way to write the statements which could in fact skew frequency counts. For example, \(x^2\) could be x squared, x to the second power, x to the power of 2, x to the power of two, x times x, and so on. Symbols (e.g. \(\£\), \(\Delta\), \(\Sigma\), \(<\), \(\beta\)) were also converted to \(<\text{symbol}>\) for similar reasons. The \(<>\) characters allow software programs (e.g. Antconc) to ignore these characters if desired, or to specifically target them in analysis.

The .txt files were saved using a transparent and expandable file name system: u/g (undergraduate or graduate) + h/s (hard or soft) + p (pure) + 1/2/3/4/5 (corresponding to the field in each category in alphabetical order) + 2 digit number of the file. Hence,
uhp101 would be an undergraduate, hard, pure file specifically in biology, and it is the first file.

The system allows for expansion should the corpus be used in future studies. Each field has room for 99 files, leaving the 00 distinction for concatenated files. There is also room to add applied fields from the taxonomy in future studies or up to four more disciplines in the hard-pure and soft-pure fields.

2.2.5 Tagging and Manipulating Text

The corpus was part of speech (POS) tagged using the Stanford Log-Linear POS Tagger v.3.0.1 (Toutanova, Klein, Manning, & Singer, 2003), an automatic text tagger which reports 97% accuracy when tagging a corpus of Wall Street Journal articles. The Stanford tagger uses the Penn Treebank tag set, consisting of 36 tags for punctuation and parts of speech. Specifically, it delineates between nouns, verbs, modals, adjectives, adverbs, pronouns, particles, possessives, prepositions, determiners, and coordinating conjunctions (see the Appendix for a complete list). While the tag set is not as robust as some (e.g. the CLAWS tag sets which vary between 60 and 180 tags [Garside & Smith, 1997]), it distinguishes the features that are most important to this study in a simple and clear fashion. For instance, nouns have four distinctions: NN for singular or mass nouns, NNS for plural nouns, NNP for proper nouns, NNPS for plural proper nouns. In contrast, the CLAWS C7 tag set has 22 distinctions at the noun level, including plural and singular month nouns, which may be useful in some studies but will not add much to this one.

The POS tagged files were then processed using Morpha (Minnen, Carroll, & Pearce, 2001), a morphology analyzer that, based on the tags, lemmatizes text with 99%
accuracy. Lemmatization refers to the base form of a word without any inflected forms.

Consider the following sentence taken from an undergraduate geography text.

The purpose of this chapter is to show how architecture and urban design are linked into the dynamics of urban change and spatial organization.

When it is POS tagged, it will look like the following.

\texttt{the\_DT purpose\_NN of\_IN this\_DT chapter\_NN is\_VBZ to\_TO show\_VB how\_WRB architecture\_NN and\_CC urban\_JJ design\_NN are\_VBP linked\_VBN into\_IN the\_DT dynamics\_NNS of\_IN urban\_JJ change\_NN and\_CC spatial\_JJ organization\_NN}.

Finally, the morphology analyzer takes the POS tagged text and outputs the following.

\texttt{the\_DT purpose\_NN of\_IN this\_DT chapter\_NN be\_VBZ to\_TO show\_VB how\_WRB architecture\_NN and\_CC urban\_JJ design\_NN be\_VBP link\_VBN into\_IN the\_DT dynamics\_NNS of\_IN urban\_JJ change\_NN and\_CC spatial\_JJ organization\_NN}.

The POS tags are optional on the lemmatized text, but as they can be ignored when using most analyzing software, it is usually helpful to keep the tags.

The result of tagging and lemmatizing is in effect three different corpora: bare text, POS tagged, and lemmatized. During analysis, different types of corpora are needed. For instance, lexical bundles are best analyzed in the bare corpus as lemmatized forms may obscure meaning. For example, \textit{on the other hand} and \textit{on the other hands} do not have the same meaning at all, but would be combined in a lemmatized corpus. On the other hand, a lemmatized corpus is far more useful in analyzing vocabulary as \textit{hand} and \textit{hands} are, in effect, the same word.

Text was analyzed and manipulated using a number of software programs and through Python scripts. These programs will be explained in more detail in the chapters in which they were used.
2.3 Summary

While there is no firm definition of academic English, we are allowing the context of an American university to define it for us. That is to say, we are focusing on the readings that are required in courses, thus looking at text that students in theory (although perhaps not in practice) are reading. Moreover, those texts were selected by the course syllabuses provided by the instructors for these courses.

The corpus is a little small at only two million words, but it is balanced and representative, having drawn on syllabuses for courses to determine which texts should be included, similar sized texts, and a similar number of texts for each discipline. In other words, while the corpus is small, it should be reliable and thus we should be able to draw conclusions about frequent vocabulary and hint at areas that may need further study, determining if the corpus should be expanded or if, in fact, there is no real difference between graduate and undergraduate academic English and thus no need to add more texts.
Chapter 3

Vocabulary

The English language has, arguably, 54,000 base words, excluding proper names, archaic words, compound nouns, abbreviations, and inflected and derivational forms (Goulden, Nation, & Read, 1990). On the other hand, when one is more inclusive, the English word count is much higher. Indeed, the Global Language Monitor (2009) reports that there are over one million words in English as of June 2009, but this is including all of the aforementioned exclusions as well as slang and newly coined or accidental words (e.g. *misunderestimate*, often referred to as a *Bushism*). Somewhere in the middle, *Websters Third International Dictionary* has about 465,000 headwords. However, the average educated native English speaker knows less than half of these words; Nation and Waring (1997) estimate that the average educated native English speaker only knows about 20,000 of the 54,000 word families.

For a learner of English, the numbers are still daunting, but less so when one considers the frequency of these words. For example, in 1972 Francis and Kucera found that the first 1000 most frequent words in the Brown corpus, a corpus of general written English comprised of five hundred 2,000 word text samples from 15 genres in 1961, account for 72% of the corpus, and that the first 3,000 words accounted for 84% of the corpus. In other words, if a person knows the 3,000 most frequent words in English, he or she will know all but 3 words in 20. Indeed, in a corpus of fiction for younger readers
(Hirsh & Nation, 1992), 2,600 words cover 96% of the corpus. So the level of the text and the register play a role in determining just how many words a reader must know.

It bears mentioning that these studies are not entirely comparable. Francis and Kucera use lemmas while Hirsh and Nation use word families. Lemmas, or lemmata, reduce words to their lexeme, or uninflected form. Inform, informs, and informed would all be reduced to INFORM, and go, goes, going, and went would all be reduced to GO.

Word families, on the other hand, include inflected forms and obvious derivational forms, so the word family for INFORM would include inform, informs, and informed, and also uninformed. Thus, studies which use lemmas will report a higher number of needed words to reach a percentage of coverage because there are fewer overall words included in a lemmatized word than in a word family. This point is made particularly clear when one looks at The Teacher’s Word Book of 30,000 Words (Thorndike & Lorge, 1944), which contains 30,000 lemmatized words, but only about 13,000 word families (Goulden, Nation, & Read, 1990), or less than half the number of lemmatized forms.

Lemmas and word families are the two traditional ways to combine multiple forms into one. There are advantages to both methods. Word families assumes that derivational and inflectional forms of words are mentally stored together, or at the least that one can figure out the meaning of a word given knowledge of a closed set of derivational affixes. However, as noted by Hyland and Tse (2007), words are used quite differently in their verbal and nominalized forms. For example, attribute (v.) means ‘to ascribe to’ while attribute (n.) means ‘feature’. Using lemmas avoids this issue.

Another, more practical, feature of using lemmas is that there are several software packages that will lemmatize a text, but the software packages that deal with word
families are far more restricted and rely on baseword lists that, as we will see, may not be entirely accurate.

Once a learner of English has learned these first most frequent words, lemmas, or word families, what should a non-native speaker who wants to use English in a university context learn next? Laufer (1988) found that one needs to know 95% of the tokens in a text to understand it. His study included 100 native Arabic and Hebrew first year students who were taking an EAP course. These students were asked to read a general academic text, underline the words they did not know, and then answer several questions about the text in a comprehension exam. Laufer found that students that knew 95% of the words in the text scored significantly better on the comprehension exam than those who knew a lesser percentage of the words.

Hu and Nation (2000) found that number to be somewhat higher at 98% for comprehension. In their study, 66 students enrolled in EAP courses who had tested well in vocabulary were randomly divided into four groups and asked to read a fiction passage. The passage was the same for each group except that, depending on the group, 0, 5, 10, or 20% of the least frequent words had been replaced with nonsense words, and the rest of the passage had been simplified to only include the 2000 most common words in English based on West’s (1953) General Service List. The students were then given a comprehension exam in two parts. Based on the scores of the students on the comprehension exams, Hu and Nation found that 95% was not an adequate threshold for vocabulary knowledge, and that around 98% would yield better results.

Schmitt, Jiang, and Grabe (2011) confirm that the greater the text coverage of known vocabulary, the better the comprehension of the text in their study of 661 college
students of varying backgrounds, ages, and native languages. Each student completed a vocabulary questionnaire that was created based on the vocabulary found in two passages of academic text. The students then read the two passages and answered several comprehension questions about them, but were not allowed to look back at the vocabulary questionnaire. The found that students who knew 95% of the tokens of the text scored an average of 60% on the comprehension exam while those who knew 98% scored on average a 70%. Those who knew all of the words scored a 75%, indicating that while there is a linear relationship between vocabulary knowledge and comprehension, knowing all of the words does not mean that students will necessarily understand everything.

Based on these findings, we can assume a student minimally should know 95% of the words in a text, with a greater coverage logically preferable. If common words account for 70-80% of a text, is there a set of words that can account for the remaining 15-28%, or some portion therein?

We take for granted that there is some form of an academic vocabulary, a set of words that would be particularly useful in the university context as they have a higher frequency there. Within this context, Nation (2001) suggests that vocabulary can be categorized into three subgroups: high frequency words (e.g. *the, of, to, and*) that are common across genres, academic vocabulary which accounts for 8-10% of tokens in academic discourse, and technical vocabulary that differs across disciplines and accounts for 5% of tokens. Some (e.g. Hyland & Tse, 2007; Chen & Ge, 2007; Martinez, Beck, & Panza, 2009) argue that this middle group should be excluded in favor of a stronger focus on specialized vocabulary. This leads us to our primary question: whether there is a
general academic vocabulary at all, that is to say, is it economical financially or
temporally (Strevens, 1988), to learn a general academic vocabulary? This question is
not trivial, nor is it easy to answer. First we must consider what academic vocabulary is.

Frequency and token variation, as mentioned previously, is one way to study
vocabulary (e.g. Francis & Kucera, 1972; Goulden, Nation, & Read, 1990; Hirsh &
Nation, 1992). In his study of university language, Biber (2006) looks at the frequency of
words, token variation, and part of speech between classroom English and textbooks and
between different disciplines. His corpus contains 1.67 million words in the spoken
portion (class sessions, labs/in-class groups, office hours, study groups, service
encounters) and 1.07 million words in the written portion (textbooks, course packs,
course management, institutional writing). He compares business, education,
ingineering, humanities, natural science, and social science, and has variable quantities
of text for each (e.g. education = 187,300 words, social science = 507,400 words) as
noted in Table 3.1. Despite this relatively small sample for vocabulary analysis, Biber
finds some notable results.

Table 3.1: Y2k-SWAL Corpus tokens, adapted from Biber (2006)

<table>
<thead>
<tr>
<th></th>
<th>Business</th>
<th>Education</th>
<th>Engineering</th>
<th>Humanities</th>
<th>Natural S.</th>
<th>Social S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classroom</td>
<td>236,400</td>
<td>137,200</td>
<td>171,300</td>
<td>248,600</td>
<td>160,800</td>
<td>294,400</td>
</tr>
<tr>
<td>Textbook</td>
<td>116,200</td>
<td>50,100</td>
<td>72,000</td>
<td>164,100</td>
<td>145,200</td>
<td>213,000</td>
</tr>
<tr>
<td>Total</td>
<td>352,600</td>
<td>187,300</td>
<td>243,300</td>
<td>412,700</td>
<td>306,000</td>
<td>507,400</td>
</tr>
</tbody>
</table>

First, he finds that classroom teaching in his corpus uses 14,500 different words,
or types, while textbooks use 27,000, despite the fact that the total number of running
words for classroom teaching exceeds that of textbooks. The majority of these words are
low frequency words (occurring fewer than 20 times per million). Classroom teaching
and textbooks use relatively few words extremely frequently (over 200 times per million
words) and while classroom teaching uses certain words with regularity (e.g. *get* was used over five thousand times per million words), textbooks have no content words used even one thousand times per million words. Thus, one can conclude that textbooks in this study use a wider range of types than classroom teaching. Not surprisingly, given these findings, while there are many words that are not found in classroom teaching, very few words are exclusively found in the classroom, and these same few words are primarily nouns that would not be discussed or used in a textbook, such as *bagel* or *fluff*.

With respect to disciplines, Biber found that business and engineering use far fewer word types than did natural science, social science, or especially the humanities. These differences in number of types can be, in part, attributed to the variation in disciplines found in the humanities, social sciences, and natural sciences, while there is less variation in business and engineering and more overlap. While types are less frequent in these areas, there is nonetheless little overlap between areas, indicating that there is a particularly specialized vocabulary for each.

Given the notion that there is in fact an academic vocabulary, several lists have been compiled based on corpora of academic texts, either by hand or through computational methods. In theory, these lists are a next step in vocabulary learning for EAP students after the most common word families are learned. Xue and Nation (1984) built the University Word List (UWL) based on four previously constructed lists of academic vocabulary (Campion & Elly, 1971; Praninskas, 1972; Lynn, 1973; and Ghadessy, 1979), which had been created by looking at either student annotations above words in textbooks or by identifying words across disciplines that occurred in a range of texts. It contains 836 word families, or 3,707 types.
Coxhead (2000) proposed a new academic word list (AWL) based on an academic corpus of her own design. The corpus was comprised of articles from academic journals, complete university textbooks, chapters from university textbooks, lab manuals, books from the MicroConcord academic corpus, and text segments from the Learned and Scientific sections of the Brown Corpus, Lancaster-Oslo/Bergen Corpus (LOB), and the Wellington Corpus of Written English, for a total of 414 academic texts and 3,513,330 tokens. Coxhead divided the corpus into four subsections (arts, commerce, law, and science), each comprised of seven disciplines and roughly the same number of tokens. Word families that a) did not appear on the General Service List (GSL) (West, 1953), b) occurred in at least 15 of the 28 subject areas and at least 10 times in each subsection, or approximately 13 times per million words, and c) occurred at least one hundred times in her academic corpus were included in the AWL, for a total of 570 word families. When Coxhead tested the AWL on a separate academic corpus, she had 10% text coverage, while the UWL had only 9.8% with 266 more word families.

However, the true usefulness of a list of general academic vocabulary has been questioned. Hyland and Tse (2007) argue that while the AWL served as a good guide for a general academic vocabulary, students actually need a more specified academic vocabulary to reflect their field of study. They also take issue with the construction of Coxhead’s corpus, calling it “opportunistic,” (p. 239) as it uses some 2000 word text fragments from academic sections of other corpora, has uneven numbers of text in the subcorpora, and has no samples of student writing. Through the use of their own academic corpus comprised of three subsections (engineering, sciences, and social sciences), each with equal numbers of whole texts (research articles, textbooks, book
reviews, and scientific letters) and some samples of student writing (master’s theses, doctoral dissertations, and final year project theses), Hyland and Tse demonstrate that while the AWL covers 10.6% of their corpus, the coverage was inconsistent, favoring engineering and especially computer science and putting science and particularly biology at a disadvantage. Moreover, they redefine “frequent” to be if an item “occurred above the mean for all AWL items in the corpus (i.e., 597)” (p. 240). Using this measure, only 192 of the AWL word families occur frequently in their corpus.

Ge and Chen (2007) took similar issue when they tested the AWL’s coverage on a corpus of medical research articles. While the AWL does cover on average 10.07% of their corpus, only 292 word families occurred more than 10 times. On the other hand, these results may be somewhat skewed as the corpus is comprised of only 50 articles, or 190,425 running words, a relatively small corpus particularly with respect to inquiry into vocabulary. Using a slightly larger corpus (826,416 running words), Martinez, Beck, and Panza (2009) tested the coverage of the AWL in a corpus of agricultural science research articles. They find that the coverage is only 9.06%, and that using Hyland and Tse’s method to measure frequency, only 92 words are frequent. It should be noted that Ge and Chen (2007) and Martinez, Beck, and Panza (2009) only compared the AWL to published research articles and excluded textbooks. Li and Qian (2010) took another approach and tested the coverage of the AWL in the Hong Kong Financial Services Corpus (HKFSC), a corpus of ~6.3 million running words of professional English used in the financial services industry reasoning that language used in academia should prepare one for the real world. In this study, the AWL covered 10.4% of the corpus, with only 162 of the word families being frequent according to Hyland and Tse’s measure.
Indeed, the one study that used the AWL exclusively on undergraduate textbooks did so not as a test of the AWL but rather as a measure of comparison. Miller (2010) compared the coverage of the AWL in the undergraduate textbook portion of the Y2KSWAL (252,100 tokens total) to the coverage of texts in ESOL course materials and found that the AWL’s coverage of the undergraduate texts was 8.8%, lower than the average coverage in either the Coxhead corpus or corpora of academic articles. The range of the words, or the occurrence of tokens across the disciplines, in the Y2KSWAL was not part of the study.

These studies show that the AWL has variable coverage and range, the question of how the AWL would cover and range across a general academic corpus at the undergraduate level remains unanswered, and it is in this undergraduate context that a general academic vocabulary would be most useful.

Despite this previous research into academic vocabulary, the differences between graduate and undergraduate vocabulary have not yet been explored using frequency and token variation. Therefore, the first step in this analysis of academic vocabulary is to consider how many words are high frequency (occurring more than 200 times per million words) and how many are low frequency (fewer than 20 times per million words)? What is the range of words used in the different disciplines and levels, and how do they compare? What are the most frequent types in the two different levels, and do they overlap? With respect to the range of words, is the range of types found in graduate input greater than that of undergraduate input? Previous studies have either lumped the same materials together (e.g. Biber, 2006; Coxhead, 2000) or have considered only what one might expect a graduate student to read (e.g. Ge & Chen, 2007; Martinez, Beck, & Panza,
2009). If the AWL worked well for a general corpus of academic English, how would it fare on a corpus specifically geared toward undergraduate input? In Miller (2010) the AWL did not have a particularly high coverage, but that corpus was small. It has also not fared as well when faced with single discipline graduate level corpora. Many of the findings in these other studies can be either confirmed or questioned with the study of the vocabulary within these two different levels of academic study.

3.1 What types are frequent and what is their range?

3.1.1 Methodology

For this portion of the analysis we will need to lemmatize the texts and then consider frequency and the type/token ratio across corpora of various sizes. To determine frequency, one needs to look at the number of lemmas per, for example, one million words. As the corpus and subcorpora do not neatly contain exactly one million words, the frequencies must be normalized so that comparisons can be made. To normalize the frequency, one must divide the number of lemmas by the total number of lemmatized tokens and then multiply by one million.

Biber (2006) defined lemmas that occur fewer than 20 times per million words as low frequency while lemmas that occurred more than 200 times per million words as high frequency. His designation was arbitrary, based on initial findings; however, I have found that his cut off points work neatly with my corpora as well.

A type is a unique word or lemma within the corpus, or within the domain of the corpus being considered. For example, orange would be one type, and even if it were found in the corpus fifteen times, it would still only count as one type, but as fifteen
tokens. The type/token ratio (TTR) is often used as a measure of lexical variation, in which the higher the TTR, the greater the diversity of words.

Like frequency, the TTR must be normalized if one is to compare several corpora. Unlike frequency, the relationship between the number of types and the size of the corpus is not linear (Richards & Malvern, 1997; Ardaud, 1992; Richards, 1987). There are a limited number of unique words in the English language as discussed above. Thus, a corpus of 100,000 tokens might have 10,000 types, but a corpus of one million tokens would not likely have 100,000 types – it might have 35,000 or 40,000. Therefore a larger corpus could well have a smaller TTR than a smaller corpus even if the corpora were created using the exact same texts.

There have been a number of ways established to normalize the TTR, some quite complex (e.g. Covington & McFall, 2010). One attempt, called the standardized TTR or mean segmental type token ratio (MSTTR) (Johnson, 1944), determines the mean TTR by finding the TTR of every subsequent 2000 words, or the total number of tokens from the smallest text sample in the corpus, and then averaging them together. Unfortunately, unless the text sample sizes are the same or one uses the same subsequent sample size, cross-corpora comparison will still be biased toward the corpus using the smaller text sample size. To create a statistic that is comparable across corpora and studies there have been several suggestions for transformations of the TTR. The corrected type token ratio (CTTR) developed by Carroll (1964) is the ratio of types to two times the square root of the number tokens. The bilogarithmic type token ration (LogTTR) developed by Herdan (1964) is the ratio of the logarithm of the types to the logarithm of the tokens and the root
type token ratio (RTTR) (Guiraud, 1960) is the ratio of the root of the types to the root of the tokens. However, none of these transformations truly account for variable text length.

Another transformation, D (McKee, Malvern, & Richards, 2000), is derived from an algorithm that fits the averaged TTR of non-replaced sub-samples to an ideal curve for increasing values of N (N=35 to N=50). While Malvern and Richards (2000) found the D measure to be a better measure for cross-corpora analysis, it is a far more complex mathematical derivation than its predecessors.

Alternatively, one can use Biber’s (2006) method. He found through a series of experiments that the normalized number of word types is roughly equal to the number of word types divided by the square root of the corpus size, and then multiplied by the square root of the size to which one wants to norm the corpora. The experiments entailed determining the number of types within a full corpus, half corpus, quarter corpus, and so on, and comparing percentages of representation, which he found were typically 70% of the corpus twice its size. The square root of one half is .70, which is why his formula appears to work. Because many of Biber’s methods are emulated in this study and because I have access to all of the corpora being analyzed, using Biber’s method for normalizing TTR would continue that parallel if the method used for the T2K-SWAL Corpus also applied on the American Undergraduate and Graduate Expected Readings (AUGER) corpus. Indeed, it does within a couple of percentage points. Therefore I normalized my TTR based on Biber’s method, but instead of using one million, as he does, I use one hundred thousand for many of the analyses in this chapter, the target size of the individual discipline corpora. Moreover, the smaller normalized number allows for
variation in the TTR to be more obvious; the smaller the corpus typically the greater the TTR and vice versa.

Once the normalized number of types has been established, one can compare the different sections of the corpora. Because undergraduate students are expected to take courses across disciplines and graduate students take courses in only one, the number of types in one area of the graduate corpus would need to be compared to the types found in the whole of the undergraduate corpus. However, comparing the types between disciplines at both levels may also net interesting results. Biber (2006) and Hyland (2008) both note that there is significant variation between disciplines, but assume this variation is universal across levels. In addition, the types in the undergraduate disciplines will need to be compared. Hyland’s findings were based on a corpus of academic journal articles while Biber’s were based on a mixed corpus of both spoken and written text across both levels of university study. Thus, one cannot assume that this variation will also exist at the undergraduate level of this corpus, although it is expected at the graduate level.

Not only is diversity significant, but also overlap. Types that are found across disciplines at any level are words which may define a “common core” of academic vocabulary. However, these words do need to be treated with some caution. As noted in Hyland and Tse (2007), not all words are used in the same way across disciplines. Even so, Granger and Paquot (2009) argue that many words, while used differently, still maintain a common meaning; moreover, while these same words may be used in one way in one discipline, these same words are also used in the same way across disciplines. For example, while the verb *analyze* may, in engineering, refer to “methods of determining
the constituent parts or composition of a substance” and in the social sciences refer to “considering something carefully” (Hyland & Tse, 2007, p. 244), *analyze* can sometimes carry the social science meaning in engineering contexts and the engineering meaning in social science contexts. It has multiple meanings within one domain. Thus, once a list of overlapping words is established, a more qualitative analysis of key words in context will need to be carried out to see if, in fact, there are both common elements and differentiating usage.

### 3.1.2 Results

The frequency of lemmatized types is shown in Figure 3.1. On the surface, the undergraduate and graduate corpora use the same number of lemmatized types extremely frequently; 674 types occur 200 or more times per million words in the undergraduate corpus compared to 671 in the graduate. 71% (70.9% of undergraduate and 71.2% for graduate) of those types are the same, while ~29% differ. Of the 389 types that do not co-occur (196 undergraduate and 193 graduate), 132 of the graduate types appear 100+ times per million for undergraduates, and 117 undergraduate types appear 100+ times per million for graduates. That is to say, while there is no complete overlap, the most frequently occurring types appear in both corpora frequently.
If we compare these findings with a dissimilar corpus of American literature, in this case a corpus of 13 popular novels written by American authors in the 20th century with the front and back matter removed, the most frequently occurring words are far less similar. The graduate corpus shares only 246 (37%) of its most frequent types with the 526 total types that occur more than 200 times per million in the literature corpus, and the undergraduate corpus shares only 302 (46%).

Zipf (1935) determined that there is an inverse relationship between frequency and rank in any corpus of natural language, an observation that came to be known as Zipf’s law. Rank is assigned to types in a corpus by decreasing frequency, so the most frequent type, in the AUGER corpus the, is given the rank of 1 for its 13,853 tokens; the next most frequent token, in this case of, is given the rank of 2 for its 8,745 tokens. The lower the rank, the greater the difference in frequency, so a type that ranks 300 may occur only slightly more frequently than the type with a rank of 301. Thus, there will be a relatively small set of words that occur frequently (e.g. the, of, be, to) but a very large number of words that occur infrequently, creating a logarithmic distribution. We can see this displayed in Figure 3.1. The most frequent words tend to be prepositions,
determiners, and other closed classed words with grammatical functions. The significance of having such deviation in the most frequent words in the academic corpus compared to the literature corpus is that we can see variation in the vocabulary used in the two registers. There is a set of vocabulary that is more frequent in academic contests.

Using Biber’s (2006) method to normalize types, the undergraduate corpus had 26,073 types per million words while the graduate corpus had 29,739 types. However, as noted in the Introduction, graduate students are not expected to read across fields extensively, so the number of types across the entire graduate corpus is not particularly relevant. If we look at normalized types by discipline, we see that, somewhat contrary to what Biber found, the soft fields have more types than the hard fields at both the graduate and undergraduate level. Unlike Biber, these are not whole categories of disciplines but rather disciplines in and of themselves, so discipline variation is less likely a cause of the type disparity. It is, of course, possible that within the disciplines we find greater topic variation. History encompasses over 3,000 years, and topics ranging from wars to technology to the soap operas of royalty, but ask a mathematician if the topics in his field are limited and he will tell you all about logic, number theory, and mathematical analysis, and claim ties to philosophy, natural science, engineering, and yes, even history.
We can clearly see in Figure 3.2 that, with the exception of chemistry, undergraduate texts contain more types than graduate texts. Similarly, if we compare the normalized type/token ratio of the whole of the undergraduate corpus to the normalized type/token ratio of the individual disciplines at the graduate level, we see that the type/token ratio is again higher with the exception of history. One might assume that given that the type/token ratio is consistently lower for the graduate level that the whole of the graduate corpus would also be lower, but it is actually highest at 9.4%, compared to the whole of the undergraduate corpus at 8.3%, shown in Figure 3.3. The greater type/token ratio in the graduate corpus implies there is more lexical diversity in the graduate corpus as a whole than in the undergraduate corpus, meaning that at the graduate level there is more lexical variation between disciplines than at the undergraduate level. As we saw when we looked at general lexical frequency, this variation is due in part to the more specialized vocabulary exhibited in these disciplines.
A key point is that there is still considerable variation in the frequent vocabulary and how it is used at the undergraduate level. If we look at vocabulary used at least 200 times per million words, in other words “frequently”, mathematics again stands out for having the fewest lemmas, suggesting that even though mathematics has the fewest types, it does not make up for a lack of types by using them much more frequently.

Perhaps more interesting is that the total lemmas used at least 200 times per million words determined by each undergraduate discipline is 2,478, but only 178 of those lemmas are used in all 10 disciplines frequently, or 7%, whereas 1,238 lemmas occur in only one discipline frequently, or fully 50% of the total frequent lemmas. Similarly, if we look at even more frequent lemmas, those appearing 400 times per million words, we find that of the 1,320 total extremely frequent lemmas, only 110 of them are extremely frequent in all ten disciplines, or 8%, while 698, or 53%, of the extremely frequent lemmas only appear in one discipline extremely frequently. While these extremely frequent and frequent lemmas may appear in the other disciplines, they
are not used nearly as frequently. Thus we can see that there is variation in how vocabulary is used between disciplines even at the undergraduate level. Preparing an undergraduate student going to an American university for one specific discipline would only prepare the student for courses within that discipline, and as was explained in the Introduction, the student would be expected to perform in several other disciplines that use English differently.

If we compare the disciplines at the graduate and undergraduate level, we find that, despite the fact that at the undergraduate level disciplines typically had more types than at the graduate level, the two levels are about even in terms of the number of frequent types, with half of the disciplines having slightly more at the graduate level, and the other half slightly more at the undergraduate. When we look at the actual words that are frequent, we see that only about half of the lemmas are shared at both levels in the hard fields, and about two thirds in the soft fields, as shown in Table 3.2. Moreover, if we consider that 120 of the overlapping lemmas appear frequently in all disciplines at both levels, we find that the percentage of words that may be associated with a discipline that overlap at the two levels is closer to 44% in the hard fields and 56% in the soft fields. These 120 common words include, for example, *the, of, and be*, as well as words like *example, however, and without*. There are also some notable exceptions, lemmas that are only found frequently at one level or the other (but may still be found at both levels), as seen in Table 3.2.
Table 3.2: Number of lemmas that appear 200 times per million or more.

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Under</th>
<th>Grad</th>
<th>Overlap</th>
<th>Notable exclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>biology</td>
<td>748</td>
<td>752</td>
<td>384</td>
<td>apply, date, finally, later, question, significant, top</td>
</tr>
<tr>
<td>chemistry</td>
<td>678</td>
<td>774</td>
<td>384</td>
<td>among, function, overall, procedure, standard</td>
</tr>
<tr>
<td>earth science</td>
<td>801</td>
<td>743</td>
<td>406</td>
<td>approximate, discussion, probably, throughout</td>
</tr>
<tr>
<td>mathematics</td>
<td>644</td>
<td>616</td>
<td>344</td>
<td>answer, approach, important, list, think, want</td>
</tr>
<tr>
<td>physics</td>
<td>750</td>
<td>769</td>
<td>405</td>
<td>comparison, estimate, great(er), process, quality, until</td>
</tr>
<tr>
<td>geography</td>
<td>734</td>
<td>711</td>
<td>441</td>
<td>before, close, indeed, main, modern, sometimes</td>
</tr>
<tr>
<td>history</td>
<td>703</td>
<td>663</td>
<td>461</td>
<td>concept, define, else, focus, instead, let</td>
</tr>
<tr>
<td>philosophy</td>
<td>790</td>
<td>613</td>
<td>457</td>
<td>actually, american, analyze, nevertheless, none, table</td>
</tr>
<tr>
<td>political science</td>
<td>729</td>
<td>760</td>
<td>424</td>
<td>context, end, explain, figure, nearly, recognize</td>
</tr>
<tr>
<td>sociology</td>
<td>697</td>
<td>791</td>
<td>465</td>
<td>basis, critical, differ, necessary, perhaps, present</td>
</tr>
</tbody>
</table>

At the undergraduate level a student is expected to take courses across disciplines, but it is significant that we can see differences even within the same discipline between levels. Students at both levels are expected to have a certain knowledge of English, and that underlying knowledge will account for many of the overlapping words. Recall that the graduate corpus shared 246 lemmas with a dissimilar corpus of American literature, and the undergraduate shared 302. Moreover, we can assume that it is only economical to teach words that are found with some frequency in a given discipline if we should want to take a discipline specific approach. So here we find a discrepancy. Within the frequent lemmas found in a discipline at two levels, there are only two or three hundred (at most) that might be considered a) not a word the student should already know, b) possibly specific to that discipline, and c) a word that a student, regardless of level, should know.

In summary, with respect to frequency and range, there is empirical evidence that there is an academic vocabulary; that is to say that vocabulary used in academic contexts differs from that used in other registers. There is also greater discipline variation between fields at the graduate level than at the undergraduate level where there is also discipline variation, but where that variation has less need to be addressed in the EAP
classroom because undergraduate students will be required to take courses across disciplines. Finally, 50% of lemmas that are used frequently, 200 times or more per million words, in both the graduate and undergraduate corpus overlap, more so in the soft fields than in the hard fields. However, many of these overlapping words are grammatical, including prepositions, determiners, and auxiliary verbs, or words that are quite common even outside of academic contexts, implying that the difference between the undergraduate and graduate levels lexically is greater than 50% might imply.

3.3 How well does the AWL cover the AUGER?

3.3.1 Methodology

The corpora were explored using Nation’s (2002) RANGE program, a program that allows the user to compare the frequency of words also found in a base word list across several subcorpora, in this case the 10 disciplines in the two levels of the AUGER corpus. The results are organized by frequency, alphabet, and range, that is, how many subcorpora in which the particular type or word family is found. The program comes with base word files for the first 998 word families in the General Service List (GSL), the second 960 word families in the GSL, and the complete AWL, 570 word families.

It is important to note that I altered the base word files that came with the program because I found them to be flawed. If a word family includes inflected forms as well as obvious derivational forms, then these base word files should also include all of these forms under each head word. However, the head words included in the base word files often did not include all of the derivative or inflected forms. For example, the head word equivalent was included in the AWL, with equivalence as the only additional word in the word family, so that words like equivalences, equivalents (inflectional forms) and
equivalency, equivalencies, and equivalently (derivational forms) were not included in the AWL inclusion numbers. Under uniform, uniformity and uniformly were both included, but not uniforms, uniformities, or uniformed. The discrepancies were not exclusively within the AWL; the GSL was also inconsistent, including the omission of widen and widens from the head word wide, which did include widened, widening, and width, for example.

According to Hirsh and Nation (1992), obvious derivational forms are derived from adding a certain prefix (in-, non-, un-) or suffix (-able, al, -ation, -er, ess, -ful, -ish, -ism, -ist, -ity, -ize, -less, -ly, -ment, -ness, -th, -y). Using this guideline, a python script compared the base words in the GSL and AWL to a word list generated from the AUGER corpus and extracted possible excluded words. I then manually went through these words to update the GSL and AWL to be more inclusive. The revised lists should not be considered comprehensive, as they were revised from the words in the AUGER corpus, but for the purpose of this study, the revised lists are quite useful.

3.3.2 Results

The coverage of the AWL of the AUGER corpus was varied. At the graduate level, the GSL accounted for 74.8% of the total word families and the AWL covered a respectable 10.5%, for a total of 85.3%. At the undergraduate level, the GSL accounted for 78.5% of the represented word families, and the AWL only accounted for 7.8%, for a slightly higher coverage of 86.3%, but not because of the AWL. We can see that the AWL covers the overall graduate corpus far better than it does the undergraduate corpus, and it also covers most of the individual disciplines at the graduate level better than the
undergraduate corpus as a whole, the exceptions being history (6.5%) and philosophy (6.3%).

Table 3.3: Coverage of the GSL and AWL at the graduate level

<table>
<thead>
<tr>
<th>Hard-pure</th>
<th>GSL 1</th>
<th>GSL 2</th>
<th>AWL</th>
<th>Overall</th>
<th>AWL Families</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td>60.1%</td>
<td>4.7%</td>
<td>11.8%</td>
<td>76.6%</td>
<td>493</td>
</tr>
<tr>
<td>Chemistry</td>
<td>61.3%</td>
<td>4.2%</td>
<td>10.4%</td>
<td>75.9%</td>
<td>447</td>
</tr>
<tr>
<td>Earth Science</td>
<td>63.8%</td>
<td>5.2%</td>
<td>9.6%</td>
<td>78.6%</td>
<td>461</td>
</tr>
<tr>
<td>Mathematics</td>
<td>74.1%</td>
<td>3.9%</td>
<td>9.9%</td>
<td>87.9%</td>
<td>345</td>
</tr>
<tr>
<td>Physics</td>
<td>66.6%</td>
<td>6.1%</td>
<td>10.6%</td>
<td>83.3%</td>
<td>455</td>
</tr>
<tr>
<td>Soft-pure</td>
<td>73.7%</td>
<td>4.9%</td>
<td>10.5%</td>
<td>89.1%</td>
<td>568</td>
</tr>
<tr>
<td>Geography</td>
<td>67.8%</td>
<td>5.3%</td>
<td>14.1%</td>
<td>87.2%</td>
<td>548</td>
</tr>
<tr>
<td>History</td>
<td>75.5%</td>
<td>5.1%</td>
<td>6.5%</td>
<td>87.1%</td>
<td>541</td>
</tr>
<tr>
<td>Philosophy</td>
<td>81.2%</td>
<td>4.2%</td>
<td>6.3%</td>
<td>91.7%</td>
<td>504</td>
</tr>
<tr>
<td>Political Science</td>
<td>71.1%</td>
<td>5.0%</td>
<td>14.0%</td>
<td>90.1%</td>
<td>526</td>
</tr>
<tr>
<td>Sociology</td>
<td>70.7%</td>
<td>5.1%</td>
<td>13.3%</td>
<td>89.1%</td>
<td>529</td>
</tr>
<tr>
<td>Graduate Total</td>
<td>69.9%</td>
<td>4.9%</td>
<td>10.5%</td>
<td>85.3%</td>
<td>569</td>
</tr>
<tr>
<td>Undergraduate Total</td>
<td>73.3%</td>
<td>5.3%</td>
<td>7.8%</td>
<td>86.4%</td>
<td>570</td>
</tr>
</tbody>
</table>

With respect to the AWL favoring certain disciplines over others, the evidence clearly shows that it does at the graduate level. While philosophy only shows a 6.3% coverage by the AWL (but an 85.3% coverage by the GSL for a total of 91.6% total coverage by the GSL and AWL), geography shows an impressive 14.1% coverage by the AWL (but only 73.1% coverage by the GSL for a total of 87.2% overall coverage), more than twice the coverage of philosophy or history. The span within the hard-pure disciplines is far less dramatic, ranging from 9.6% in earth science to 11.8% in biology, but still notable. All of the results can be seen in Table 3.3.

All 570 word families were represented in the undergraduate corpus, while only 569 were in the graduate corpus (interestingly, *clause* was the only word family not found). Of these word families, 258 were found across all 10 disciplines in the undergraduate corpus, 253 in the graduate corpus, 119 across 9 disciplines in the...
undergraduate corpus and 112 in the graduate corpus, and 83 across 8 in the undergraduate corpus, and only 63 in the graduate. Based on these numbers, it could be said that while the AWL may have better coverage at the graduate level, it has better range at the undergraduate level.

Perhaps more interesting are the word families that were excluded from the GSL and AWL, but were found to be in all disciplines. As can be seen in Table 3.4, 95 word families were found across all 10 disciplines and appeared at least 13 times per million words, the threshold set by Coxhead (2000). Admittedly, this threshold is extremely low as each subcorpus consists of ~100,000 words, so a word family only needs to appear once to technically qualify. For that reason the threshold was raised to 2, a still notably low threshold. Nonetheless, words like *reference*, *routine*, *original*, or *height* are not included in the GSL or AWL, yet would seem intuitively like common words. While age has often been an argument against the GSL, it hardly seems a factor for most of the words on this list with the possible exception of *tech*. As for the AWL, it is possible that with around 100,000 words for each of the 28 disciplines that some of these words were left out, similar to the exclusion of AWL words in this AUGER corpus.
Table 3.4: Word families found in all 10 disciplines but not in the GSL or AWL, regardless of level, excluding geographical references.

<table>
<thead>
<tr>
<th>AMPLIFY</th>
<th>COUNTERPART</th>
<th>INTIMACY</th>
<th>REFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABUNDANCE</td>
<td>DENSE</td>
<td>INTUIT</td>
<td>RENDER</td>
</tr>
<tr>
<td>ACCELERATE</td>
<td>DEPICT</td>
<td>INVERSE</td>
<td>RESEMBLE</td>
</tr>
<tr>
<td>ACCOMPLISH</td>
<td>DEPOSIT</td>
<td>IRONIC</td>
<td>RESIDUAL</td>
</tr>
<tr>
<td>AMAZE</td>
<td>DESIGNATE</td>
<td>LABORATORY</td>
<td>RIGOR</td>
</tr>
<tr>
<td>ARRAY</td>
<td>DICTATE</td>
<td>LATIN</td>
<td>ROUTINE</td>
</tr>
<tr>
<td>ARTEFACT</td>
<td>DISPERSE</td>
<td>MAGNITUDE</td>
<td>SCRUTINY</td>
</tr>
<tr>
<td>ASSERT</td>
<td>DOSE</td>
<td>OPTIMAL</td>
<td>SIMULTANEOUS</td>
</tr>
<tr>
<td>ATOM</td>
<td>EMBED</td>
<td>ORIGINAL</td>
<td>SOPHISTICATE</td>
</tr>
<tr>
<td>BINARY</td>
<td>ENTAIL</td>
<td>OVERWHELM</td>
<td>SPAN</td>
</tr>
<tr>
<td>BIOLOGY</td>
<td>EXHAUST</td>
<td>PHRASE</td>
<td>SPATIAL</td>
</tr>
<tr>
<td>CANDIDACY</td>
<td>EXPEND</td>
<td>PIONEER</td>
<td>SUBTLE</td>
</tr>
<tr>
<td>CELL</td>
<td>FEASIBLE</td>
<td>PLOT</td>
<td>SUPPRESS</td>
</tr>
<tr>
<td>CHEMISTRY</td>
<td>FEED</td>
<td>PRIMITIVE</td>
<td>SYNTHESES</td>
</tr>
<tr>
<td>CHRONIC</td>
<td>GROSS</td>
<td>PROFILE</td>
<td>TECH</td>
</tr>
<tr>
<td>CLUSTER</td>
<td>HEIGHT</td>
<td>PROFOUND</td>
<td>THRESHOLD</td>
</tr>
<tr>
<td>CONCRETE</td>
<td>HOMOGENATE</td>
<td>PROMINENCE</td>
<td>TRIVIAL</td>
</tr>
<tr>
<td>CONFIGURE</td>
<td>HUGE</td>
<td>PROPAGATE</td>
<td>UNCLEAR</td>
</tr>
<tr>
<td>CONJUNCTION</td>
<td>IMPRESS</td>
<td>QUANTITATIVE</td>
<td>UNEXPECTED</td>
</tr>
<tr>
<td>CONSERVE</td>
<td>INSIST</td>
<td>REALM</td>
<td>VAST</td>
</tr>
<tr>
<td>CONVERGE</td>
<td>INSPIRATION</td>
<td>RECALL</td>
<td>VERIFY</td>
</tr>
<tr>
<td>CORRELATE</td>
<td>INTERIOR</td>
<td>RECONCILE</td>
<td>VICE</td>
</tr>
<tr>
<td>COUNTER</td>
<td>INTERSECT</td>
<td>RECUR</td>
<td></td>
</tr>
</tbody>
</table>

Some of the word families included are surprising; biology, chemistry, and mathematics are all disciplines themselves so unexpected to be found in all other disciplines. However, when we consider their derivative forms, it is understandable how these word families could find themselves in other disciplines, as can be seen below.

1. Still, the discovery of life of any kind would be profound, both to our understanding of biology and to philosophical considerations of our place in the universe. (physics)

2. Interestingly, it seems that men regard Women's subordination as something natural, something that is effectively determined by biology. (geography)
3. Hyperbolas occur frequently as graphs of equations in **chemistry**, physics, **biology**, and economics (Boyle's Law, Ohm's Law, supply and demand curves). (mathematics)

4. Indeed, some point to the neighborhood as one of the key elements, along with the family, in the **chemistry** of the Italian American ethnic group (political science)

5. If a **chemist** tells you that something is a metal, that tells you a lot about its likely behaviour. (philosophy)

6. You could show them every book and ask every expert, but you could not provide them with direct evidence without a telescope, a lot of time and a lot of **mathematics**. (history)

7. For example, Harold, a 68-year-old grandfather, was frustrated by his inability to help his granddaughter with her **math** homework: (sociology)

8. This example demonstrates that the **mathematical** constraint of orthogonality within the set of spatial maps does not necessarily imply that large areas of 'activation' which overlap significantly between maps can no longer be extracted. (biology)

What we have found is that first, the GSL has greater coverage of undergraduate texts than it does most graduate level texts, with the exceptions of history and philosophy, implying that undergraduate texts may be written using less technical language. While all 570 word families did appear in the undergraduate corpus, less than half were found in all 10 disciplines, and only two thirds of the AWL words were found in at least 9 disciplines. While the range was greater than that of the graduate corpus, it was still quite low. Moreover, the words that were excluded from the GSL and the AWL were surprising and prolific. In other words, while the AWL may be an excellent idea, the list itself is probably not economical for a student attending an American university.
3.4 The America-centric use of language

During the analysis above, one characteristic of the language stood out: the use of nation-centric language in the undergraduate corpus. It was mentioned that an impetus for this study is the fact that many of the arguments made for discipline specific academic English do not consider a general education model of education as one would find in the United States, and that for that reason the texts included in the corpus would be those required by an American institution. Not unsurprisingly then, we find the use of American-centric language throughout both levels of the corpus; however, it is significantly (df = 4, $\chi^2 = 34.76, p < .001$) more prevalent in the undergraduate corpus, as can be seen in Table 3.5.

Table 3.5: Use of nation-centric language

<table>
<thead>
<tr>
<th></th>
<th>Undergraduate</th>
<th>Graduate</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>663</td>
<td>254</td>
</tr>
<tr>
<td>U.S./US</td>
<td>347</td>
<td>104</td>
</tr>
<tr>
<td>USA</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>America</td>
<td>218</td>
<td>99</td>
</tr>
<tr>
<td>American</td>
<td>576</td>
<td>302</td>
</tr>
<tr>
<td>Total</td>
<td>1,813</td>
<td>775</td>
</tr>
</tbody>
</table>

* frequencies per million words

It should be noted that the numbers above are not taken directly from the corpus. Instances of *US* that were not referring to the United States were excluded, as were instances of *America* and *American* that were preceded by *Central, Latin, South,* and *Spanish.* No instances of U.S.A. were found.

The prevalence of this language will make itself particularly known in the study of lexical bundles in the next chapter. It may also be significant because it shows a direction in which undergraduate studies take. Thus, while there may not be a specific vocabulary that undergraduate students need to learn, there may be a need for cultural
instruction. While to many this may seem like common sense, the nation-centric leanings of undergraduate vocabulary provide linguistic evidence to back up this assumption.

3.5 Summary

In our analysis of the vocabulary we see that there is in fact some difference between graduate and undergraduate academic English. Given the drive for discipline specific ESOL instruction, it is of particular importance to know that frequent vocabulary at the graduate and undergraduate level vary greatly, with only about 50% of the words overlapping, many of these closed class grammatical words. In general, there were more types at the undergraduate level, but the number of frequent words was about the same as with the graduate level. Discipline variation was also clearer at the graduate level than at the undergraduate.

The AWL proved to be less useful at the undergraduate level, with just a 7.8% coverage of the undergraduate portion of the AUGER. Thus, again (e.g. Hyland and Tse, 2007), the true usefulness of the AWL comes into question. The AWL was created as a general academic word list, not discipline specific. However, if a student intends to study at the graduate level, a more discipline specific academic word list might well serve him or her better than a general list, while an undergraduate student, who actually needs to have a general vocabulary list, is not well served by the AWL.

Finally, we see that location plays a significant role in topic selection, at least in the United States, at the undergraduate level but much less so at the graduate level.
Chapter 4
Lexical Bundles

Studying vocabulary at the “word” level is somewhat isolationist. Words do not exist alone but rather in context. We also know that words regularly appear with other specific words (collocations). There is abundant evidence that there are groups of words that commonly appear in a certain order (Nattinger & DeCarrico, 1992; Scott, 1996; Biber et al., 1999; Schmitt, 2000). Idioms, narrowly defined as expressions which cannot easily be understood by the meaning of their constituent words (e.g. *bite off more than one can chew, easy as pie*), are one example, although they are fairly rare in English usage (Moon, 1998). Phrasal verbs (e.g. *turn on, turn up, turn around*) are another example. It has even been suggested that up to 80% of language makes use of these word sets (Alternberg, 1998).

These word sets are significant for a couple of reasons. First, research in cognitive linguistics suggests that these word groupings, or constructions, much like words, have form-meaning mappings (Langacker, 1987; Goldberg, 2006; Robinson & Ellis, 2009). Secondly, native-like language use relies on the use of these collocations, where the lack of such collocations is often a marker of novice writers (Haswell, 1991; Hyland, 2008).

These groups of words have been studied under numerous names, with slightly different meanings, including multi-word items (Moon, 1997), lexical bundles (Biber et al., 1999), lexical chunks (Schmitt, 2000), lexical phrases (Nattinger & DeCarrico, 1992),
and clusters (Scott, 1996). A multi-word item is a sequence of two or more words which forms a meaningful unit through lexical or semantic fossilization and not by grammatical means (Moon, 1997). Thus go on and dog gone are multi-word items, while had gone and was going are not. A multi-word unit should exhibit at least one of the following criteria: institutionalization, fixedness (can it be inflected, is there variable lexis?), and non-compositionality (can the string be interpreted on a word by word analysis?). A phrasal verb, a verb plus a preposition, is a multi-word item because it cannot be interpreted on a word by word analysis. Fixed phrases such as of course and by far are also non-compositional, while good morning and you know are institutionalized. Multi-word items also include lexical bundles and prefabs, semi-fixed strings or stereotyped chunks found with some frequency in English (e.g. the thing/point/fact is, at the same time).

Schmitt (2000) defines lexical chunks or phrases inclusively, noting that multi-word items, or units, do not account for the longer strings of words that frequently pattern together. In his definition, he includes prefabricated and fixed expressions and does not discount multi-word items, although he does not discuss frequency as a means of determining inclusion, as with lexical bundles.

Lexical bundles, sometimes referred to as n-grams, are sequences of three or more words which appear repeatedly in a corpus (Biber et al., 1999). These words are not necessarily institutionalized, fixed, or non-compositional, and they may in fact stretch across syntactic boundaries. In other words, lexical bundles are groups of words that are defined empirically instead of intuitively (Cortes, 2004).
For the purpose of this study, lexical bundles will be the unit of analysis. While multi-word units have historically been the unit of analysis, and lexical chunks allows for a looser interpretation, the study is an empirical one seeking what sequences are common and how these vary across levels.

There have been a handful of recent studies on lexical bundles in academic English, beginning with Biber et al.’s (1999) comparison of lexical bundles in conversation and academic prose. It was in this study, the *Longman Grammar of Spoken and Written English*, that Biber first coined the term ‘lexical bundle’. At the core of its meaning, lexical bundles rely on frequency or recurrence, an arbitrary attribute. For the purpose of this particular study, Biber et al. determined that a word sequence would only be counted as recurrent if it appeared ten times per million words and across at least five different texts. They use a lower threshold for the less common five and six word lexical bundles.

Biber et al. found that 21% of the words in academic prose, defined here as primarily journal articles with a few book chapters, almost none of which written for a lay or non-scholarly audience, consisted of recurrent lexical bundles, the vast majority of which (18%) were three word lexical bundles. Moreover, three word bundles occurred 60,000 times per million words, and four word bundles 5000 times per million words. Partial noun phrases and prepositional phrases accounted for 60% of all lexical bundles. It is important to note that lexical bundles are rarely complete structural units (e.g., noun phrase, verb phrase), but rather span between two. This finding is significant as much of the research on lexical chunks and multi-word units prior to this study assumed structural dependence. When lexical bundles were complete structures in academic prose, they
usually formed prepositional phrases used as discourse markers (e.g. *on the other hand*, *in the same way*).

Necessarily, many three-word bundles were part of four word bundles, as the common three-word lexical bundle *on the other* is usually part of the common four-word lexical bundle *on the other hand*. However, an interesting finding showed that the some common three word bundles might be part of more than one four-word bundle, as in *I don’t know*, a common three-word bundle particularly in conversation, and *I don’t know what* or *I don’t know why*.

Biber et al. went on to categorize lexical bundles by their structural nature, creating twelve distinct groups (see below). While Biber et al. do not specifically denote functional categories yet, they do comment on the functions of some of the structural categories. For example, the first category, noun phrase with an *of*-phrase fragment, is often used for physical description, marking existence, identifying abstract qualities, and describing events that occur over an extended period of time.

1. Noun phrase with *of*-phrase fragment
2. Noun phrase with other post-modifier fragment
3. Prepositional phrase with embedded *of*-phrase fragment
4. Other prepositional phrase fragment
5. Anticipatory *it* + verb phrase/adjective phrase
6. Passive verb + prepositional phrase fragment
7. Copula *be* + noun phrase/adjective phrase
8. (Verb phrase +) *that*-clause fragment
9. (Verb/adjective phrase +) *to*-clause fragment
10. Adverbial clause fragment
11. Pronoun/noun phrase + *be* (+ …)
12. Other expressions

(Biber, et al., 1999, pp. 1014-1015)

Cortes (2004), in her study of student and published writing in history and biology, did categorize lexical bundles by function. She used two corpora, one of history
and biology journal articles with around two million words, and one of student writing (lower and upper division undergraduate writing and graduate writing) in the same fields with roughly one million words. Looking at four word lexical bundles, which she defined as occurring more conservatively at least 20 times per million words, she categorized the lexical bundles based on a taxonomy she created previously and that Biber, Conrad, and Cortes (2003) revised. This taxonomy included three main headings: referential bundles (time makers, place markers, descriptive bundles, quantifying bundles, and subject bound bundles), text organizing bundles (contrast/comparison, inferential, focus, framing), and stance bundles (epistemic-impersonal/probable-possible and other). Basing their findings on the T2K-SWAL corpus, Biber, Conrad, and Cortes (2004) and reprinted in Biber (2006) also used this taxonomy, but with different subheadings (below).

- Stance Expressions
  - Epistemic stance (e.g. I don’t know if)
  - Attitudinal/modality (e.g. it is important to)
- Discourse organizers
  - Topic introduction/focus (e.g. if you look at)
  - Topic elaboration/clarification (e.g. has to do with)
- Referential
  - Identification/focus (e.g. that’s one of the)
  - Imprecision (e.g. or something like that)
  - Specification of attributes (e.g. the size of the)
  - Time/place/text reference (e.g. at the same time)
    (Biber, Conrad, & Cortes, 2004; Biber 2006; Biber & Barbieri, 2007)

Stance expressions reflect one’s interpretation of an upcoming proposal, and can be divided into two categories: Epistemic and attitudinal/modal. Epistemic expressions, what Cortes called impersonal/probable-possible, express certainty, or a lack of certainty. Personal epistemic expressions usually express uncertainty, while impersonal epistemic expressions express certainty. On the other hand, attitudinal/modality expressions reflect
one’s feelings on the following proposition, not one’s confidence in it. These expressions can be further broken down into expressions of desire (e.g. *I don’t want to*), obligation/directive (e.g. *it is important to*), intention/prediction (e.g. *what we’re going to do*), and ability (e.g. *to come up with*).

Discourse organizers serve three functions, to introduce a new topic and focus, to clarify or elaborate a point, or to identify a point. Identifying a point can also be a referential point as its function is to mark the following text as especially important. Other referential bundles reflect imprecision (e.g. *and things like that*), specify attributes (e.g. *the size of the*), or reference a time, place or text-deixis (e.g. *As shown in figure, the end of the*).

Following the principles of these earlier studies, Hyland (2008) created his own functional taxonomy, in part because his corpus of written academic English (including published journal articles, Ph.D. theses, and M.A. theses) was not entirely comparable to the T2K-SWAL corpus, which included spoken English as well as institutional written English. He grouped them inductively based on the research driven writing in his corpus (see below).

- **Research oriented**
  - Location (e.g. in the present study)
  - Procedure (e.g. the role of the)
  - Quantization (e.g. a wide range of)
  - Description (e.g. the structure of the)
  - Topic (e.g. the currency board system)
- **Text oriented**
  - Transition signals (e.g. on the other hand)
  - Resultative signals (e.g. as a result of)
  - Structuring signals (e.g. in the next section)
  - Framing signals (e.g. in the case of)
- **Participant oriented**
  - Stance features (e.g. are likely to be)
Engagement features (e.g. it should be noted)  
(Hyland, 2008, pp 13-14)

Hyland found text and research oriented lexical bundles tended to be equally popular, while participant oriented lexical bundles were considerably less so. He also found that with his taxonomy, structural patterns more or less correlated with the three different orientations. Noun phrases + of tended to be research oriented constructions, while prepositional phrase patterns were text oriented, and anticipatory it was participant oriented.

The main focus of Hyland’s study was to look at four-word lexical bundles across disciplines. What he discovered was that different disciplines tend to use different lexical bundles, and that less than half of the top 50 four-word lexical bundles in the four disciplines he studied (applied linguistics, business studies, electrical engineering, and biology) occurred in any of the other disciplines. Indeed, fewer than 30% of the lexical bundles occurred in three or more of the fields, implying that there is enough differentiation between disciplines that EAP programs need to focus more on discipline specific instruction over general academic English.

While Cortes (2004) did look at student writing at the graduate and undergraduate level, she compared all of the student writing to academic articles, texts that undergraduate students are unlikely to be accessing. Moreover, while Biber and his colleagues have looked at lexical bundles in academic prose and textbooks, neither corpus differentiated between graduate and undergraduate levels. That is to say, while Biber and Barbieri (2007) compare the lexical bundles found in the academic prose
portion of the LGSWE corpus and the textbook section of the T2K-SWAL corpus, both sections consist of texts from all levels of academia.

This leaves investigation into the use of lexical bundles at the graduate and undergraduate level open. Not only is it important to look at which lexical bundles are frequent, as previous studies have involved relatively small corpora for research in vocabulary, but also to look at how they are used. Because the corpus is comprised exclusively of written texts, it can be hypothesized that Hyland’s (2008) taxonomy for the functions of lexical bundles will still apply. However, with the inclusion of textbooks, that may not be the case. Therefore, one of the questions of this study is to analyze how lexical bundles are being used at the graduate and undergraduate level and to determine if they are being used in the same ways across disciplines.

4.2 Analysis of Lexical Bundles

Lexical bundles, or n-grams, as suggested by the name, are derived through a frequency-driven approach. Previous research (e.g. Cortes, 2004, Hyland 2008) tends to confine these bundles to four or five word groupings out of necessity. Smaller groupings are particularly common and yield unwieldy results, and often are parts of larger lexical bundles. Larger groupings are particularly rare. Indeed, Cortes (2004) confined her analysis to only four-word groupings as five-word lexical bundles were just not as common. Similar results have been found within the AUGER corpus; while five-word lexical bundles do exist with frequency in the corpus as a whole, at the discipline level they are quite rare.

Because lexical bundles are determined by frequency, there must be a cut-off, a point at which a group of words is not considered frequent. Biber (2006) chose the
“conservative” (p. 134) number of 40 times per million, while Cortes (2004) and Hyland (2008) both chose the “conservative” (p. 404 and p. 8 respectively) cut-offs of 20 times per million words, and Biber et al. (1999) allowed all lexical bundles greater than 10 times per million words. The cut-off is arbitrary, but for the purpose of this study 20 times per million will typically be used.

Moreover, the lexical bundles must be found in a range of texts to account for author bias. Biber (2006) considered five texts enough to account for author bias, while Hyland (2008) used 10%. When looking at an entire level, one is considering 132 (undergraduate) or 140 (graduate) texts, but when looking at a specific discipline within a level, there are only 13-15 texts. Therefore, using 10% of the texts would not actually account for author bias in a discipline level analysis.

4.2.1 What is the frequency of lexical bundles across corpora?

Interestingly, Biber, Conrad, and Cortes (2004) found that comparatively, textbooks do not exhibit heavy usage of lexical bundles, particularly in comparison to classroom discourse. However, Hyland (2008) found lexical bundles to be quite common, if different, across disciplines, in academic journal writing. The question, then, is whether the undergraduate corpus, comprised of textbooks, will exhibit far fewer lexical bundles than the upper level corpora which contain academic journals?

The answer is essentially no, the graduate corpus does not contain significantly more lexical bundles than the undergraduate corpus. While there are a few more lexical bundles that occur 20 times per million words or more (see Table 4.1), the numbers are not significant for any of the three lengths of lexical bundles using the chi-squared test ($p = .39$ for three words, $p = .12$ for four words, and $p = 1$ for five words).
However, while the number of different lexical bundles is not significant, three-word lexical bundles that meet the requirement for inclusion occur 52,601 times in the graduate corpus and only 42,838 times in the undergraduate corpus. Four-word lexical bundles exhibited a similar disparity, with the 131 undergraduate lexical bundles occurring 3,748 times and the 145 graduate lexical bundles occurring 5,511 times. Thus the graduate level texts made greater use of the lexical bundles that were used.

<table>
<thead>
<tr>
<th></th>
<th>Lexical Bundles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>three word</td>
</tr>
<tr>
<td>Graduate</td>
<td>1,308</td>
</tr>
<tr>
<td>Undergraduate</td>
<td>1,271</td>
</tr>
</tbody>
</table>

### 4.2.2 Lexical bundles at the undergraduate and graduate levels

A second, and related, question asks which lexical bundles are frequent at the graduate and undergraduate levels. Are they the same, or are there different bundles associated with different levels of writing? This is noteworthy as Hyland (2008) notes that it is a lack of appropriate lexical bundles that often marks a writer as novice. However, if this novice writer is not exposed to the lexical bundles expected in his or her discipline, how is said learner expected to know which bundles are common and which ones are not?

In the case of both three-word lexical bundles and four-word lexical bundles, six lexical bundles are shared in the top ten most frequent in the graduate and undergraduate corpus, or over half. However, this is comparing the whole of both corpora, and as has been mentioned many times previously, a graduate student would really only be exposed to his or her area of study while an undergraduate student would have this more
comprehensive exposure. Moreover, the frequencies of the common lexical bundles are still quite different in some cases.

Table 4.2: Top three-word lexical bundles across both corpora

<table>
<thead>
<tr>
<th>Undergraduate</th>
<th>Graduate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the united states</td>
</tr>
<tr>
<td>2</td>
<td>one of the</td>
</tr>
<tr>
<td>3</td>
<td>as well as</td>
</tr>
<tr>
<td>4</td>
<td>the number of</td>
</tr>
<tr>
<td>5</td>
<td>in the united</td>
</tr>
<tr>
<td>6</td>
<td>some of the</td>
</tr>
<tr>
<td>7</td>
<td>as a result</td>
</tr>
<tr>
<td>8</td>
<td>part of the</td>
</tr>
<tr>
<td>9</td>
<td>of the world</td>
</tr>
<tr>
<td>10</td>
<td>for example the</td>
</tr>
</tbody>
</table>

Table 4.3: Top four-word lexical bundles across both corpora

<table>
<thead>
<tr>
<th>Undergraduate</th>
<th>Graduate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>in the united states</td>
</tr>
<tr>
<td>2</td>
<td>on the other hand</td>
</tr>
<tr>
<td>3</td>
<td>at the same time</td>
</tr>
<tr>
<td>4</td>
<td>the end of the</td>
</tr>
<tr>
<td>5</td>
<td>as a result of</td>
</tr>
<tr>
<td>6</td>
<td>one of the most</td>
</tr>
<tr>
<td>7</td>
<td>in the case of</td>
</tr>
<tr>
<td>8</td>
<td>the united states and</td>
</tr>
<tr>
<td>9</td>
<td>as shown in figure</td>
</tr>
<tr>
<td>10</td>
<td>in the form of</td>
</tr>
</tbody>
</table>

As we can see in Table 4.2, *the united states* is by far the most frequent three-word lexical bundle in the undergraduate corpus, occurring 605 times per million words. It occurs only 205 times in the graduate corpus, nearly a third as frequent. It could be argued that this lexical bundle is part of a larger lexical bundle, and perhaps should not be included as a three-word lexical bundle; however, it occurs in several different four-word lexical bundles, including *in the united states* (262 occurrences in the undergraduate corpus, 88 in the graduate), *the united states and* (72 occurrences in the undergraduate corpus, 27 in the graduate) and *of the united states* (56 occurrences in the undergraduate corpus, 29 in the graduate). Even excluding these instances, *the united states* would rate
in the top ten most frequent three-word lexical bundles. Indeed, it is noteworthy just how prevalent the lexical bundle is; *the united states* occurs 605 times, the second most frequent three-word bundle, *one of the*, occurs only 374 times, or not even two thirds as frequent. See section 3.4 for further discussion on this point.

If we consider Hyland’s (2008) taxonomy, we see that at the graduate level the lexical bundles are more textual while the ones at the undergraduate level are more research-oriented with both groups of lexical bundles. With the exception of the bundles referring to the United States, the overlapping bundles are either text-oriented or research-oriented with respect to quantification. Recall that these are names of functions of the lexical bundles, thus a research-oriented lexical bundle does not necessarily mean it is used strictly in the reporting of research but rather are used to “help writers to structure … activities and experiences of the real world” (Hyland, 2008, p. 13). Thus, while useful in the reporting of research, these lexical bundles would also be used to report simple facts and potentially used very differently at the undergraduate level, as will be addressed below.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Undergraduate</th>
<th>Graduate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>at the end of the 34</td>
<td>in the case of the</td>
</tr>
<tr>
<td>2</td>
<td>by the end of the 27</td>
<td>the same procedure was used 30</td>
</tr>
<tr>
<td>3</td>
<td>in the united states in 24</td>
<td>on the other hand the 27</td>
</tr>
<tr>
<td>4</td>
<td>the west bank and gaza 23</td>
<td>at the same time the 22</td>
</tr>
<tr>
<td>5</td>
<td>as a result of the 22</td>
<td>in such a way that 22</td>
</tr>
<tr>
<td>6</td>
<td>at the same time the 22</td>
<td>the second half of the 20</td>
</tr>
</tbody>
</table>

Also not surprising is the extreme lack of five word lexical bundles. In the case of both undergraduate and graduate levels, only six five-word lexical bundles were found that met the threshold requirement. All of these lexical bundles necessarily contain a three or four-word lexical bundle, although these smaller lexical bundles are found with
other words as well. What might be surprising is that only one of these lexical bundles is shared between levels, unlike the three and four-word lexical bundles.

The functions of these lexical bundles are also different. While both graduate and undergraduate lexical bundles are text or research oriented, when we look at the lexical bundles in context we can see that the undergraduate lexical bundles are more summative, while the graduate lexical bundles are more analytical. That is to say, the graduate lexical bundles frame an original process of thought, while the undergraduate lexical bundles present “known” information. Consider the examples below, which compare the use of the five-word lexical bundle *at the same time the*, a research-oriented lexical bundle according to Hyland’s (2008) taxonomy.

**Undergraduate**
1. Each pair of s orbitals overlaps to give VI bonding and one antibonding molecular orbital, while *at the same time the* p orbitals overlap to give three bonding and three antibonding molecular orbitals. (chemistry)
2. *At the same time, the* strength of the subtropical jet stream takes a hit because the reversed temperature gradient works against strong westerly flow. (earth science)
3. *At the same time, the* consequences of the Renaissance and the Reformation drastically altered intellectual, political, religious, and social life within Europe. (geography)
4. *At the same time, the* war had forced many European manufacturers to become more efficient and had expanded the demand for automotive and air transport, electrical products, and synthetic goods. (history)
5. *At the same time, the* country is firmly anticommunist. (political science)
6. *At the same time the* demands for constant alertness and for instant obedience are continuously stepped up, and the standards by which the dress and behavior of the recruits are judged become steadily more unforgiving. (sociology)

**Graduate**
1. This would avoid the need for an interfacial layer and *at the same time the* high permittivity of the material could be fully realized. (physics)
2. *At the same time, the* roles of non-state actors in the process of regime formation and policy implementation are increasingly acknowledged. (geography)
3. *At the same time, the* roles of non-state actors in the process of regime formation and policy implementation are increasingly acknowledged. (history)
4. In this way the unity of the thing is preserved, and, *at the same time, the* otherness is preserved outside the thing, as well as outside consciousness. (philosophy)
5. *At the same time, the* sentence indicates how uncertain the researcher is about the estimated quantity of interest. (political science)
6. *At the same time, the* simple fact that these forms of resistance exist suggests there is something out there to avoid and that there are powerful constraining forces at work. (sociology)
Given the function of textbooks, the primary source of text in the undergraduate corpus, and the function of research articles and monographs, this finding is expected, but appears to have been neglected in previous research that discusses function and frequency, but does not discuss polysemy in usage. When meaning was considered, the most common meaning was determined and that was used to decide the function.

4.2.3 Which disciplines use lexical bundles most frequently at the undergraduate and graduate levels?

To compare lexical bundles within disciplines, I chose to only look at four-word lexical bundles. It was clear from the analysis of the corpus as a whole that there were already few five-word lexical bundles, but an overwhelming number of three-word lexical bundles which were often simply part of a larger four-word lexical bundle. There is also a precedent already set for four-word bundle analysis (e.g. Hyland 2008; Cortes, 2004).

As with the analysis of the whole corpus, frequencies were normalized to one million words to maintain the standard set previously, although each discipline only contains around 100,000 words. The cut off for analysis is necessarily higher; 20 times per million words means that the lexical bundle appears in the discipline only two or three times, arguably not frequently at all and almost surely not across several texts. Thus, the cut off for analysis has been raised to a more conservative and yet still low level of 40 times per million, but found in at least three texts, or 20-25% of the texts in the discipline. The results are reported in Table 4.5.
### Table 4.5: four-word bundle frequency information

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Different bundles</th>
<th>Undergraduate</th>
<th>Graduate</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total cases</td>
<td>% total words in bundles</td>
<td>Total cases</td>
<td>% total words in bundles</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td>110</td>
<td>692</td>
<td>2.8%</td>
<td>97</td>
<td>517</td>
<td>2.1%</td>
<td></td>
</tr>
<tr>
<td>Chemistry</td>
<td>272</td>
<td>1,696</td>
<td>7.0%</td>
<td>124</td>
<td>1,066</td>
<td>4.2%</td>
<td></td>
</tr>
<tr>
<td>Earth Science</td>
<td>101</td>
<td>798</td>
<td>3.1%</td>
<td>78</td>
<td>512</td>
<td>1.8%</td>
<td></td>
</tr>
<tr>
<td>Mathematics</td>
<td>375</td>
<td>2,126</td>
<td>9.0%</td>
<td>449</td>
<td>2,672</td>
<td>11.1%</td>
<td></td>
</tr>
<tr>
<td>Physics</td>
<td>225</td>
<td>1,281</td>
<td>5.3%</td>
<td>109</td>
<td>721</td>
<td>2.8%</td>
<td></td>
</tr>
<tr>
<td>Geography</td>
<td>68</td>
<td>589</td>
<td>2%</td>
<td>93</td>
<td>833</td>
<td>2.6%</td>
<td></td>
</tr>
<tr>
<td>History</td>
<td>30</td>
<td>307</td>
<td>0.8%</td>
<td>59</td>
<td>493</td>
<td>1.3%</td>
<td></td>
</tr>
<tr>
<td>Philosophy</td>
<td>114</td>
<td>770</td>
<td>2.7%</td>
<td>113</td>
<td>1,059</td>
<td>2.9%</td>
<td></td>
</tr>
<tr>
<td>Political Science</td>
<td>72</td>
<td>810</td>
<td>2.4%</td>
<td>114</td>
<td>717</td>
<td>2.5%</td>
<td></td>
</tr>
<tr>
<td>Sociology</td>
<td>31</td>
<td>254</td>
<td>0.9%</td>
<td>45</td>
<td>366</td>
<td>1.0%</td>
<td></td>
</tr>
</tbody>
</table>

What immediately is apparent is that some disciplines are more dependent on lexical bundles than others. Mathematics, for example, makes particularly heavy use of lexical bundles, with one tenth of its text on average being part of a frequent four-word lexical bundle, and without considering three-word lexical bundles or p-frames, sets of words with a wild card somewhere within the set (e.g. theorem <wild card> let x). I would hypothesize that if one were to study texts in mathematics and the other hard disciplines using p-frames, one would find an even heavier dependence of formulaic language, as equations, numbers, and symbols are often inserted into the middle of sentences. Mathematics is often noted as a field in which non-native speakers can excel because of the heavy use of numbers, a universal among most languages. However, with a smaller set of vocabulary and heavy use of formulaic language, it may not only be for the sake of numbers that it is easier for non-native speakers to do extremely well.

On the other end of the spectrum are history and sociology, which hardly make use of lexical bundles at either the undergraduate (30 and 31, respectively) or graduate (59 and 45, respectively) level. Indeed, it would seem in general that the soft fields are
less reliant on lexical bundles than the hard fields, although not all hard fields use as many lexical bundles as mathematics.

Another major difference between the hard and soft fields is the inverse relationship they have with the frequency of lexical bundles at the undergraduate and graduate level. The hard fields typically have more lexical bundles at the undergraduate level, the exception being with mathematics, while the soft fields have more at the graduate level. It is unclear why the two areas would use lexical bundles so differently. Hard fields often make use of formulaic language to refer to figures, graphs, charts, and formulae, visual representations found far less frequently in the soft fields. This could account in part for the greater number of lexical bundles in the hard fields, but not for the lower numbers at the graduate level when the primary function of academic writing is to present data. One hypothesis (Xiaofei Lu, personal communication), as yet untested, is that at the graduate level the same lexical bundle used at the undergraduate level might be expressed in several ways, thus lowering the frequency of these lexical bundles potentially below the threshold set in this study. Another hypothesis, again untested, is that data is more readily included within the text of graduate level texts, and that the use of numbers, symbols, and formulae actually break up potential lexical bundles creating p-frames. Both hypotheses may be tested in future research.

The numbers do not tell the whole story, as can be seen in Table 4.6. Consider philosophy, which has essentially the same number of lexical bundles at both levels (114 undergraduate, 113 graduate) and, although the graduate level lexical bundles are used more regularly (2.9% of text compared to the undergraduate 2.7% of texts), one could assume that the two levels are quite similar. However, if we look at the lexical bundles
themselves, we find that only 17 of them overlap. In other words, there are actually 210 lexical bundles represented in philosophy as a whole, only 17 of which occur at least 40 times per million words at both levels, or 8% of the total lexical bundles available in that discipline. Similar numbers are found throughout all of the disciplines. Earth science only shares 3 lexical bundles between levels, and while mathematics shares 36 bundles, they only comprise 5% of the total lexical bundles found repeatedly in mathematics texts. It is history, with its relatively low number of lexical bundles, that has the highest percentage of coverage at 17% with only 13 lexical bundles.

Table 4.6: Shared lexical bundles between undergraduate and graduate levels

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Shared lexical bundles</th>
<th>% of total lexical bundles</th>
</tr>
</thead>
<tbody>
<tr>
<td>biology</td>
<td>14</td>
<td>7%</td>
</tr>
<tr>
<td>chemistry</td>
<td>16</td>
<td>4%</td>
</tr>
<tr>
<td>earth science</td>
<td>3</td>
<td>1.7%</td>
</tr>
<tr>
<td>mathematics</td>
<td>36</td>
<td>5%</td>
</tr>
<tr>
<td>physics</td>
<td>12</td>
<td>4%</td>
</tr>
<tr>
<td>geography</td>
<td>10</td>
<td>7%</td>
</tr>
<tr>
<td>history</td>
<td>13</td>
<td>17%</td>
</tr>
<tr>
<td>philosophy</td>
<td>17</td>
<td>8%</td>
</tr>
<tr>
<td>political science</td>
<td>11</td>
<td>6%</td>
</tr>
<tr>
<td>sociology</td>
<td>5</td>
<td>7%</td>
</tr>
</tbody>
</table>

It is important to caution that while these findings may seem significant, they are based on relatively small corpora. Each discipline only contains around 100 thousand words and thus the data can be easily skewed. However, as preliminary results, it does appear that the lexical bundles used at the undergraduate level are quite different from those used at the graduate level, meaning that herein lies one of the largest differences between the two levels so far.
4.2.4 What are the most frequent lexical bundles in disciplines, and how do they compare between levels?

Although we saw that when we compared the undergraduate and graduate corpora as a whole there were several lexical bundles that were shared by the two levels, we have just discovered that these shared lexical bundles do not carry over to the different disciplines. What we do find, shown in Tables 4.7 – 4.16, is that among the most frequent lexical bundles in each discipline there is occasionally overlap. At best, four lexical bundles are shared by both levels in some disciplines (i.e. biology, mathematics, sociology), but not surprisingly in some cases there were no shared lexical bundles (i.e. chemistry, earth science, physics). What is of particular interest, though, are the lexical bundles that are not shared.

Based on Hyland’s (2008) taxonomy of lexical bundle functions, there are three types of lexical bundle: research oriented, text oriented, and participant oriented. If we look at the undergraduate lexical bundles, we will see that that they are primarily research oriented, and more specifically they are mostly locative, quantitative, descriptive, or topical, but not procedural. On the other hand, if we look at the graduate lexical bundles, they are both research oriented and text oriented, and the research oriented bundles are quite often procedural. Consider the lexical bundles found in earth science (Table 4.9). All but one of the undergraduate lexical bundles are research oriented, the exception being as a result of, a bundle not included as most frequent at the graduate level. At the graduate level, five of the top ten lexical bundles are textual, and of the five remaining research oriented bundles, three are quantitative, compared to none at the undergraduate level, one is process oriented, where no undergraduate ones were, and one is topical.
Given that many graduate level texts are conveying information on a study, we would expect to find such language, but it is a definite distinction between the two levels.

Moreover, the shared lexical bundles are almost all text oriented, with the notable exceptions of *in the united states*, *of the human genome*, and *the end of the*. With respect to *in the united states*, as this study is based on texts required in classes at an American university, it is not surprising to find this lexical bundle prevalent at both levels. *Of the human genome* most likely reflects the current age of biology in which the mapping of the human genome and cloning are the most current innovations. As for *the end of the*, it is technically a research oriented lexical bundle representing location, but it can also be text oriented as a structuring signal, as in *at the end of the chapter*. The categories are not strict; even Hyland blurs the lines by putting *in the present study* as both a research oriented locative lexical bundle and a text oriented structuring signal.

Participant oriented lexical bundles, according to Hyland, are the rarest. That has proven the case with this data as well. While some disciplines do include participant oriented lexical bundles in their most frequent bundles at the graduate level (e.g. political science and sociology), these bundles are missing from the undergraduate level with the exception of sociology (see Table 4.16). Perhaps this is unsurprising as textbooks are not written to convince the reader that a finding in a research project is significant, but rather to inform in what is theoretically an unbiased way.
Table 4.7: Biology

<table>
<thead>
<tr>
<th>rank</th>
<th>Undergraduate</th>
<th>occurs</th>
<th>Graduate</th>
<th>occurs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>of the digestive tract</td>
<td>229</td>
<td>has been shown to</td>
<td>143</td>
</tr>
<tr>
<td>2</td>
<td>of the small intestine</td>
<td>209</td>
<td>on the basis of</td>
<td>143</td>
</tr>
<tr>
<td>3</td>
<td>the end of the</td>
<td>179</td>
<td>as a result of</td>
<td>133</td>
</tr>
<tr>
<td>4</td>
<td>can be seen in</td>
<td>139</td>
<td>in the human genome</td>
<td>123</td>
</tr>
<tr>
<td>5</td>
<td>as a result of</td>
<td>129</td>
<td>of the human genome</td>
<td>123</td>
</tr>
<tr>
<td>6</td>
<td>in the united states</td>
<td>129</td>
<td>in the united states</td>
<td>113</td>
</tr>
<tr>
<td>7</td>
<td>of the human genome</td>
<td>129</td>
<td>the end of the</td>
<td>113</td>
</tr>
<tr>
<td>8</td>
<td>the surface of the</td>
<td>129</td>
<td>in the case of</td>
<td>102</td>
</tr>
<tr>
<td>9</td>
<td>be seen in figure</td>
<td>119</td>
<td>in the presence of</td>
<td>92</td>
</tr>
<tr>
<td>10</td>
<td>in the form of</td>
<td>119</td>
<td>can be used to</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>in the number of</td>
<td>119</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>the length of the</td>
<td>119</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.8: Chemistry

<table>
<thead>
<tr>
<th>rank</th>
<th>Undergraduate</th>
<th>occurs</th>
<th>Graduate</th>
<th>occurs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>as shown in figure</td>
<td>405</td>
<td>in the context of</td>
<td>404</td>
</tr>
<tr>
<td>2</td>
<td>of the periodic table</td>
<td>353</td>
<td>in the case of</td>
<td>345</td>
</tr>
<tr>
<td>3</td>
<td>the number of protons</td>
<td>291</td>
<td>the same procedure was</td>
<td>316</td>
</tr>
<tr>
<td>4</td>
<td>in the periodic table</td>
<td>249</td>
<td>same procedure was used</td>
<td>296</td>
</tr>
<tr>
<td>5</td>
<td>the same number of</td>
<td>249</td>
<td>procedure was used as</td>
<td>276</td>
</tr>
<tr>
<td>6</td>
<td>in the united states</td>
<td>197</td>
<td>which was isolated from</td>
<td>276</td>
</tr>
<tr>
<td>7</td>
<td>is shown in figure</td>
<td>187</td>
<td>confirmed by total synthesis</td>
<td>266</td>
</tr>
<tr>
<td>8</td>
<td>of the chemical elements</td>
<td>187</td>
<td>which is a metabolite</td>
<td>197</td>
</tr>
<tr>
<td>9</td>
<td>can be used to</td>
<td>166</td>
<td>been confirmed by total</td>
<td>178</td>
</tr>
<tr>
<td>10</td>
<td>on the other hand</td>
<td>166</td>
<td>in the presence of</td>
<td>178</td>
</tr>
<tr>
<td></td>
<td>the number of neutrons</td>
<td>166</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.9: Earth Science

<table>
<thead>
<tr>
<th>rank</th>
<th>Undergraduate</th>
<th>occurs</th>
<th>Graduate</th>
<th>occurs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the earth’s surface</td>
<td>314</td>
<td>as a function of</td>
<td>134</td>
</tr>
<tr>
<td>2</td>
<td>in the northern hemisphere</td>
<td>255</td>
<td>are given in table</td>
<td>125</td>
</tr>
<tr>
<td>3</td>
<td>of the earth’s</td>
<td>196</td>
<td>can be used to</td>
<td>125</td>
</tr>
<tr>
<td>4</td>
<td>in the united states</td>
<td>176</td>
<td>as a result of</td>
<td>107</td>
</tr>
<tr>
<td>5</td>
<td>as a result of</td>
<td>157</td>
<td>on the other hand</td>
<td>98</td>
</tr>
<tr>
<td>6</td>
<td>in the southern hemisphere</td>
<td>157</td>
<td>a large number of</td>
<td>89</td>
</tr>
<tr>
<td>7</td>
<td>in the form of</td>
<td>147</td>
<td>is the number of</td>
<td>89</td>
</tr>
<tr>
<td>8</td>
<td>the center of a</td>
<td>137</td>
<td>the land and ocean</td>
<td>89</td>
</tr>
<tr>
<td>9</td>
<td>the end of the</td>
<td>137</td>
<td>on the basis of</td>
<td>80</td>
</tr>
<tr>
<td>10</td>
<td>the surface of the</td>
<td>137</td>
<td>the fact that the</td>
<td>80</td>
</tr>
</tbody>
</table>
Table 4.10: Mathematics

<table>
<thead>
<tr>
<th>rank</th>
<th>Undergraduate occurs</th>
<th>Graduate occurs</th>
<th>Undergraduate occurs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a b and c</td>
<td>304</td>
<td>on the other hand</td>
</tr>
<tr>
<td>2</td>
<td>as shown in figure</td>
<td>220</td>
<td>if and only if</td>
</tr>
<tr>
<td>3</td>
<td>a and b are</td>
<td>199</td>
<td>in the sense of</td>
</tr>
<tr>
<td>4</td>
<td>in terms of the</td>
<td>168</td>
<td>with respect to the</td>
</tr>
<tr>
<td>5</td>
<td>on the other hand</td>
<td>168</td>
<td>let x be a</td>
</tr>
<tr>
<td>6</td>
<td>b and c are</td>
<td>157</td>
<td>in terms of the</td>
</tr>
<tr>
<td>7</td>
<td>the equation of a</td>
<td>147</td>
<td>in this section we</td>
</tr>
<tr>
<td>8</td>
<td>in this section we</td>
<td>136</td>
<td>it is clear that</td>
</tr>
<tr>
<td>9</td>
<td>symmetric with respect to</td>
<td>136</td>
<td>in the case of</td>
</tr>
<tr>
<td>10</td>
<td>in the previous section</td>
<td>126</td>
<td>the proof of the</td>
</tr>
<tr>
<td></td>
<td>in this chapter we</td>
<td>126</td>
<td>the right hand side</td>
</tr>
<tr>
<td></td>
<td>the length of the</td>
<td>126</td>
<td></td>
</tr>
<tr>
<td></td>
<td>with respect to the</td>
<td>126</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.11: Physics

<table>
<thead>
<tr>
<th>rank</th>
<th>Undergraduate occurs</th>
<th>Graduate occurs</th>
<th>Undergraduate occurs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the star's surface</td>
<td>226</td>
<td>as a function of</td>
</tr>
<tr>
<td>2</td>
<td>in our solar system</td>
<td>195</td>
<td>on the other hand</td>
</tr>
<tr>
<td>3</td>
<td>the magnitude of the</td>
<td>185</td>
<td>are shown in fig</td>
</tr>
<tr>
<td>4</td>
<td>the laws of quantum</td>
<td>154</td>
<td>in the limit of</td>
</tr>
<tr>
<td>5</td>
<td>kinetic energy of the</td>
<td>144</td>
<td>for the case of</td>
</tr>
<tr>
<td>6</td>
<td>potential energy of the</td>
<td>144</td>
<td>in the case of</td>
</tr>
<tr>
<td>7</td>
<td>is equal to the</td>
<td>134</td>
<td>as a result of</td>
</tr>
<tr>
<td>8</td>
<td>of the solar system</td>
<td>123</td>
<td>as shown in fig</td>
</tr>
<tr>
<td>9</td>
<td>the laws of physics</td>
<td>123</td>
<td>is shown in fig</td>
</tr>
<tr>
<td>10</td>
<td>the mass of the</td>
<td>123</td>
<td>of the order of</td>
</tr>
</tbody>
</table>

Table 4.12: Geography

<table>
<thead>
<tr>
<th>rank</th>
<th>Undergraduate occurs</th>
<th>Graduate occurs</th>
<th>Undergraduate occurs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>in the united states</td>
<td>353</td>
<td>the politics of scale</td>
</tr>
<tr>
<td>2</td>
<td>at the same time</td>
<td>215</td>
<td>the ways in which</td>
</tr>
<tr>
<td>3</td>
<td>of the world's</td>
<td>215</td>
<td>in the case of</td>
</tr>
<tr>
<td>4</td>
<td>as a result of</td>
<td>181</td>
<td>on the other hand</td>
</tr>
<tr>
<td>5</td>
<td>of the united states</td>
<td>138</td>
<td>as a result of</td>
</tr>
<tr>
<td>6</td>
<td>on the other hand</td>
<td>129</td>
<td>of the politics of</td>
</tr>
<tr>
<td>7</td>
<td>parts of the world</td>
<td>129</td>
<td>as well as</td>
</tr>
<tr>
<td>8</td>
<td>in the northern hemisphere</td>
<td>112</td>
<td>in relation to the</td>
</tr>
<tr>
<td>9</td>
<td>one of the most</td>
<td>112</td>
<td>in terms of the</td>
</tr>
<tr>
<td>10</td>
<td>the middle east and</td>
<td>112</td>
<td>of the study area</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>on the one hand</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>the extent to which</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>the politics of scale</td>
</tr>
</tbody>
</table>
Table 4.13: History

<table>
<thead>
<tr>
<th>rank</th>
<th>Undergraduate occurs</th>
<th>Graduate occurs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>one of the most 124</td>
<td>in the united states 164</td>
</tr>
<tr>
<td>2</td>
<td>the end of the 118</td>
<td>one of the most 144</td>
</tr>
<tr>
<td>3</td>
<td>at the same time 98</td>
<td>at the same time 123</td>
</tr>
<tr>
<td>4</td>
<td>for the most part 91</td>
<td>as well as the 103</td>
</tr>
<tr>
<td>5</td>
<td>in the face of 91</td>
<td>of the united states 96</td>
</tr>
<tr>
<td>6</td>
<td>support for the war 91</td>
<td>second half of the 82</td>
</tr>
<tr>
<td>7</td>
<td>in the united states 85</td>
<td>the second half of 82</td>
</tr>
<tr>
<td>8</td>
<td>the beginning of the 78</td>
<td>in the name of 75</td>
</tr>
<tr>
<td>9</td>
<td>at the end of 72</td>
<td>the history of the 75</td>
</tr>
<tr>
<td>10</td>
<td>for the first time 72</td>
<td>of the nineteenth century 68</td>
</tr>
<tr>
<td></td>
<td>of the seventeenth century 72</td>
<td>the manner in which 68</td>
</tr>
<tr>
<td></td>
<td>on the other hand 72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>the rest of the 72</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.14: Philosophy

<table>
<thead>
<tr>
<th>rank</th>
<th>Undergraduate occurs</th>
<th>Graduate occurs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the state of nature 241</td>
<td>at the same time 344</td>
</tr>
<tr>
<td>2</td>
<td>in the state of 138</td>
<td>on the other hand 248</td>
</tr>
<tr>
<td>3</td>
<td>in the united states 129</td>
<td>on the basis of 179</td>
</tr>
<tr>
<td>4</td>
<td>in the case of 112</td>
<td>in the case of 172</td>
</tr>
<tr>
<td>5</td>
<td>on the basis of 112</td>
<td>one and the same 131</td>
</tr>
<tr>
<td>6</td>
<td>on the spiritual path 112</td>
<td>that is to say 124</td>
</tr>
<tr>
<td>7</td>
<td>on the other hand 103</td>
<td>in the course of 117</td>
</tr>
<tr>
<td>8</td>
<td>to the welfare of 103</td>
<td>in the same way 117</td>
</tr>
<tr>
<td>9</td>
<td>does not mean that 86</td>
<td>with a view to 117</td>
</tr>
<tr>
<td>10</td>
<td>it is not just 86</td>
<td>in so far as 110</td>
</tr>
<tr>
<td></td>
<td>not be able to 86</td>
<td>in the way of 110</td>
</tr>
<tr>
<td></td>
<td></td>
<td>so far as it 110</td>
</tr>
<tr>
<td></td>
<td></td>
<td>the greater part of 110</td>
</tr>
</tbody>
</table>

Table 4.15: Political Science

<table>
<thead>
<tr>
<th>rank</th>
<th>Undergraduate occurs</th>
<th>Graduate occurs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the united states and 283</td>
<td>as well as the 105</td>
</tr>
<tr>
<td>2</td>
<td>in the united states 276</td>
<td>it is easy to 105</td>
</tr>
<tr>
<td>3</td>
<td>at the same time 179</td>
<td>on the other hand 96</td>
</tr>
<tr>
<td>4</td>
<td>to the united states 134</td>
<td>a large number of 87</td>
</tr>
<tr>
<td>5</td>
<td>and the soviet union 127</td>
<td>as a function of 87</td>
</tr>
<tr>
<td>6</td>
<td>the civil rights movement 127</td>
<td>in the case of 87</td>
</tr>
<tr>
<td>7</td>
<td>on the other hand 119</td>
<td>in the context of 87</td>
</tr>
<tr>
<td>8</td>
<td>for the first time 112</td>
<td>are likely to be 79</td>
</tr>
<tr>
<td>9</td>
<td>in the middle east 112</td>
<td>in the sense that 79</td>
</tr>
<tr>
<td>10</td>
<td>the end of the 112</td>
<td>one of the most 79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>the effects of the 79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>the probability of a 79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>the ways in which 79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>there is no reason 79</td>
</tr>
</tbody>
</table>
Table 4.16: Sociology

<table>
<thead>
<tr>
<th>rank</th>
<th>Undergraduate occurs</th>
<th>Graduate occurs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>in the united states 481</td>
<td>in the united states 253</td>
</tr>
<tr>
<td>2</td>
<td>are more likely to 193</td>
<td>to the extent that 116</td>
</tr>
<tr>
<td>3</td>
<td>percent of the time 79</td>
<td>the magnitude of the 108</td>
</tr>
<tr>
<td>4</td>
<td>are less likely to 70</td>
<td>more likely to be 101</td>
</tr>
<tr>
<td>5</td>
<td>at the same time 70</td>
<td>are more likely to 94</td>
</tr>
<tr>
<td>6</td>
<td>on the other hand 70</td>
<td>less likely to be 79</td>
</tr>
<tr>
<td>7</td>
<td>the number of people 70</td>
<td>on the basis of 72</td>
</tr>
<tr>
<td>8</td>
<td>the united states in 70</td>
<td>at the same time 65</td>
</tr>
<tr>
<td>9</td>
<td>and the united states 61</td>
<td>in the case of 65</td>
</tr>
<tr>
<td>10</td>
<td>are more likely than 61</td>
<td>in the laboratory experiment 65</td>
</tr>
<tr>
<td></td>
<td>over the life course 61</td>
<td>on the other hand 65</td>
</tr>
<tr>
<td></td>
<td>the united states the 61</td>
<td>the strength of ties 65</td>
</tr>
</tbody>
</table>
4.3 Summary

We can see from the various forms of analysis that there are in fact differences between the graduate and undergraduate levels. While both levels do make use of a similar number of lexical bundles, lexical bundles are used more frequently at the graduate level. The most frequent lexical bundles at each level are also quite different, even within the same discipline. The lexical bundles at the undergraduate level tend to be more research-oriented while those at the graduate level are both research and text-oriented. These differences can be attributed to discipline as well as the different functions that the texts at these two levels have: one is to primarily instruct while the other is to primarily disseminate research. It should come as no surprise that novice writers at the graduate level so rarely use the lexical bundles that are most frequent in their disciplines if their input previously, even within the same discipline, was at the undergraduate level.
Chapter 5

Grammatical Variation

While EAP has been extensively debated, most research in the area has focused on its pragmatics and overall structure, or gone to the opposite extreme and focused on the building blocks, the words, both individually as vocabulary or together as lexical bundles or clusters. The grammar of academic English has only been the focus of a few studies, notably by Biber and his colleagues.

The Longman Grammar of Spoken and Written English (LGSWE) (Biber, Johansson, Leech, Conrad, & Finegan, 1999) was the first comprehensive look at grammar in academic English, albeit part of a larger study of English in general. The purpose of the study was to create an inclusive description and explanation of American English grammar based, in large part, on Quirk (1985). The authors noted that "the overriding goal [is] to use categories and terms that are familiar and unobjectionable to the widest range of grammar users" (p. 7). The authors also explained that while the terminology was borrowed, the grammar itself was determined empirically, not based on a previous model such as Chomsky’s generative grammar (1956) or minimalist program (1995), or Halliday’s systemic functional linguistics (1973).

The academic prose section of the LGSWE corpus was comprised of 333 journal articles and extracts from 75 academic books, of which only a small margin were written for "a lay audience, including student textbooks" (p. 33), and ranged across 13 disciplines, for a total size of a little over five million words. The academic prose portion
of the LGSWE is comparable to the graduate level of the AUGER corpus. While Biber et al. do not explicitly explain criterion for inclusion in the LGSWE, it is clear that student textbooks were largely excluded, and it is these books that make up the majority of undergraduate input and the undergraduate level of the AUGER corpus.

The Biber et al. study is a comparative one, in which academic prose is one of four registers analyzed, the others being news, fiction, and conversation, the latter section representing spoken English. While there is an in-depth study of lexical bundles in academic prose (see Section 4.1), most of the volume merely references academic prose, giving occasional charts and comments if something seems interesting. Noun phrases are used more in written genres than in spoken genres, where verbs are far more frequent, for example. As such, it is perhaps unsurprising that adjectives are also more common in written registers, and in particular in academic prose. Moreover, the study claims that most noun phrases have some form of modifier, and goes on to give examples, but does not specifically target academic prose.

By far the most comprehensive study of academic English is Biber’s (2006) study of university language, both spoken and written. He included textbooks, academic program brochures, university catalogs, student handbooks, course packs, and course management texts like syllabuses, as well as spoken texts from classroom teaching, office hours, classroom management, study groups, lab and in-class groups and service encounters. In sum, he had approximately one million words in each of the spoken and written registers. While he did take care to include texts from different levels of university study, he did not perform any analysis based on these levels.
With respect to grammar, his analysis focused primarily on word classes (nouns, verbs, adjectives, adverbs) and found that there were greater type variation in nouns and adjectives than in verbs and adverbs, that is to say that the majority of types, whether frequent or infrequent, occurred in the noun phrase, across disciplines and in both classroom teaching and textbooks. However, when he looked at the spoken registers as a whole, verbs were more common than nouns in labs, office hours, study groups, and service encounters, and adverbs share an increase in frequency as well, while nouns remained common across the written registers.

5.1 Analysis of Grammar

The four content parts of speech, nouns, verbs, adjective, and adverbs, provide an excellent base from which to begin a general analysis and comparison of grammar. This section will begin with a general comparison of the content word classes across the two levels. Then we will look at noun frequency and semantic type before comparing the use of pronouns, which are often considered a subclass of nouns, between the two levels and the two groups of fields (soft and hard). Finally, we will address verbs, considering semantic class, tense, aspect, and voice.

In order to conduct these analyses, the corpus must be minimally part-of-speech (POS) tagged. As was mentioned in section 2.2.5, the corpus was tagged using the Stanford log-linear POS tagger (Toutanova et al, 2003) and lemmatized using Morph (Minnen, Carroll, & Pearce, 2001). The 36 POS tags distinguish between word classes and inflected forms (see Appendix A for a complete list of tags).

Additionally, nouns and verbs were semantically tagged based on Biber’s (2006) semantic classes of nouns and Biber et al’s (1999) semantic classes of verbs. Biber
(2006) divided all common nouns, those that occurred 20 times per million words or more, into eight semantic classes. He acknowledged that some nouns have multiple meanings, but that “this perspective is useful for comparing the general patterns of use across registers” (Biber, 2006, p 52). An explanation of each classification can be seen in Table 5.1. Biber’s semantic tags for nouns only accounted for roughly half of the common nouns in the AUGER corpus and thus those nouns that were not accounted for in Biber’s initial tag set were given tags based on Biber’s definitions. I then used a python script to tag nouns in the corpus.

Table 5.1: Types of nouns (Taken from Biber, 2006, pp 248-250)

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animate</td>
<td>humans or animals</td>
<td>woman, family, American, wolf</td>
</tr>
<tr>
<td>Cognitive</td>
<td>mental/cognitive processes or perceptions</td>
<td>fact, reason, reaction, effort</td>
</tr>
<tr>
<td>Concrete</td>
<td>inanimate objects that can be touched</td>
<td>water, body, table, food, leaf</td>
</tr>
<tr>
<td>Technical/concrete</td>
<td>tangible objects that are not normally perceived and/or cannot normally be touched</td>
<td>atom, element, cell, air, chapter</td>
</tr>
<tr>
<td>Place</td>
<td>places, areas, or objects in a fixed location</td>
<td>world, surface, area, region</td>
</tr>
<tr>
<td>Quantity</td>
<td>nouns specifying a quantity, amount, or duration</td>
<td>time, number, energy, year</td>
</tr>
<tr>
<td>Group/institution</td>
<td>nouns that denote a group or institution</td>
<td>people, group, government, society</td>
</tr>
<tr>
<td>Abstract/process</td>
<td>intangible, abstract concepts or processes</td>
<td>state, system, life, war, form, power</td>
</tr>
</tbody>
</table>

Based primarily on core meaning, Biber et al. (1999) divided verbs into seven semantic classes: occurrence, mental, facilitation and causation, existence and relational, communicative, aspectual, and activity. Occurrence verbs are those which “primarily report events (typically physical events) that occur apart from any volitional activity” (Biber et al, 1999, p 364). Mental verbs are those which represent states and activities that are not necessarily volitional and are not actively physical. Verbs of facilitation and causation are used when the subject “brings about a new state of affairs (p 363). Existence and relationship verbs, as indicated by the name, indicate the relationship or state between entities. Aspectual verbs describe the progress of an activity or event.
Activity verbs are those which “denote actions and events that could be associated with choice” (p 361), and communication verbs are a subset of activity verbs specifically detailing forms of communication.

Verbs with multiple meanings were classified based on their most common use. For example, *keep* can have the activity verb meaning of holding on to something, but it is more typically used aspectually, as in continuing to do something. Thus *keep* was classified as an aspectual verb. As with the nouns, frequent verbs in the AUGER corpus that were not included in Biber’s initial list of verbs were manually categorized based on Biber et al.’s definitions, and then a python script was written to tag verbs within the AUGER corpus.

### 5.2. Content Word Classes

Previous studies (e.g. Biber et al., 1999) have shown a preference for complex noun phrases in academic writing, and as such one would expect nouns to be particularly prolific within the corpus, and accordingly adjectives as descriptors of nouns. But these findings are salient primarily in academic writing, and the question is whether the case remains true when analyzing lower level textbooks. Verbs, on the other hand, are particularly common in spoken discourse. Are these introductory textbooks, written to be assessable to young academics, emulating the trends of classroom discourse or academic writing?

Biber (2006) found that nouns were far more common in written registers than they were in spoken registers, where nouns and verbs were more equally distributed. Accordingly, adjectives were more common in written registers, while adverbs were used far more frequently in spoken registers. Domains within his written register (textbooks,
course packs, syllabi, and institutional writing) showed little deviation from this trend. Institutional writing made greater use of nouns than the other three domains, syllabi used adjectives less frequently than the other three domains, and textbooks and course packs used adverbs more than syllabi and institutional writing.

Figure 5.1: Comparison of content word classes by level (per 1 million words)

Comparing undergraduate and graduate domains, we find very little variation, and similar results to those published by Biber. While the difference between the nouns and adverbs is not statistically significant, the greater use of verbs at the undergraduate level ($\chi^2=232.1, p <.001$) and the greater use of adjectives at the graduate level ($\chi^2=280.7, p <.001$) is. It can be hypothesized that the texts at the undergraduate level are written to be more accessible, perhaps even more informally or colloquially, which would explain the greater use of verbs while text at the graduate level will indeed make greater use of complex noun phrases, including greater use of adjectives.

5.3 Nouns

Nouns, at both the undergraduate and graduate level, are by far the most frequently occurring content word class, accounting for 291 (undergraduate) and 285
(graduate) words in every thousand, over twice as frequent as the next most frequent word class, verbs.

Table 5.2: Frequency of nouns by level

<table>
<thead>
<tr>
<th></th>
<th>Undergraduate</th>
<th>Graduate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>tokens</td>
<td>types</td>
</tr>
<tr>
<td>&gt;200</td>
<td>128,005</td>
<td>251</td>
</tr>
<tr>
<td>20-200</td>
<td>149,065</td>
<td>2,106</td>
</tr>
<tr>
<td>&lt;20</td>
<td>69,840</td>
<td>19,245</td>
</tr>
<tr>
<td>Total</td>
<td>346,910</td>
<td>21,602</td>
</tr>
</tbody>
</table>

As can be seen in Table 5.2, extremely frequent nouns, those occurring more than 200 times per million words, were expectedly few in number. There were more nouns that occurred extremely frequently and infrequently, fewer than 20 times per million, at the graduate level, but there were also more noun types in general at the graduate level. For more on specific differences in vocabulary, including nouns, see Chapter 3.

With respect to semantic classification of nouns, Biber (2006) found that abstract/process nouns were by far the most common across registers, accounting for at least 50% of nouns across disciplines. Abstract/process nouns were also more common in engineering and business textbooks than in other textbooks. Technical-concrete and concrete nouns were particularly common in natural science and engineering textbooks, where animate nouns were notably rare, while mental nouns were found to be more common in business and humanities texts.

In contrast, in the AUGER corpus abstract/process nouns only accounted for close to 50% of nouns at the graduate level in the soft-pure disciplines (49.3%), and at the undergraduate level of the same only accounted for 40.3% of nouns. There was a similar difference between levels with the hard-pure fields; the undergraduate level accounted for 29.0% of abstract/process nouns while the graduate level accounted for 37.5%. Concrete
and technical nouns were much more common in the hard-pure fields than the soft-pure fields, both concrete and technical nouns are more common at the undergraduate level (15.9% and 25% respectively) than at the graduate level (11.2% and 19.4%). Animate nouns were rare at both levels for the hard-pure fields. In the soft-pure fields, animate nouns were far more common at the undergraduate level than at the graduate level (14.7% and 9.1%), and the same could be said about nouns of place (13.3% and 8.7%).

Figure 5.2: Nouns by group
Looking at the whole of the undergraduate corpus compared to the whole of the graduate corpus, we find that the differences in use of abstract, animate, group, cognitive, concrete, and place nouns are all extremely statistically significant \( (p < .001) \), but that the use of quantity and technical nouns is not. On the other hand, if we compare the whole of the undergraduate corpus to just the hard-pure fields in the graduate corpus, only the concrete nouns are not extremely statistically significant \( (p < .05) \). Comparing undergraduate texts to the soft-pure fields at the graduate level shows that only animate and group nouns are not at all statistically significant, while the other types of nouns are all extremely significantly significant. While these comparisons continue to uphold discipline variation, they also definitively show the differences between graduate and undergraduate texts.

These differences are not altogether surprising. The use of more abstract/process nouns at the graduate level and more concrete and technical/concrete nouns at the undergraduate level of hard-pure fields harkens back to purpose. Undergraduate textbooks serve to define and explain known phenomena, while at the graduate level there is more theory and speculation.

Consider the following examples. Abstract/process nouns are bolded, concrete and technical/concrete nouns are underlined, and other nouns are italicized.

1. In other words, the sediments lithify and become sedimentary rocks. If the continents were at sea level, there would be virtually no driving force for weathering and erosion, and thus no rock cycle. (undergraduate – earth sciences)

2. Plutonium-239 is one of the by-products present in the expended fuel rods. It is formed by absorption of a neutron by uranium-238, followed by two successive beta emissions. (undergraduate – chemistry)
3. Furthermore, the remaining vegetated habitat often contains low plant diversity as a result of erosion, trampling, pollution, invasion or cultivation of a few normative species, and many other human disturbances.

(graduate – biology)

4. The discussion on the effect of slow growth rate on the charge-carrier mobility and the absorption spectra of the films follows.

(graduate – physics)

In the first two examples processes are being explained using concrete and technical nouns while the second two examples, somewhat denser in the sense that each contains only one verb phrase, use more abstract/process nouns as entities within a description.

Consider also the top 15 most frequent abstract/process nouns at each level, given below in Table 5.3. First, we can see that at the undergraduate level certain words are used extremely frequently, but the numbers quickly drop off, while on the graduate side the decrease is far less dramatic. We can also see some similarities; state and system are both in the top three. On the other hand, model, the most frequent graduate abstract noun, does not make the top 15 list for undergraduate abstract nouns; effect is used twice as often at the graduate level, and result almost twice as much. Nouns like model, effect, result, and study are all indicative of experimental or theoretical research while nouns like life, war, and law are concepts that can be discussed. We can see a general trend that with the exception of the top three most frequent abstract nouns at the undergraduate level, abstract nouns are simply not used as frequently as they are at the graduate level, even though, interestingly enough, there were exactly 805 frequent abstract/process nouns at both the graduate and undergraduate levels.
Personal pronouns are a closed, subclass of nouns used to reference a noun or noun phrase. Much of the previous research on pronouns in academic writing has focused on the pragmatic function (Harwood, 2005; Hyland, 2001; Kuo, 1999) or author identity (Hyland, 2002; Tang & John, 1999; Vassileva, 1998). These studies have included research articles, student writing or both, but have not included textbooks. Even Biber’s (2006) study only glosses over pronouns, focusing on first and second person pronouns and emphasizing their existence in spoken academic contexts. First and second pronouns are also “relatively common in written course syllabi and other written class materials, where they serve a similar directive function” (p. 51), but further elaboration as to which class materials is not included.

However, pronouns can be a tricky class of words for the novice writer. A common complaint that instructors who teach undergraduate courses requiring essays make is that students misuse pronouns. Advice on the use of pronouns in academic

<table>
<thead>
<tr>
<th>Rank</th>
<th>Undergraduate</th>
<th>Frequency</th>
<th>Graduate</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>state</td>
<td>1,895</td>
<td>model</td>
<td>1,583</td>
</tr>
<tr>
<td>2</td>
<td>example</td>
<td>1,508</td>
<td>state</td>
<td>1,288</td>
</tr>
<tr>
<td>3</td>
<td>system</td>
<td>1,199</td>
<td>system</td>
<td>1,171</td>
</tr>
<tr>
<td>4</td>
<td>way</td>
<td>1,095</td>
<td>effect</td>
<td>1,166</td>
</tr>
<tr>
<td>5</td>
<td>life</td>
<td>1,045</td>
<td>result</td>
<td>1,126</td>
</tr>
<tr>
<td>6</td>
<td>war</td>
<td>914</td>
<td>study</td>
<td>1,027</td>
</tr>
<tr>
<td>7</td>
<td>law</td>
<td>835</td>
<td>example</td>
<td>985</td>
</tr>
<tr>
<td>8</td>
<td>point</td>
<td>761</td>
<td>structure</td>
<td>984</td>
</tr>
<tr>
<td>9</td>
<td>pressure</td>
<td>759</td>
<td>process</td>
<td>935</td>
</tr>
<tr>
<td>10</td>
<td>form</td>
<td>737</td>
<td>way</td>
<td>848</td>
</tr>
<tr>
<td>11</td>
<td>change</td>
<td>707</td>
<td>form</td>
<td>845</td>
</tr>
<tr>
<td>12</td>
<td>power</td>
<td>692</td>
<td>function</td>
<td>801</td>
</tr>
<tr>
<td>13</td>
<td>problem</td>
<td>655</td>
<td>condition</td>
<td>767</td>
</tr>
<tr>
<td>14</td>
<td>result</td>
<td>604</td>
<td>point</td>
<td>760</td>
</tr>
<tr>
<td>15</td>
<td>process</td>
<td>567</td>
<td>change</td>
<td>683</td>
</tr>
</tbody>
</table>
writing varies. The writing program at Dartmouth, an Ivy League university in the United States, has published online tips for writing in college that say that it is acceptable to use I and you in academic writing as long as “you use them with care” (Gocsik, 2005). The Purdue OWL, an often cited writing source for students, says that while using I is often discouraged, if the use of I is “vital” it may be used sparingly. However, the use of we is discouraged as it is vague. You should only be used when writing direct instructions, and even then it should be avoided (Cimasko, 2012), a somewhat disingenuous piece of advice as the page giving advice on how to write in “North American Higher Education” uses you 77 times. Hoyt (2004) agrees with the Purdue OWL with respect to the use of you, but suggests graduate students use we in place of I in research writing, even if the text is written by only one author. She also recommends that students speak with their advisor or mentor about the norms in their discipline, a point that Swales and Feak (2009) also make. With respect to third person pronouns, only the Purdue OWL included a discussion on avoiding gender bias.

If the argument is made that student writing is influenced by the input students receive, the use of pronouns at the two levels might provide some insight into why students use pronouns the way they do. It may also determine whether using first person and second person pronouns is, in fact, acceptable in academic writing.

Pronoun usage at the two levels is, in fact, distinct. Overall, pronouns are used more often in undergraduate texts (23049 per million words versus 19549 at the graduate level), but first person pronouns are more common at the graduate level. Second and third person pronouns, on the other hand, are far more frequent at the undergraduate level, as is shown in Figure 5.3.
While the number of pronouns used varies between the levels, if we look at the difference between hard-pure and soft-pure disciplines, we see that there is an even greater distinction (Figure 5.4). The first person singular (*I, me, my, myself*) is almost exclusively used in the soft fields, while the first person plural (*we, us, our, ourselves*) is used in both the soft and hard fields, but still favors the soft fields.
Second person (you, your, yourself) is far more common in the undergraduate texts, but also favors some soft-pure fields. There are two instances of you in hard-pure graduate texts, both in the same text within the same sentence, and it is used in hypothetical quoted speech:

But from the point of view of quantum interference, it is not counterintuitive. One could ask, Would you make a tunnel, (that is, an interference) and then go through it, or would you first go through it, and then make it? (graduate – physics)

The instances of second person in the graduate soft-pure fields are, with the exception of the history texts, quoted speech. History is a noted exception in which many of the texts include journal entries, letters, and treatises; the genre of writing is arguably not “academic” and is more conducive to the use of second person. In contrast, at the undergraduate level the second person is commonly used to, as Hyland (2001) would call it, bring in the reader. At the undergraduate level we see the subject you overtly used as a means to personalize the text for the reader. Consider the examples below.

1. As you can see, the volume of the formerly hypertonic solution has increased (because its solutes cannot diffuse out). (undergraduate – biology)

2. You can see good examples of such landforms at Shiprock, New Mexico (Fig. 6.11b), and at Devil's Tower, Wyoming (Fig.9.26d). (undergraduate – earth sciences)

3. If you recall that as little as parts-per-million level of impurities can change the conductivity of silicon by a millionfold, it becomes clear why clean rooms are necessary. (undergraduate – chemistry)

4. Even when you're using your computer for some seemingly nonmathematical activity, components known as gates, based on logical operators that you'll learn about in this chapter, are making it all possible. (undergraduate – mathematics)
5. If you stand on a mattress, Earth pulls you downward, but you remain stationary. (undergraduate – physics)

6. You are driving down the interstate, on your way home from school on vacation. You think you packed enough food for the trip. (undergraduate – geography)

In the first example, instead of writing the more impersonal “as can be seen” the author instead uses “as you can see”, as is also shown in the second example. One could imagine that in the third example, were the text written for a more expert audience, the author might have written “recall that as…”, using the imperative. The final three examples all posit hypothetical situations with which the reader can identify and, in theory, draw connections to the content of the text. Texts at the graduate level rarely do the same, and in those rare occasions might opt to use the more formal one, as in “one might argue that water…” (graduate – geography), or use the inclusive we, as in “We can again imagine Franklin and Weberif…” (graduate – history), where the author is including him or herself along with the reader. Alternatively, like example three, the author might use the imperative, as in “consider, now, any two arbitrary…” (graduate – sociology).

The gendered third person pronouns (he, him, his, himself, she, her, herself) also prefer the soft fields, and are more common at the undergraduate level. There are at least three explanations: topic, form, and manner. The soft fields, including philosophy, history, political science, sociology, and geography, refer to humans regularly, be it as entities within a society or figures of some significance, far more than the hard-pure fields like chemistry, physics, math, and earth science, which tend to focus more on the physical world. Gendered pronouns are used to refer back to a human in some context,
thus giving the soft-pure fields a distinct advantage. Moreover, at the graduate level soft-pure texts tend to include more explanatory literature reviews than hard-pure texts. The few gendered pronouns used at the graduate level in the hard disciplines were found referring to scholars of past research that was being detailed and specifically targeted by the new research, but such a review was only found in a handful of texts. Finally, undergraduate texts are written to be accessible to the lay person and often include anecdotes, history, and referential information, even in the hard-pure fields.

Below we see an introduction to the laws of motion in an introductory physics book followed by an excerpt from a graduate level physics text. References to Newton are bolded while other pronouns are italicized.

1. The relation between a force and the acceleration it causes was first understood by Isaac Newton (1642-1727) and is the subject of this chapter. The study of that relation, as Newton presented it, is called Newtonian mechanics. We shall focus on its three primary laws of motion. … Before Newton formulated his mechanics, it was thought that some influence, a "force," was needed to keep a body moving at constant velocity… (undergraduate-physics)

2. As a consequence, gravity lives in more than four dimensions at macroscopic distances leading to potentially measurable deviations from Newton’s inverse square law at sub-mm distances. (graduate-physics)

The introductory text begins using Newton’s full name and includes his year of birth and death before going on to mention him several more times, even beyond the text segment included here. The second example makes only one mention of Newton in the entirety of the text, and it is assumed that the reader not only knows which Newton, Isaac as opposed to Huey or Wayne, and that the reader is also familiar with his laws.
5.5 Verbs

5.5.1 Semantic Classes of Verbs

Biber (2006) used semantic verb classes, explained in section 5.2, excluding existence/relational, to analyze verb type across registers in university language. His most striking finding was the difference in use between writing and speaking. As has already been discussed, spoken language makes greater use of verbs than written; however, the spoken registers also relied heavily on activity and mental verbs, more so than the written registers. Mental verbs were particularly marked as in some areas, such as study groups and office hours, their use was more frequent than even activity verbs. Aspectual, occurrence, causative, and communication verbs were used similarly in both registers.

While Biber et al. (1999) classify the most frequent verbs, the lists were created from common verbs found across four domains: academic prose, conversation, newspapers, and fiction, and only included 338 total verbs. As such, many (1077) of the frequent, and some extremely frequent, verbs in the AUGER corpus were unclassified. To increase the coverage and accuracy of this part of the study, those verbs that were found at least 20 times per million words were manually classified in addition to the Biber et al. (1999) list. The results are reported in Table 5.4.

<table>
<thead>
<tr>
<th>Category</th>
<th>Under</th>
<th>Grad</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>activity</td>
<td>37,089</td>
<td>29,924</td>
<td>put, follow, send, produce, open</td>
</tr>
<tr>
<td>aspectual</td>
<td>2,263</td>
<td>1,557</td>
<td>start, stop, end, begin, continue</td>
</tr>
<tr>
<td>communicative</td>
<td>8,605</td>
<td>8,691</td>
<td>say, propose, suggest, insist, argue</td>
</tr>
<tr>
<td>existence/relational</td>
<td>9,812</td>
<td>11,042</td>
<td>seem, look, indicate,</td>
</tr>
<tr>
<td>facilitation/ causation</td>
<td>5,153</td>
<td>4,019</td>
<td>help, allow, cause, affect</td>
</tr>
<tr>
<td>mental</td>
<td>16,024</td>
<td>15,706</td>
<td>think, suggest, consider, hope</td>
</tr>
<tr>
<td>occurrence</td>
<td>8,410</td>
<td>6,001</td>
<td>become, change, develop, occur</td>
</tr>
</tbody>
</table>
In the AUGER corpus, the type of verbs used between levels was not even, in part because the total number of verbs used was not even; verbs were more frequent at the undergraduate level than at the graduate level. Activity, aspectual, and occurrence verbs were far more frequent in the undergraduate corpus than in the graduate corpus, while communicative and existence/relational verbs were more frequent in the graduate corpus. Activity verbs were used far more than mental verbs in both corpora, although there does seem to be a trend to use mental verbs more in the graduate corpus than in the undergraduate, as shown in Figure 5.5.

Figure 5.5: Percent of tokens for semantic classes of verbs by level

As we did before, if we compare the tokens in the soft fields and hard fields at the two levels we see more variation (Table 5.5). While both graduate hard and soft disciplines used fewer verbs than their undergraduate counterparts, undergraduate hard fields used far fewer verbs than the graduate soft fields. Activity verbs comprised more of the verbs at the undergraduate level than those at the graduate level for both groups, as did facilitation and occurrence verbs to a lesser degree. Existence (e.g. seem, appear,
include) and mental verbs (e.g. see, find, consider, think), on the other hand, accounted for more verbs at the graduate level. Communicative verbs (e.g. say, argue, suggest) were more frequent in the soft fields, and more frequent at the graduate level in general. We should expect these results as communicative, existence, and mental verbs are all used in referring to past research, argumentation, and the presenting of results – elements common in research papers and less so in textbooks. Aspectual verbs (e.g. begin, continue, keep), while used less at the graduate level, accounted for the fewest verbs for both groups and the smallest change.

Table 5.5: Semantic classes of frequent verbs

<table>
<thead>
<tr>
<th>Category</th>
<th>Undergraduate hard token</th>
<th>Type</th>
<th>%</th>
<th>Graduate hard token</th>
<th>Type</th>
<th>%</th>
<th>Undergraduate soft token</th>
<th>Type</th>
<th>%</th>
<th>Graduate soft token</th>
<th>Type</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>activity</td>
<td>19,869</td>
<td>285</td>
<td>44.9</td>
<td>16,346</td>
<td>236</td>
<td>41.5</td>
<td>24,328</td>
<td>321</td>
<td>40.6</td>
<td>200,045</td>
<td>291</td>
<td>37.5</td>
</tr>
<tr>
<td>aspectual</td>
<td>1,135</td>
<td>22</td>
<td>2.6</td>
<td>715</td>
<td>16</td>
<td>1.8</td>
<td>1,582</td>
<td>21</td>
<td>2.6</td>
<td>1,110</td>
<td>21</td>
<td>2.1</td>
</tr>
<tr>
<td>communicative</td>
<td>3,751</td>
<td>52</td>
<td>8.4</td>
<td>3,575</td>
<td>52</td>
<td>9.1</td>
<td>6,566</td>
<td>82</td>
<td>11.0</td>
<td>6,828</td>
<td>85</td>
<td>12.8</td>
</tr>
<tr>
<td>existence</td>
<td>5,346</td>
<td>55</td>
<td>12.1</td>
<td>5,566</td>
<td>49</td>
<td>14.1</td>
<td>6,400</td>
<td>65</td>
<td>10.7</td>
<td>7,274</td>
<td>68</td>
<td>13.6</td>
</tr>
<tr>
<td>facilitation</td>
<td>2,278</td>
<td>33</td>
<td>5.2</td>
<td>2,104</td>
<td>35</td>
<td>5.1</td>
<td>3,850</td>
<td>47</td>
<td>6.4</td>
<td>2,791</td>
<td>45</td>
<td>5.2</td>
</tr>
<tr>
<td>mental</td>
<td>7,164</td>
<td>97</td>
<td>16.2</td>
<td>7,971</td>
<td>88</td>
<td>20.3</td>
<td>11,905</td>
<td>133</td>
<td>19.9</td>
<td>11,411</td>
<td>132</td>
<td>21.3</td>
</tr>
<tr>
<td>occurrence</td>
<td>4,723</td>
<td>68</td>
<td>10.7</td>
<td>3,166</td>
<td>53</td>
<td>8.0</td>
<td>5,273</td>
<td>64</td>
<td>8.8</td>
<td>4,051</td>
<td>58</td>
<td>7.6</td>
</tr>
</tbody>
</table>

5.5.2 Verb Tense, Aspect, and Voice (Inflections)

Biber (2006) noted that the past tense was marked in all registers and disciplines, but that in education and humanities textbooks the past tense was used comparatively frequently, more than 30% and 40% respectively, but in different ways. Education used the past tense narratively, while the humanities used it to report past events. Engineering, on the other hand, only used the past tense in 5% of its textbook verb phrases, overwhelmingly preferring to use the present tense.

Also in Biber (2006), in 95% of all written verb phrases, simple aspect was used. Perfect and progressive verb phrases comprised the remaining verb phrases, with progressive aspect being slightly more common. Similarly, active voice was by far the
most common voice, at 95% in spoken registers and around 80% in written registers. Engineering textbooks used passive voice the most, and natural science textbooks to a lesser extent, while humanities textbooks used passive voice the least.

These findings are both surprising and expected, given Tarone et al.’s (1998) study of the use of active and passive voice in astrophysics journal articles. The impetus for the study was based on earlier findings by other scholars showing that the use of passive voice was particularly common in scientific texts, and indeed some found that passive voice was used more than active voice. Tarone et al. used a notably small corpus, two articles, for their analysis, but found that indeed passive voice was not used more than active voice, and only accounted for 11-28% of voice in verbs, depending on how and what one was counting. This discrepancy needs to be addressed. If passive voice is not a tool used in undergraduate writing, should it be a focus in an undergraduate EAP classroom? Or, it may be the case that contrary to Biber’s findings, passive voice will be found to be prevalent even in textbooks.

Extracting verbs and verb forms is fairly simple, but not trivial. While the Penn Treebank tagset does code for past tense, past participles, and -ing forms, it is the combination of these forms with forms of be and have that express voice and aspect. For example, to extract passive voice constructions, the script must look for some conjugation of the verb to be + (adverb) + (negation) + (adverb) + past participle, where the adverbs and negation in parenthesis are optional.

The findings in the AUGER corpus are mostly consistent with Biber’s findings, but show a distinct difference in the popularity of the past tense between disciplines. History is of particular interest as at both the graduate and undergraduate level the present
tense is actually the marked tense. In undergraduate history texts the past tense is used 81.5% of the time, while in graduate texts it is used 61.7% of the time. Philosophy, which historically has also been considered among the humanities with history, only used the past tense 26.8% of the time at the undergraduate level and 14.5% of the time at the graduate level, suggesting that it is possible that Biber’s results may have been skewed by the probable inclusion of both history and art history into his humanities subcorpus. Which disciplines are included in which subcorpus in the Y2KSWAL corpus is not actually delineated. The past tense is also used differently at the two levels in philosophy. At the graduate level primary sources are used almost exclusively as the required reading for courses, while at the undergraduate level some primary sources may be used but quite often there is a textbook or secondary source also assigned. Thus, at the graduate level we actually see the past tense used far more narratively while at the undergraduate level it is in fact used to report historical events. Nonetheless, the use of the past tense is far more common at the undergraduate level across all of the disciplines in the soft-pure fields than at the graduate level (see Figure 5.6).
We find the same statistically significant trend in physics (27.2% undergraduate, 13.7% graduate, $\chi^2=90.1$, $p < .001$) and to a lesser degree in mathematics (8.9% undergraduate, 6.9% graduate, $\chi^2=20.9$, $p < .001$) and earth sciences (25.3% undergraduate, 21.3% graduate, $\chi^2=28.8$, $p < .001$). However, the opposite is the case in chemistry (12.0% undergraduate, 38.3% graduate, $\chi^2=1298.9$, $p < .001$) and biology (14.2% undergraduate, 30.5% graduate, $\chi^2=519.51$, $p < .001$), where the past tense is significantly more frequent at the graduate level. See Figure 5.7 for comparison. The difference in past tense usage, whether it was more or less, was statistically significant in every discipline in both the hard and soft fields.
Figure 5.7: Past tense and present tense in the hard-pure fields

Some of this variation in tense and voice between levels may be attributed to the style guides used by the respective disciplines. The American Psychological Association (APA, 2009), used by education and many social sciences, and the related American Sociological Association (ASA, 2010) recommend using the past tense when writing a literature review, while the Chicago style (2010), used in history, philosophy, and geography, and the related American Political Science Association (APSA, 2006), suggest using present tense. Clearly, history did not use the present tense much, but political science and geography both use the past tense only 11.9% of the time at the graduate level, the least of the five soft disciplines. Sociology, on the other hand, uses it 27% of the time, second only to history in the soft disciplines.

In the hard disciplines, the Council of Science Editors (CSE), used in biology, recommends using the past tense (2006) which corresponds with biology using verbs in the past tense 30.5% of the time. Chemistry, with 38.5% of its verbs in the past tense,
uses the American Chemical Society (ACS) style, and in the guide it specifically states that authors should not only use the past tense when possible but also use passive voice (ACS, 1986). Perhaps this is why at the graduate level we see 39.6% of the verb phrases written in passive voice, the most of any of the hard or soft disciplines (see Figure 5.8). Past tense usage is not addressed in the United States Geological Survey (USGS, 1991), the American Institute of Physics (AIG, 1997), or the American Mathematical Society (AMA, 2010), but all three offer advice on the use of passive voice. The AIP specifically notes the usefulness of the passive voice, but discourages its use if it “makes the syntax inelegant” (p. 14). The USGS also notes that the passive is necessary, but does say that the active voice is preferred when it makes sense to do so, which is fairly arbitrary advice. The AMA recommends using the active voice if the actor is present, and it is possible that is why mathematics has the smallest use of passive voice in the hard disciplines at 15.7%. Indeed, it appears that, with the exception of mathematics (20.8% passive at the undergraduate level, 15.7% at the graduate, $\chi^2=64.3, p < .001$) and political science (14.9% passive at the undergraduate level, 14.7% at the graduate), passive voice is used far more often at the graduate level, but even then far more in the hard-pure disciplines than in the soft-pure. Again, style guides may account for this difference as both the APA and Chicago manuals encourage the use of active voice, the APA going so far as to call the passive voice “non-preferred”. The difference between the use of active and passive voice at the graduate and undergraduate level was statistically significant ($p < .001$) in every discipline except political science.
Figure 5.8: Percent of passive voice by discipline and level

With respect to aspect, simple aspect is unsurprisingly the unmarked form. However, contrary to Biber’s (2006) findings, only 83.9% of tensed verb phrases were simple in undergraduate soft disciplines, while perfect and progressive aspect were equally represented at 8.0%. At the graduate level, 90.3% of tensed verbs were simple, and perfect aspect was clearly preferred in other forms; 7.8% of verbs used perfect aspect while only 1.7% used progressive. Indeed, the hard-pure disciplines at both levels used perfect aspect far more than progressive aspect when simple aspect was not used; however, at the undergraduate level 94.1% of tensed verbs used simple aspect while 92.9% did at the graduate level. Perfect progressive aspect appeared marginally and at best accounted for .1% of tensed verb phrases at any given level or discipline.

5.6 Summary

While some of the differences between the graduate and undergraduate levels may be minute, there are some differences that are quite significant. In general, nouns,
verbs, adjectives, and adverbs are used equally at the two levels, but the types of nouns and verbs are different. Abstract/process nouns, while common at both levels, were significantly more frequent at the graduate level than at the undergraduate level, while concrete and concrete/technical nouns were more common at the undergraduate level. Second person pronouns were used primarily at the undergraduate level with the exception of reported speech and hypothetical questions, and even then second person pronouns were rare at the graduate level. First person pronouns were more frequent at the graduate level, but singular forms tend to be in the soft-pure fields. Pronouns, in general, were more frequent in undergraduate texts, perhaps in part due to the explanatory nature of the texts. With respect to verbs, activity, aspectual, and occurrence verbs were more common at the undergraduate level, while communicative and mental verbs were more common at the graduate level.

Tense and voice displayed the greatest differences between the two levels. Past tense is used significantly more at the undergraduate level than at the graduate level in the soft-pure fields, while being used more at the graduate level than the undergraduate level in the hard-pure fields. Overall, 30.8% of the undergraduate verb phrases were written in the past tense, twice as much as graduate level physics, mathematics, political science, philosophy, or geography. Passive voice was used 20.0% of the time in verb phrases at the undergraduate level, half as frequently as it was used in graduate level chemistry, 50% less frequently than it was used in graduate level physics, earth science, and biology, and more frequently in general than it was used in any of the soft fields with the exception of geography.
In other words, while the vocabulary may vary slightly between the levels, it is how the parts of speech are used that truly varies.
Chapter 6

Conclusions

In this chapter I will discuss the findings from the previous chapters on vocabulary, lexical bundles, and grammar. I will then address the implications these findings have on EAP programs, American universities, and students themselves. I will conclude with the limitations of this study and future directions that this research may take.

6.1 Vocabulary

Chapter three focused on vocabulary at the undergraduate and graduate levels. Words are the smallest part of discourse that can stand alone. Moreover, academic vocabulary has already been widely studied, although not with respect to academic level. There were several research questions that were drawn from the literature:

1. What types are frequent and what is their range?
2. How many words are high frequency and how many are low frequency?
3. What is the range of words used in the different disciplines and levels, and how do they compare?
4. What are the most frequent types in the two different levels, and do they overlap?
5. If the AWL worked well for a general corpus of academic English, how would it fare on a corpus specifically geared toward undergraduate input?

Regarding the first four questions, the most frequent words are shared between the undergraduate and graduate levels if the two are compared as a whole, but these
words include many function words including determiners, prepositions, and pronouns, words that would not even be considered specifically academic but rather would be found across a range of registers. When we compare disciplines, we find that there are consistently more types at the undergraduate level than at the graduate level in every discipline except chemistry. If we emulate an undergraduate student’s discipline expectations compared with a graduate’s, then we need to compare the whole of the undergraduate corpus to single disciplines at the graduate level. Unsurprisingly, the whole of the undergraduate corpus displayed greater lexical variation than single disciplines at the graduate level with the exception of history. However, comparing the type/token ratio of the whole undergraduate corpus to the whole graduate corpus, the graduate corpus showed a higher lexical variation. In other words, there is greater discipline variation at the graduate level, which accounts for the greater lexical variation when the whole graduate corpus is compared to the whole undergraduate corpus. For undergraduate students this finding is both positive and negative.

Since undergraduate students are expected to take a large variety of courses, the needed vocabulary is naturally going to be greater than if one were to take courses in only one discipline. Fortunately, vocabulary overlaps more between disciplines at this level reducing the overall load. Unfortunately, frequent lemmas did not have an even distribution across the disciplines, even at the undergraduate level, with 50% of frequent lemmas appearing in only one discipline frequently.

The question remains as to why there is typically greater lexical diversity at the undergraduate level than at the graduate level. One possible explanation is that the courses used in creating the undergraduate corpus were introductory or survey courses,
courses that typically cover a wide breadth of topics. Graduate courses, on the other hand, have a narrower focus, even in survey courses, so the lexical variation might be less even though there is more discipline specific language. Another factor might be tied to register. While textbooks at the undergraduate level are still written media, they are written to address a completely novice audience whereas the texts at the graduate level are written to address a more expert audience. Thus textbooks are written more informally, colloquially, and draw on examples that resonate with a wider variety of readers.

Question five essentially asks if the AWL, a list of academic words designed to facilitate vocabulary acquisition for future college students, is useful for undergraduates planning to attend American universities, particularly given the results discussed above? Essentially, no. While the GSL and AWL combined covered more of the undergraduate corpus than the graduate corpus, the AWL only covered 7.8% of the types in the undergraduate corpus, well under the 10.5% covered at the graduate level, 10.6% in Hyland and Tse (2007), and 10% in Coxhead (2000). Graduate students entering certain fields might find the AWL very useful as it covered 13.3% of sociology texts, 14.0% of political science texts, and 14.1% of geography texts, but history and philosophy each had less than 7% coverage. Nevertheless, if students already know which fields they intend to study in graduate school, it makes sense to use a more discipline specific vocabulary list over the AWL given that the AWL was created from a corpus of 28 disciplines, even in disciplines where the AWL had respectable coverage.

Undergraduate students benefited from the GSL the most, and particularly the first 1000 word families, which accounted for 73.3% of the total undergraduate corpus.
At present, there may be no quick solution as to what a student should study next. However, it is important to remember that undergraduates, native speakers and non-native speakers alike, are often in the same situation of having to learn new, academic language their freshman year. It is entirely possible that Dudley-Evans and St. John (1998) had it right when they suggested EAP focus on the “common core”, at least for future undergraduate students. Indeed, some instruction on the culture of American universities would seem prudent, particularly given the American-centric nature of the content in courses and the general education model, a model that is not universal around the world.

6.2 Lexical Bundles

In chapter four the frequency of co-occurring sets of words, called n-grams or lexical bundles, were considered since it is the logical next step in a comparative analysis of two levels as we expand the study of words to words in their environment. Again, the research questions were drawn from previous research on lexical bundles, research that had not yet included any analysis between levels.

1. What is the frequency of lexical bundles in general across corpora?
2. Which lexical bundles are frequent at the graduate and undergraduate levels?
3. Which disciplines use lexical bundles frequently?
4. What are the most frequent lexical bundles in disciplines and how do they compare between levels?
5. What is the function of these lexical bundles?

Again we find a disparity between the undergraduate and graduate levels. In response to the first two questions, while similar numbers of lexical bundles were found
at each level, graduate level texts had far greater diversity in their co-occurring word groups. The functions of the four-word lexical bundles also varied, with those at the undergraduate level favoring Hyland’s (2008) research-oriented type and graduate level including far more text-oriented lexical bundles. Even those lexical bundles which were frequent at both levels were not necessarily used the same way, and context played a role in overall meaning and function. Thus, teaching certain lexical bundles, although more contextualized than individual vocabulary words, could potentially exclude the function of the lexical bundle.

It is very clear that, when we compare lexical bundles in specific disciplines, the hard-pure fields made greater use of frequent lexical bundles overall, with mathematics particularly relying on them. Oddly, there was less reliance on lexical bundles at the graduate level in the other four hard-pure fields (biology, chemistry, earth science, physics), although this could potentially be explained by varying means of expressing the same concept or the incorporation of more data within the text itself at the graduate level. Although each discipline did show a very slight uptick in usage at the graduate level, the soft-pure fields generally made little use of lexical bundles, a result that conflicts to some extent with Hyland’s (2008) findings. In his study he considered applied linguistics and business studies social sciences, and both made greater use of lexical bundles than did biology. While this study did not include applied fields, in which applied linguistics and business studies would have been included, biology does not appear to use lexical bundles more or less than the soft-pure fields as a whole.

With question four we found that the number of shared frequent lexical bundles between the levels is not encouraging. History, with only 17% of the frequent lexical
bundles overlapping between levels, had the highest percentage of commonality. Earth science had only 1.7%, or 3 frequent four-word lexical bundles. Even mathematics, the field most reliant on formulaic language, only shared 5% of its frequent four-word lexical bundles at the graduate and undergraduate levels.

Here again we have significant differences between the language used at the undergraduate level and that used at the graduate. Based on the results from this study, it appears that teaching the same lexical bundles to all EAP students regardless of their future collegiate plans would be uneconomical.

Simpson-Vlach and Ellis (2010) created an Academic Formulas List (AFL) that separates frequent lexical bundles by function and distinguishes between primarily spoken and written language. The researchers used MiCASE (Simpson et al. 2002) and the academic speech files from the BNC (British National Corpus, 2006) for the spoken registers and parts of the BNC’s written academic corpus and Hyland’s (2004) corpus for the written registers. Both MiCASE and the BNC cover a wide range of levels in academia, and in fact the BNC includes some non-university texts within its spoken education portion. However, the use of the BNC provides a British slant to the results, good for an international English context but less helpful for American university contexts. Also, the use of Hyland’s corpus skews the data away from undergraduate input.

Even if the compilation of the corpus was perfectly even in representing both undergraduate and graduate levels in the American academy, the question has to be asked if it would be worth teaching such a list in EAP classes given the results from the analyses of lexical bundles in this dissertation. Not only is there discipline variation at
both levels, but the reliance on lexical bundles appears to be greater at the undergraduate level where students will be taking classes across the curriculum, not just in one field. Moreover, the lexical bundles that appear most frequently at the undergraduate level are not the same ones used at the graduate level.

This leads to another issue that affects native and non-native speakers alike. If the lexical bundles used at the undergraduate level are not those used at the graduate level, how is it that graduate students beginning their program should know the formulaic language associated with their field, particularly if they have received their bachelor’s degree at an American university? To be fair, it is possible that courses taught later at the undergraduate level, in the third and fourth years which were not represented in the AUGER corpus, might better emulate the graduate input. Even so, the vastly different input in the first year of graduate and undergraduate study suggests that while lexical bundles are important for native-like fluency, they may not be important in a general EAP classroom.

6.3 Grammar

In chapter five the analyses considered grammar in a broad sense, focusing on content word classes and the major functions of those classes using Biber (2006) as its primary model. The major research questions included:

1. What are the most frequent word classes at the two levels?
2. What types of nouns are used frequently at the two levels?
3. What types of verbs are used at the two levels?
4. How are personal pronouns used at the graduate and undergraduate levels?
5. How might verb tense, aspect, and voice vary between graduate and undergraduate academic English?

Regarding question one, analyzing the four major content word classes (nouns, verbs, adjectives, and adverbs) shows that undergraduate text use verbs significantly more than graduate, while adjectives are used more at the graduate level than at the undergraduate. These results point to two distinct differences between the two levels. First, graduate texts are tightly packed, written by and for experts and future experts in the field. The greater use of adjectives may indicate more complex noun phrases, a lexico-grammatical feature found to be common in previous research on academic English (e.g. Halliday & Martin, 1993). On the other hand, the undergraduate texts were written to be accessible to a novice, thus these texts might be written more colloquially, more like an instructor teaching rather than an expert reporting.

To answer questions two and three, nouns, the most abundant word class, are also treated differently at the two levels. Only concrete nouns were found not to be significant when comparing the undergraduate corpus with both the hard-pure and soft-pure graduate sub-corpora. Abstract nouns showed a pronounced increase at the graduate level and cognitive nouns also increased marginally. On the other hand, technical, animate, and place nouns accounted for a larger portion of the nouns at the undergraduate level. Because undergraduate texts typically serve in an expository capacity, the larger proportion of such nouns is logical while graduate texts, more theory or research oriented, would have greater need of abstract and cognitive nouns. Verb types show a similar disparity. Activity, existence, and occurrence verbs were frequent at the undergraduate
level, while communicative, facilitation, and mental verbs accounted for more verbs at the graduate level.

Personal pronouns, addressed in question four, had some of the greatest and most significant variation. The second person personal pronoun, you, displays the starkest contrast between the two levels of text; despite being the least used personal pronoun at either level, in undergraduate texts it is used repeatedly to address the reader while in graduate texts, with the exception of history, it is used exclusively in hypothetical and quoted speech. Perhaps this discrepancy is one reason why undergraduate novice writers use second person address. Textbooks and other assigned materials would be, for most freshmen and sophomores, considered formal or academic text. It is reasonable to suggest that these same students would use these texts as models in their own academic writing. The other personal pronouns showed less pronounced variation. As a whole, pronouns were more frequent in the undergraduate corpus, but first person and third person neutral were used more at the graduate level in the soft-pure fields.

Finally, question five addresses the grammatical inflections carried by verbs. Aspect, previously found to be overwhelmingly simple (Biber, 2006), had a bit more variation at the undergraduate level where 7% of tensed verbs used perfect aspect and another 7% used progressive. At the graduate level, simple aspect accounted for 90.3% of tensed verbs and perfect aspect accounted for 7.8%. While there was no significant change in the use of perfect aspect, simple aspect was clearly the preferred aspect over progressive.

Present tense was also the preferred tense of verbs, with the clear and obvious exception of history. Even so, the use of past tense varied significantly between the two
levels in almost every discipline, and appeared to be heavily influenced at the graduate level by the standard formatting conventions adopted by each discipline. If present tense was the preferred tense at the graduate level, then past tense appeared more frequently at the undergraduate level and vice versa, although past tense was rarely preferred at the graduate level. Beyond the writing conventions, another reason for this discrepancy is that undergraduate texts often contextualize subject matter, explaining the history of a theory or topic. Recall the example of Newton's laws in an undergraduate textbook. The author took the time to explain who Newton was before going into detail about the laws themselves. In graduate texts this knowledge is often assumed and therefore not included. Although previous studies relevant to the current research may be addressed, particularly in the soft fields, writing conventions of many disciplines require these studies be referred to in the present tense.

The use of passive was also significantly different in most disciplines between the two levels, and again writing conventions seemed to dictate the norms at the graduate level while the undergraduate level seemed to no follow any specific guide. The general trend showed that passive voice was used far more at the graduate level, with exceptions in mathematics and political science. While no discipline used passive voice the majority of the time, graduate chemistry did use it almost 40% of the time in tensed verb phrases, and the graduate hard sciences in general, again with the exception of mathematics, used passive voice in 30% or more of their tensed verb phrases.

This does lead one to question why writing conventions adopted by each discipline are not followed at the undergraduate level. It is possible that undergraduate texts are not truly considered formal academic writing and thus are not held to the
conventional rules of a discipline. This would seem a logical conclusion given that many undergraduate textbooks do not overtly cite resources, use the second person pronoun, and use tense, aspect, and voice differently, yet there has been little outcry from the academic community or from editors of textbooks to conform to those standards used at the graduate level. Likely this is a result, again, of function. The purpose of undergraduate texts is to inform, to begin inquiry, and to explain concepts that are likely new only to the learner, not necessarily to the field. On the other hand, graduate level texts often do impart knowledge that is new to the field, at least at the time of publication, and while graduate students are beginning their own inquiry they do so starting at a much higher point that undergraduate students.

6.4 Implications for Research

Previous studies of academic English have focused on student work, research articles, monographs, or a combination of resources including, in addition to the other sources, textbooks, syllabuses, and institutional writing. One reason for this variety is that academic English has not been absolutely defined; it is a fairly nebulous concept involving language used in the academy, produced by students, produced by faculty, used by students, and used by faculty. Thus when a study is completed it is necessary to explicitly explain the purpose of the study and to meet that purpose.

We have to consider fundamentally why we are doing the study, and what is the real purpose of the text (based on spoken or written data) included in the corpus? If we want to make claims about a specific discipline, is it fair to include undergraduate level textbooks, research articles, student writing, and monographs since all of these sources of text have different audiences? At the same time, is it fair to make sweeping claims about
a discipline based solely on research articles when there is so much more published
within the discipline? Comparing the two levels within a single discipline demonstrated
that there is clearly variation, sometimes extremely significant variation. A corpus of
research articles will allow for results on published expert writing, but it is important to
be aware that such writing is not that which low level undergraduates will read nor is it
what students should be expected to produce. Indeed, outside the academy it likely will
not even be read, which brings into the picture an entirely different set of texts involving
non-fiction for non-academics, often written by academics themselves. Defining
academic English has been an issue for a while, but defining what should be included as
discipline specific text is also debatable as texts will vary depending on intended
audience.

Corpus-based research allows for very detailed analyses of language and how it is
used. However, the findings of such studies are only reliable in contexts that mirror the
corpus from which the data was drawn. In other words, while a corpus may contain a
segment of language defined as academic English, it may not actually provide results that
are generalizable to all academic English. Moreover, as a general limitation of corpus
analysis, multimodal analyses are limited. Yet figures, pictures, sidebars, and the general
way such added information is organized on a page are all part of the representation of
academic English, much as Powerpoint or Prezi slides are part of teaching and
presenting. These elements are difficult to include in a strictly corpus driven study.

6.5 Implications for Pedagogy

With respect to pedagogy, there are several implications. Hyland (2006) argues
for an across the board discipline specific approach to academic English. Indeed, there
appears to be a recent surge in research detailing the differences between disciplines (see Kuo, 1999; Hewings, 2004; Hyland, 2002a, 2002b; Biber, 2006 for a few examples). While practice often lags behind research, the calls for changes in institutions that offer EAP courses have begun. Should these institutions listen?

There is no simple answer to that question. The findings of this dissertation support both a discipline specific approach as well as a general approach; the difference is in the level. As was mentioned in the introduction, there are very few institutions that offer graduate and undergraduate specific EAP courses. Those that make a distinction between the two usually group students together for most classes and then offer a couple of classes specific to graduate students. Very few, as yet, offer discipline specific EAP coursework, although it has been implemented in some schools in various ways. One of the most common is to have students attend courses in their discipline, and then work with an ESOL instructor on individual assignments, for example.

One argument used to counter a discipline specific approach is that institutions that offer EAP courses lack the resources to adequately teach thusly (Bhatia, 2002). ESOL teachers are trained in teaching a second language, but not in the intricacies of individual disciplines. Moreover, teaching several individual curricula in one classroom is taxing and extremely difficult to do well, while hiring enough teachers to teach discipline specific classes is not cost effective. Indeed, financial gain is often a greater driving force in institutions than sound pedagogic practice.

The findings in this study appear to complicate the situation. Undergraduate students planning to attend American universities may actually benefit from the status quo, but graduate students would be far better served with discipline specific EAP
courses. In other words, an institution cannot completely convert to discipline specific coursework as it would be a disservice to undergraduates, but the status quo is a disservice to graduate students.

One of the limitations of this dissertation is the fact that the corpus is comprised only of expected readings. Thus, pedagogical implications can only reliably extend to the receptive skill of reading and cannot necessarily assume implications for writing (see Hedgcock & Atkinson, 1993). In some ways this limitation further justifies a non-discipline specific approach for graduate students. While the vocabulary and reading skills acquired would likely be useful, it might have little significant effect on their writing skills which is often an important means of assessment in graduate school.

Regardless of how the institutions choose to address the problem, those that admit both graduate and undergraduate students do need to take heed that some form of variation in instruction would be useful.

6.6 Limitations

There are two more serious limitations with this research and two smaller issues. The first is the relatively small corpus that was used for the studies. While the total number of tokens exceeded two million, the subcorpora consisted of only 100,000 tokens, a relatively small number particularly for analyses of vocabulary. A comparison of the whole graduate corpus to the whole undergraduate corpus would likely yield reliable results, but the smaller discipline specific subcorpora need to be expanded in future work. For now, these subcorpora provide insight into what might be worth researching further.

The second major limitation is the general focus of the dissertation. A comparative analysis could include a number of studies, and many of those studies were
not included or conducted within the space of these pages. A study of clause structures (e.g. subordinate, relative) or sentence complexity would add to the literature as would a comparison of noun phrases. Hedging and stance features are popular research subjects in academic English because of their prevalence, and again a study of how these features vary between levels might inform how students form their arguments in their own papers. While the exclusion of these studies is a limitation of the dissertation, it does not mean that these studies will not be conducted at a future time.

Another issue relates to discipline selection. The corpus was constructed using five disciplines from the hard/pure and soft/pure fields for a total of ten disciplines. The pure fields were a requirement based on the nature of the research; it was a comparison of graduate and undergraduate first or second year texts and applied fields simply did not have many courses offered early in the career of an undergraduate. For that reason engineering and education, to name just two vast fields, were excluded entirely from the study. Many students come to the United States to study engineering both at the undergraduate and graduate levels. While the research does address many of the classes that an undergraduate would take, no graduate level engineering texts were included.

Finally, undergraduate students moving from EAP classes to university classes will begin with introductory or freshmen courses. The nature of graduate school, with only two years of coursework instead of four, dictates that coursework at that level is not delineated quite so strictly. The corpus included texts from 100 and 200 level course at the undergraduate level and 500 courses at the graduate level to reflect these coursework expectations. This means that 300 and 400 level courses were completely excluded. Ideally, upper level undergraduate courses would have been included; however, students
transition from EAP courses to university classrooms are unlikely to begin with 300 and 400 level classes so they were not included for the purpose of this study.

6.7 Future Research

The limitations of this dissertation in many ways serve as a starting point for future research. While it was clear there were many differences between the levels, the primary purpose of this dissertation was to consider a graduate student in a single discipline to an undergraduate who must take classes in many disciplines. While some analyses were made within the disciplines, they were fairly minimal.

I look forward to expanding the corpus considerably, discipline by discipline. If possible, I would like each subcorpora to minimally have 500,000 tokens as the reliability of the results would improve significantly. Many of the same analyses that were already conducted could be redone to ensure reliability. For example, preliminary results in the lexical bundle analysis showed that there were very few overlapping lexical bundles at the graduate and undergraduate levels even between disciplines, but a larger corpus would verify those results. The expanded subcorpora also allow for studies of intradisciplinary variation, another area that appears to be under-researched given how reliant previous corpora have been on specific types of texts for overarching conclusions on discipline practices.

One analysis that was completely excluded from this dissertation was the use of p-frames, or chunks of language in which a single entity may be substituted (i.e. on the one hand and on the other hand). It seems like mathematics is a field that makes extreme use of p-frames, but without having done the analysis, it is hard to say for sure.
Finally, this study was always a building block to a larger study of how students are socialized into academic writing. We saw in the study of grammar that passive voice and past tense are both used differently between levels, and it stands to reason that student writing might also reflect these differences. A common complaint among instructors requiring formal papers in undergraduate classes is students’ inappropriate use of personal pronouns, including the use of the second person. Yet, textbooks make extensive use of the second person. Could students be emulating their textbooks? A second issue, not addressed in this dissertation, is citation. The texts used at the undergraduate level, with the notable exception of sociology, typically did not cite any resources, or if they did, included superscripted numbers with references at the end. Could this be one of the reasons undergraduate students struggle to use appropriate citation methods?

There are two approaches to answering these questions. The first is to build a corpus, or use a pre-existing corpus, of student writing. This corpus should not discriminate between good and bad writing, but rather include the range that would be found in a classroom. If only the best texts are selected, then the research only studies what those who have been socialized well can produce and excludes the average student, arguably the more important target. The second approach is not corpus-based but rather involves a more ethnographic study. While we can assume that students are reading what is assigned, the fact is much of their information may come from other sources.

6.8 Concluding remarks

This study served to fill a gap in the literature that had previously, at best, been only marginally addressed. While there is extensive research on academic English and a
growing body of research on interdisciplinary differences, no research has yet been published that distinguishes the experiences of EAP students as they enter the university for either graduate or undergraduate study. While MiCUSP does include upper division undergraduate student papers with graduate student papers, there has been no comparison drawn and no lower level student papers were included. The Y2KSWAL corpus made a point to include lower division and upper division textbooks as well as graduate textbooks, but then failed to do any analysis based on those levels.

In this study I have begun to differentiate between graduate and undergraduate academic English by focusing on the expected readings of students as they enter their respective programs. I have proven that there are, in fact, differences, although a larger corpus is needed to confirm these results. Looking forward, this study serves as a starting point for further research into the differences between the two levels of academic study as well as how students are socialized into the writers they become. These studies are especially important now as American universities accept and even recruit more and more international students for both levels of study.
References


Arnaud, P. (1992). Objective lexical and grammatical characteristics of L2 written


Chen, Q. and Ge, G. (2007). A corpus-based lexical study on frequency and distribution


Harwood, N. (2005). 'We do not seem to have a theory... the theory I present here attempts to fill this gap': Inclusive and exclusive pronouns in academic writing. Applied Linguistics. 26 (3), 343-375.


Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English*


M. McCarthy (Eds.), *Vocabulary: Description, Acquisition, and Pedagogy.*
Cambridge: Cambridge University Press.

Oxford: Oxford University Press

Oxford University Press.


Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition, and Pedagogy.* Cambridge: Cambridge University Press.

http://www.corpus.bham.ac.uk/PCLC/NesiStudentWriting.doc

*Harvard Educational Review, 66* (1). Retrieved from
http://wwwstatic.kern.org/filer/blogWrite44ManilaWebsite/paul/articles/
A_Pedagogy_of_Multiliteracies_Designing_Social_Futures.htm


Appendix

The Penn Treebank Tag Set

1. CC  Coordinating conjunction
2. CD  Cardinal number
3. DT  Determiner
4. EX  Existential there
5. FW  Foreign word
6. IN  Preposition or subordinating conjunction
7. JJ  Adjective
8. JJR  Adjective, comparative
9. JJS  Adjective, superlative
10. LS  List item marker
11. MD  Modal
12. NN  Noun, singular or mass
13. NNS Noun, plural
14. NP  Proper noun, singular
15. NPS  Proper noun, plural
16. PDT  Predeterminer
17. POS  Possessive ending
18. PP  Personal pronoun
19. PPS  Possessive pronoun
20. RB  Adverb
21. RBR  Adverb, comparative
22. RBS  Adverb, superlative
23. RP  Particle
24. SYM  Symbol
25. TO  to
26. UH  Interjection
27. VB  Verb, base form
28. VBD  Verb, past tense
29. VBG  Verb, gerund or present participle
30. VBN  Verb, past participle
31. VBP  Verb, non-3rd person singular present
32. VBZ  Verb, 3rd person singular present
33. WDT  Wh-determiner
34. WP  Wh-pronoun
35. WP$ Possessive wh-pronoun
36. WRB  Wh-adverb

(Marcus, Santorini, & Marcinkiewicz, 1993)
VITA

Tracy S. Davis
Department of Applied Linguistics
Pennsylvania State University
305 Sparks Building
State College, PA 16802

Education
2012 Ph.D., Applied Linguistics, Pennsylvania State University
Dissertation: A Corpus-Based Comparison of Undergraduate and Graduate Academic English
Committee: Xiaofei Lu (advisor), Suresh Canagarajah, Robert Schrauf, and Xiaoye You
2005 M.A., Linguistics (concentrations in Linguistic Theory and TESOL), Syracuse University
Adviser: Jaklin Kornfilt
2001 B.S., Spanish (concentration in Education), Pennsylvania State University

Publications

Academic Positions
2012 – present Assistant Professor, Educational Linguistics, Central Michigan University, Mt Pleasant, MI
2011 – 2012 Postdoctoral Fellow, Educational Linguistics, Central Michigan University, Mt Pleasant, MI
2010 - 2011 Teaching Assistant, Linguistics and Applied Linguistics, Pennsylvania State University, University Park, PA
2007 – 2009 Assistant to the Editor, TESOL Quarterly, TESOL, University Park, PA
2006 – 2007 Teaching Assistant, Applied Linguistics, Pennsylvania State University, University Park, PA
2005 – 2006 Instructor, EFL, Shantou University, Shantou, China
2005 Instructor, ESOL, English Language Institute, Syracuse, NY
2004 Fluency Instructor, ESOL, English Language Institute, Syracuse, NY
2003 – 2005 Teaching Associate, Syracuse University, Syracuse NY
2001 – 2003 Spanish and ESOL Teacher, Warren County Middle School, Warrenton, NC