2016-07-01

A Corpus-Based Comparison of the Academic Word List and the Academic Vocabulary List

Jacob Andrew Newman
Brigham Young University

A Corpus-Based Comparison of The Academic Word List and The Academic Vocabulary List

Jacob Andrew Newman

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Dee Isaac Gardner, Chair
Mark Davies
K. James Hartshorn

Department of Linguistics and English Language

Brigham Young University

July 2016

ABSTRACT

A Corpus-Based Comparison of The Academic Word List
and The Academic Vocabulary List

Jacob Andrew Newman
Department of Linguistics and English Language, BYU
Master of Arts

Research has identified the importance of academic vocabulary (e.g., Corson, 1997; Gardner, 2013. Hsueh-chao & Nation, 2000). In turn, many researchers have focused on identifying the most frequent and salient words present in academic texts across registers and presenting these words in lists, such as The Academic Word List (AWL) (Coxhead, 2000). Gardner and Davies (2014), recognizing the limitations of the AWL, have developed a new list known as The Academic Vocabulary List (AVL). This present study examines the appearance of the 570 AWL word families and the top 570 AVL word families in the Academic Textbook Corpus (ATC) – a 1.9-million-word corpus created from three middle school, three high school, and three college level textbooks from the disciplines of American history, mathematics, and physical sciences. The study determined (1) word families from both the AWL and the AVL found in the ATC, (2) words families unique to the AWL in the ATC, (3) word families unique to the AVL in the ATC, and (4) characteristic differences between the AWL and AVL unique word families. The results suggest that the AWL and AVL capture high frequency academic word families that are salient across a variety of academic disciplines and grade levels, but the AVL provides a greater number of unique frequent core academic word families.

Keywords: AWL, AVL, academic vocabulary, corpus

ACKNOWLEDGEMENTS

I am grateful for the many people who have assisted me throughout my graduate school experience. Writing a master's thesis is not an easy task, but I have been grateful for those who have supported me throughout this culminating experience.

First and foremost, I would like to thank my Heavenly Parents for the blessing of pursuing a master's degree in TESOL. I would also like to thank my earthly parents, Sandy and Helen Newman, who have supported me and the rest of my wonderful siblings in our pursuit of higher education. They inspired us to do our best.

I am also thankful to my many close friends who have supported me throughout the writing process, especially Christian Larsen and Kyra Nelson, who both read my thesis probably more than they would have liked to. I am also particularly grateful for the support of my chair, Dr. Dee Gardner, who has been an inspiration to me since my time as an undergraduate at BYU. Through his mentorship and guidance, I have become a better researcher and writer. I am also grateful to my other committee members, Dr. Mark Davies and Dr. James Hartshorn, who have been supportive of my research and whose insights have been extremely valuable.

TABLE OF CONTENTS

LIST OF TABLES

PREFACE

TESOL MA guidelines at Brigham Young University have indicated the preference for some theses to be prepared as manuscripts for submissions to academic journals. Therefore, this thesis was prepared as a manuscript to be submitted to an academic journal in the field of TESOL or applied linguistics. Most journals in the field of TESOL and applied linguistics have two common requirements: (1) manuscripts should be prepared according to the Publication Manual of the American Psychological Association 6th Edition, and (2) manuscripts should have between 6,000 and 8,500 words. The final draft of this thesis has 9,116 words. Therefore, this thesis would require minimal revisions to be submitted for publication at a future date. This research would be of particular interest to *TESOL Quarterly* or *Applied Linguistics.*

CHAPTER ONE

INTRODUCTION

English language learners (ELLs) who are learning vocabulary for academic purposes and English language teachers who help prepare ELLs for academic contexts both struggle with the issue of vocabulary. Without adequate vocabulary, many ELLs feel ill-equipped to use English in academic settings (Corson, 1997; Gardner, 2013; Hsueh-chao & Nation, 2000). Likewise, English language teachers often struggle to know what words they should teach their learners. Some of these teachers have turned to research to answer a difficult question: Is there a core set of academic vocabulary that students will use across disciplines? In other words, is there a set of vocabulary that is common to all academic disciplines such as the humanities and the sciences? Researchers in applied linguistics and teachers alike have struggled with this question. In order to deal with this conundrum, corpus linguists have studied large corpora of academic texts to determine if there is a core set of academic vocabulary that spans the disciplines. In turn, these researchers have created word lists that they assert are representative of actual academic text (Campion & Elley, 1971; Ghadessy, 1979; Praninskas, 1972).

The most famous of these, the Academic Word List (AWL) (Coxhead, 2000) has been used in English for Academic Purposes (EAP) settings for over 15 years. Many have asserted that this word list is representative of academic text (Coxhead, 2011; Wang, Liang, & Ge, 2008; Ward, 2009.) However, other researchers have noted limitations in the AWL. In fact, some have proposed that the AWL is insufficient in measuring academic vocabulary and its particular senses across all disciplines (Chen & Ge, 2007). In addition to these criticisms, some researchers have identified flaws in the AWL (Gardner & Davies, 2014) and have proposed that the AWL

might not be representative of actual academic text. As a response to these concerns, Gardner and Davies have proposed another option for a list that purports to cover a greater breadth and depth of academic vocabulary: the Academic Vocabulary List (AVL).

The present study examines and compares the AWL and the AVL in a real-world context: textbooks across several disciplines and grade levels. These include middle school American history, math, and physical science textbooks; high school American history, math, and chemistry textbooks; and college-level American history, math, and physical science textbooks. Together, these texts constitute the Academic Textbook Corpus (ATC). This research aims to examine and compare the AWL and the AVL in the ATC from both a quantitative and qualitative perspective. At the outset, it is crucial to recognize important limitations of attempting to compare these two lists based on the construct of word families. Specifically, there is a tendency to exaggerate frequency counts when word families are used as the unit of measurement. This limitation will be discussed in depth later in the thesis.

Some may argue that a comparison between the AWL and the AVL is unfair because the AWL only considered word families beyond the first 2,000 most frequent words of the language as determined by the General Service List (GSL) (West, 1953), but as several experts have pointed out, the AWL actually contains high frequency word families from the top tiers of more modern corpora than the GSL (Cobb, 2010; Gardner & Davies, 2014; Hancioğlu, Neufeld, & Eldridge, 2008; Nation, 2004, 2008; Neufeld, Hancioğlu, & Eldridge, 2011; Schmitt & Schmitt, 2012). Perhaps even more importantly, the AWL purports to be a list of core academic vocabulary, and should therefore be held to that standard.

CHAPTER TWO

LITERATURE REVIEW

Importance of Vocabulary

Research has demonstrated the importance of vocabulary knowledge for both native speakers and English language learners (ELLs). To succeed academically, native speakers need to have sufficient vocabulary to meet the challenges of using English in academic contexts (Beck, Perfetti, & McKeown, 1982; Biemiller, 1999; Biemiller & Slonim, 2001; Chall, Jacobs, & Baldwin, 1990; Hart & Risley, 2003). These native learners have the benefit of time and exposure to acquire these words and to improve their vocabulary knowledge. Beyond the general academic success noted by researchers, others have recognized a reciprocal relationship between vocabulary knowledge and the development of academic reading skills (Biemiller, 2003; Corson, 1997; Nagy & Townsend, 2012). Academic vocabulary knowledge, in turn, is a determining factor in performance on academic gate-keeping exams such as the ACT, SAT, GRE, and GMAT—tests that determine future academic and professional success (Gardner, 2013).

ELLs face particular challenges in acquiring sufficient vocabulary to understand written texts and achieve similar academic and professional success. This daunting task is compounded by the fact that they do not have the luxury of time or repeated language exposure like their native-speaking peers. While these ELLs might have developed Basic Interpersonal Communicative Skills (BICS) in a relatively short time, they spend five to seven years developing Cognitive Academic Language Proficiency (CALP), including the academic literacy skills required to succeed in academic contexts (Cummins, 1979). One reason for this difficulty in developing academic literacy skills is that learners need 95-98% vocabulary knowledge of a

given text to achieve basic reading comprehension (Hsueh-chao & Nation, 2000; Nation, 2001). Therefore, understanding of a great volume of vocabulary is essential to succeed in academic contexts.

Core High-Frequency Vocabulary

Learning high frequency general vocabulary, or a core set of general high frequency words, is the first vocabulary challenge facing ELLs. In order to assist ELLs in achieving this difficult task, West (1953) created a word list known as the General Service List (GSL). The words for the GSL were drawn from a 2.5 million-word corpus. West's selection of these words was based primarily on frequency, with the assumption that these words were essential for ELLs to know. Other factors that were considered included the universality (the words are used across countries where English is the primary medium of communication), utility (the words cover a broad range of genres), and usefulness (the words are useful when attempting to define other lexical items) of the words selected (Gilner, 2011). The 2,000 headwords include other words that constitute a loosely defined "word family," or "base forms plus inflected forms and transparent derivatives" (Gardner, 2007, p.245). For example, the headword *act* includes inflected verb forms (*acts, acting, acted)* as well as derived forms (*action, active, actor).* Because of its impact, the GSL has been extensively examined to determine how representative it is of written texts (Carter, 2012; Engels, 1968; Hirsh & Nation, 1992; Nation, 2001; Nation, 2004; Nation & Kyongho, 1995; Richards, 1974; Sutarsyah, Nation, & Kennedy, 1994). Additionally, the word family paradigm set the stage for much of the research involving pedagogical word lists, as many future word lists would use this design.

Another key issue for the current study is that the GSL is not an attempt to directly address the academic vocabulary needs of ELLs. Other researchers (e.g., Coxhead, 2000) realized this deficiency and began to specifically target academic vocabulary in order to determine how to best assist ELLs in academic settings.

Academic Vocabulary

Nagy and Townsend (2012) suggest at least four characteristics of academic English that make it difficult for ELLs, as well as native speakers: First, academic vocabulary contains many Latin roots (e.g., *acquire, diverse*) and Greek roots (e.g., *analysis, economics*). Second, academic words tend to be more morphologically rich, with a plethora of derivational affixes added to create words. For example, the verb *assess* includes possible derivational variations such as *assessable, assessment, assessments, reassess, reassessed, reassessing, reassessment,* and *unassessed*. Many of these new words that are present in academic contexts are nominalizations, but there are also many more conceptually difficult adjectives than in general English. These nominalizations and other structures often represent grammatical metaphor, or the substitution of one grammatical form for another in order to compress information. Third, the information presented through academic vocabulary is often very dense (e.g., *photosynthesis, mitosis, revolution,* and *emancipation).* Finally, academic vocabulary expresses concepts that are more abstract than general high frequency vocabulary (e.g., *metacognition*, *paradigm*, *medium, aggregate*). These vocabulary features cause ELLs difficulty when they are trying to understand academic texts, such as textbooks and research articles.

Academic Word Lists

In order to assist ELLs with this academic vocabulary challenge, researchers have often turned to pedagogical word lists to assist them in prioritizing vocabulary learning and instruction. There have been various approaches in the creation of academic word lists. Without the assistance of computers, several researchers began to develop academic word lists in the 1970s. Four landmark studies provided the foundation for work in this area. Campion and Elley (1971) along with Praninskas (1972) devised academic word lists based on relatively small corpora, in which the words were counted by hand. Studies conducted by Lynn (1973) and Ghadessy (1979), on the other hand, based the contents of their academic word lists on notes made in textbooks by students, indicating words that they were not familiar with. Using the lists from these pioneering studies, Xue and Nation (1984) created a large-scale academic word list, which they named the *University Word List* (UWL). This list was widely used for several years in a variety of teaching and research contexts.

The Development of the Academic Word List (AWL)

Because the UWL was simply an amalgamation of earlier academic word lists, researchers realized the need for a stronger methodology in the creation of a more representative list of academic words. One of the most influential endeavors in this regard is Coxhead's (2000) Academic Word List (AWL). Better equipped with technology and a larger, more modern corpus of academic materials, Coxhead was able to create an academic word list that was more representative of actual academic text, using the following research questions as a guide:

1. Which lexical items occur frequently and uniformly across a wide range of academic materials but are not among the first 2,000 words of English as given in the GSL (West, 1953)?
2. Do the lexical items occur with different frequencies in arts, commerce, law, and science texts?
3. What percentage of the words in the Academic Corpus does the AWL cover?
4. Do the lexical items identified occur frequently in an independent collection of academic texts?
5. How frequently do the words in the AWL occur in nonacademic texts?
6. How does the AWL compare with the UWL (Xue & Nation, 1984)? (Coxhead, 2000, p. 218)

In order to answer these research questions, Coxhead based the word list on sound principles of corpus linguistics. First, Coxhead realized the importance of representing a wide variety of academic disciplines, noting that the linguistic features may differ considerably across these disciplines. Second, the larger corpus was divided into approximately equal sections to effectively calculate the range (dispersion) of the academic vocabulary across the entire corpus. Finally, Coxhead considered the size of the corpus. Earlier researchers in the 1970s who designed smaller-scale corpora (Campion & Elley, 1971; Ghadessy, 1979; Praninskas, 1972) only included between 300,000 and 500,000 words due to manual word counts. Coxhead determined that the size of the corpus used to design an academic word list should be large enough to allow the most salient and important words to emerge. Therefore, she gathered a much larger corpus of academic text, including articles from journals, textbooks, and texts from the

Wellington Corpus of Written English (Bauer, 1993), consisting of 28 subject areas in the domains of arts, commerce, law, and science. In total, over 3.5 million words of primarily New Zealand English from these domains were collected, with approximately 875,000 words in the disciplines of arts, commerce, law, and science.

When selecting academic core words, Coxhead decided to use the principle of word families for counting purposes. Word families, as defined by Coxhead, were the base of a word and all related forms containing affixes, including inflected forms and transparent derivational forms. For example, the AWL includes the headword *survive* and the inflectional forms of *survives, survived,* and *surviving,* as well as derivational forms of *survival, survivor,* and *survivors.* Coxhead argued that some morphological differences (especially simple inflectional endings) suggest a strong relationship between words that allows them to be be grouped into these word families.

With these family groupings in mind, Coxhead established three measurement criteria for inclusion in the AWL: (1) The word families had to be beyond the first 2,000 most frequent words in English found in West's (1953) GSL. Coxhead asserted that there is a distinction between general high frequency vocabulary and more specialized academic vocabulary; (2) word families had to be found more than 10 times in each of the four sections of the corpus and in at least 15 of 28 the subject areas represented; and (3) the members of a word family had to be present 100 times or more in the entire corpus, with an average of 25 times in the four outlined sections of the corpus (arts, commerce, law, and science). Using these measurement criteria, Coxhead produced the AWL. It has been the most commonly used word list for teaching ELLs in the last 15 years (Coxhead, 2011).

Issues with the AWL

Although the AWL had a stronger research design and a basis in contemporary advances in corpus linguistics, it still contained key limitations that needed to be addressed by future research:

(1) When Coxhead created the AWL, she did not address the concern that academic word families may be far too broad and inclusive to provide an accurate representation of a core academic vocabulary. For example, the noun *use* and the verb *use* are not only pronounced differently but the noun form is definitely more academic than the verb form. Likewise, *proceeds* (the noun) and *proceeds* (present tense verb) are also pronounced differently and have potentially different impacts in academic writing. In both of these instances, word families would not account for these distinctions.

In addition, word families make no distinction for word forms with multiple meanings (Hatch & Brown, 1995). Some word forms have vastly different meanings (homonymy) in contexts. For example, the word form *mean* (verb), as in *What does she mean?*, the adjective *mean*, as in *She's a mean person*, and the noun *mean,* as in *Find the mean of these numbers*, would have no distinction in a word family.

(2) The AWL does not consider the high frequency academic words in the first 2,000 most frequent words of the English language found in the General Service List (GSL). Several researchers have questioned this decision on the basis that some salient academic vocabulary might be in the highest frequency lists of English (Nation and Townsend, 2012; Neufeld, Hancioğlu., & Eldridge, 2012). In fact, using the much more modern *Corpus of Contemporary American English* (COCA), Gardner and Davies (2014) discovered that 451 of the 570 AWL

word families fall within the first 4,000 words of that corpus, suggesting that many AWL word families are actually high frequency words of English.

Some researchers disagree on whether general high frequency vocabulary and academic vocabulary should be treated as mutually exclusive. While many researchers distinctly separate general high frequency vocabulary and academic vocabulary (Coxhead, 2000; Praninska, 1972), others favor making no distinction between general vocabulary and high frequency academic vocabulary (Hancioğlu, Neufeld, & Eldridge, 2008; Neufeld & Billuroğlu, 2005). Finally, others have decided to keep the separation between high frequency words and academic vocabulary but allow for academic words to also be in general high frequency lists (Gardner & Davies, 2014). These paradigms greatly influence the salient academic words that may or may not appear in pedagogical word lists.

(3) the AWL draws its conclusions from a relatively small corpus of 3.5 million words, primarily of New Zealand English. In more recent years, researchers have gained access to much larger corpora for vocabulary research. Advances in designing megacorpora have led to the creation of corpora such as *The Corpus of Contemporary American English (COCA)* (520 million tokens), the *Wikipedia Corpus* (1.9 billion tokens), and many others. All of these corpora are substantially larger than the 3.5-million-word corpus used by Coxhead (2000) to create the AWL, although not all of them are academic. It is clear that larger corpora allow researchers to create word lists that more accurately reflect the language being targeted.

The Development of the AVL

In response to the serious criticisms of the AWL, Gardner and Davies (2014) created a new list of high frequency academic vocabulary, known as the Academic Vocabulary List (AVL). Their goals were as follows:

(1) The AVL used lemmas rather than word families. Lemmas are defined as "a set of lexical forms having the same stem and belonging to the same major word class differing only in inflection and/or spelling." (Francis & Kučera, 1982, p.1). For example, the verb START includes *start, starts, started,* and *starting,* but not *starter, restart*, and so forth. All of these members of the lemma are verbs related through inflectional morphology. STARTER would constitute a separate lemma because it belongs to a separate word class (noun rather than verb). By accounting for part of speech, lemmas provide some granularity in terms of potential meaning differences in words with similar forms (*use,* the noun vs. *use,* the verb). In addition, the use of lemmas more accurately represents the developing morphological abilities of ELLs for whom a particular list is intended (Gardner, 2007). In short, producing pedagogical word lists that include derivational relationships rather than inflectional relationships is problematic because metacognitive awareness of derivational morphology develops after awareness of inflectional morphology (Nation & Waring, 1997). Even after years of studying English, ELLs struggle to correctly identify and produce derived members of a word family (Schmitt & Zimmerman, 2002). While counting lemmas is not a perfect solution to issues of homonymy, because lemmas do not account for for multiple meanings of a single word form from the same part of speech (e.g., a *run* in baseball vs. a *run* in the nylons), it is still more robust than word families.

While lemmas do provide for this granularity, Gardner and Davies (2014) also converted their lemma-derived list into word families to allow for direct comparison against the AWL, which was already organized by the principle of word families (Coxhead, 2000). It would be impossible to convert the word families of the AWL to lemmas for comparison purposes in the present study.

(2) The AVL was created using a large, modern corpus of 120 million words of academic American English. These words were grammatically tagged for parts of speech (nouns, verbs, adjectives, and adverbs) by the CLAWS 7 tagger (Rayson, 1996) from Lancaster University, using the larger 450-million-word (now 520-million-word) *Corpus of Contemporary American English* (COCA). Gardner and Davies used a corpus that contained a wider variety of academic disciplines than Coxhead (2000) did in order to obtain the core academic words that are common to all these disciplines. Each of the academic disciplines in the 120-million-word corpus (e.g., education, humanities, history, social science, philosophy/religion/psychology, law/political science, science/technology, medicine/health, and business/finance) had between 8 and 22 million words. Therefore, this corpus was almost 35 times larger than the 3.5-million-word corpus designed by Coxhead (2000) for the AWL.

(3) The AVL was statistically derived without consideration of predetermined lists, such as the GSL, or any other pedagogically established lists. Using the large corpus and powerful statistics (e.g., ratio, range, dispersion, and discipline measures), Gardner and Davies separated academic words from general words of English and also technical academic words from core academic words, resulting in the AVL.

The Current Study

Coxhead (2000) created another academic corpus and a fiction corpus to examine the coverage of the AWL. Likewise, Gardner and Davies (2014) compared the AVL to COCA and the BNC to determine its coverage. In both instances, the researchers used general megacorpora to examine how well their word lists cover text. In the Gardner and Davies study, the researchers also directly compared the AWL and AVL using the megacorpora.

To date, however, there have been no studies that have compared the coverage of these two lists using a practical academic corpus that better represents materials from authentic learning and teaching contexts. The purpose of this study is to provide an analysis of the AWL and the AVL in the Academic Textbook Corpus (ATC)—a collection of academic materials at different grade levels and from different disciplines that more closely represents what ELLs might encounter in an actual classroom curriculum. With this in mind, the following questions will guide the remainder of this study.

1. What word families from the AWL and AVL are found in the Academic Textbook Corpus (ATC)?
2. What AWL word families are unique to the ATC?
3. What AVL word families are unique to the ATC?
4. Are there characteristic differences between the AWL and AVL unique word families?

CHAPTER THREE

METHODS

The Academic Textbook Corpus

Biber (1993) noted that corpus-based analyses are only reliable as long as the corpus is representative of the targeted language as a whole. Therefore, in order to create a representative corpus to answer the research question of this study, the Academic Textbook Corpus (ATC) was designed to be representative of authentic academic materials, particularly textbooks encountered in educational contexts. Three disciplines that both native English speakers and ELLs study throughout their academic careers were selected for the ATC: American history, science, and mathematics. Three texts from each of the grade levels were selected: junior high, (or approximately 8th grade); high school, (or approximately 11th of 12th grade); and college, (or general education courses) (Danzer, 2007; Dearden & Lawler, 2012; Fish, Latimer, & Souza, 2013; Garcia, 2007; Jordan & Dirga, 2013; Utah State Office of Education, 2014a; Utah State Office of Education, 2014b; Utah State Office of Education, 2014c; Stewart, 2012, see Appendix A). The complete corpus contained 1,911,307 words.

Procedures

These textbooks were first scanned (if not already in PDF format) and then converted into .txt documents per the requirements of the software program (described below). In order to ensure reliable counts, any errors (especially spelling errors) that occurred due to the scanning process were removed once these documents were converted into .txt documents.

Word Lists

The 570 word families of AWL and the top 570 word families of the AVL were the word lists compared in this study. It is crucial to remember that the AVL was created using the principles of lemmas (base form plus inflectional affixes), whereas the AWL was created using the principle of word families (base form plus transparent derivational and inflectional affixes). Due to these differences in methodology, it is impossible to reconstruct the AWL in terms of lemmas. It is possible, however, to convert the lemma-based AVL into word families. To allow for direct comparison, the AVL was organized into 1,991 word families. While counting word families is problematic in many regards (c.f. Gardner & Davies, 2014), it is an unavoidable decision in order to provide the most equitable comparison possible. Therefore, the first 570 word families of the AVL were considered when conducting the initial quantitative and qualitative analyses.

Another key issue for this study is the tendency of word families to exaggerate frequency counts. Some seemingly general words (types) often become included in academic word families, even though they themselves may not necessarily be academic. This was true in Coxhead's (2000) methodology, and it is true when AVL lemmas are converted to word families for comparison purposes in this study. For example, word families such as *find, study,* and *use* may contain some members that are more general high frequency in their coverage (e.g., *find* (verb), *study* (verb), *use* (verb)), and some that are more often found in academic materials (e.g., *finding* and *study* (nouns), and *use* (noun)). The non-academic words in these liberal word families will skew frequency counts if the focus is on academic English. This is a problematic

but unavoidable limitation of the current study because of the initial decision to use word families in creating the AWL.

Computer Programs and Procedures for Collecting Data

The Range program (Heatley, Nation, & Coxhead, 2002) was used for the quantitative analysis of the word families. Range allows users to upload lists of words to calculate frequency of a given set of words. By default, Range has three base lists: (1) the first 1,000 words of the GSL; (2) the second 1,000 words of the GSL; and (3) the 570 word families of the AWL. To analyze the AWL, these base lists were maintained. When analyzing the AVL, the base lists were modified to include two word lists: the top 570 word families of the AVL and the complete 1,991 word families of the AVL.

The Range program produced the percentage of the text covered by these words, the number of types (distinct words), tokens (the number of times types appear), word families (base form of a word plus derivational and inflectional affixes), and the range of each word (the number of texts the word appears in). After analyzing the texts individually and the ATC as a whole using the standard base lists and the modified base lists, the results were sorted in Excel worksheets. This process also separated the word families into three distinct groups that relate to the previously mentioned research questions.

1. *Word families common to both the AWL and the AVL*: These are words that are found in both the 570 word families of the AWL and the top 570 word families of the AVL.
2. *Word families found only in the AWL*: These are words that are exclusive to the AWL and are not found in the AVL.

3. *Word families found only in the AVL*: Initially, these were words found only in the top 570 word families of the AVL. After this analysis, the remainder of the word families of the AVL were also considered to see if any AWL word families in the ATC were in the expanded AVL—i.e., beyond the top 570 word families.

The words in these categories, along with data about the frequency and range of these words, were collected for each textbook and for the corpus as a whole. These data were then organized in Excel documents for analysis. It was decided to define a word in this study as a word family. While there are many issues with the organizing principle of word families (as mentioned in the literature review), the AWL was created using the principle of word families. As a result, it is best to provide a point of direct comparison by evaluating the AWL word families and the AVL word families.

CHAPTER FOUR

RESULTS AND DISCUSSION

Coverage, Frequency, Range in the ATC

The research questions for this study relate to both the unique and shared AWL and AVL word families that appear in the ATC, as well as any characteristic differences between these word families. Before discussing these issues, however, it is valuable to describe the coverage, frequency, and range of the AWL and the AVL to inform the answers to the previously discussed research questions.

The coverage in the ATC provided by these lists differs in specific ways (see Table 1). First, the percentage of types (distinct words) found in the ATC varies between the AVL and the AWL. Approximately one to three percent more AVL types appear in the 8th grade, high school, and college sub-corpora of the ATC. For sub-technical lists of words, such as the AWL and the AVL, this difference is noteworthy. The AWL, for example, covered approximately 10% of types in academic materials (Coxhead, 2000). Therefore, one to three percent greater coverage by the AVL, as observed in the ATC, is a substantial difference.

Additionally, the number of tokens (number of times a type occurs) varies substantially between the word lists. This finding suggests that the AVL types appear more frequently and consistently in the ATC across grade levels. Finally, the number of word families covered by the AWL and the AVL differs in the ATC, with the ATC containing 567 of the AWL families and all 570 of the top AVL families. Looking more closely at the individual sub-corpora of the ATC, the AVL covers more word families consistently across grade levels as well.

Table 1
AWL and AVL Coverage Across the ATC

	Academic Word List			Academic Vocabulary List		
Context	Token	Type	Families	Token	Type	Families
8th Grade	5.78%	10.11%	525	12.14%	13.14%	546
High School	6.29%	7.77%	555	13.14%	9.44%	568
College	7.17%	8.35%	562	13.93%	9.81%	567
Corpus	6.67%	5.99%	567	13.40%	7.08%	570

In addition to the overall coverage, there is additional evidence that the AVL word families in the ATC are more frequent (see Table 2). In order to calculate the average frequency of a word family, the frequency of the family is divided by the total number of word families that appear in each sub-corpus or the ATC as a whole from either the AWL or the AVL. The average frequency of an AVL word family is 449 compared to the AWL's 225. In fact, in each sub-corpus of the ATC the members of an AVL word family consistently appear almost twice as many times as the members of an AWL word family on average.

Table 2
AWL and AVL Average Family Frequency in the ATC

	Academic Word List	Academic Vocabulary List
Context		
8th Grade	30	60
High School	75	152
College	125	241
Corpus	225	449

The range (the number of texts in which a word family is present) is also substantially different for the AWL and AVL word families appearing in the ATC (see Table 3). As previously mentioned, the ATC sub-corpora (8th grade, high school, and college curricula) contain three texts each, with a total of nine texts in the complete ATC. To calculate the average range of the word families found in the ATC, the range values of each word family (from 1 to 3 for each sub-corpus and from 1 to 9 for the entire ATC) were averaged. In each sub-corpus, and the complete ATC, the differences in range favor the AVL consistently across all three grade levels. For the ATC as a whole, the average range of the AVL word families was 7.3 compared to the AWL's 6.3. Therefore, the AVL word families that appear in the ATC appear more consistently across disciplines and grade levels than the AWL word families. These figures must be qualified, however. The measurement of Range does not consider word families that might have high frequency in one of the nine texts of the ATC and very low frequency in other texts. Dispersion, another measure frequently used in corpus linguistics to measure how well words are spread throughout a corpus, would ultimately be a better measure but using dispersion fell outside the research questions of this particular study.

Table 3
AWL and AVL Average Family Range in the ATC

Context	Academic Word List	Academic Vocabulary List
8th Grade	1.9	2.3
High School	2.0	2.3
College	2.6	2.8
Corpus	6.3	7.3

These data are useful to understand the relative impact of the AWL and the AVL in the ATC as a whole. It is clear that the AVL has advantages over the AWL quantitatively, but a narrower qualitative comparison of the most salient words in the ATC is also warranted and will provide additional insight.

Words Shared in the AWL and the AVL

First, a comparison of the shared AWL and AVL word families that appear in the ATC is necessary. When considering these shared words, it is important to discuss the most salient ones, or those that have the highest possible range of nine (see Table 4). Those words with a range of nine have the greatest utility because they are found in all of the texts of the ATC and are also high frequency. For example, the word family *function* appears 4,441 times in the ATC.

Table 4:
Families Shared in the AVL and AWL with a Range of 9 in the ATC

Function	4441	Individual	460	Accurate	259
Require	2119	Source	450	Consist	259
Section	1804	Principle	447	Error	250
Affect	1723	Involve	440	Design	248
Define	1335	Construct	437	Specific	244
Region	1234	Interpret	433	Select	228
Vary	1087	Summary	423	Convert	216
Create	1057	Concept	411	Primary	204
Positive	979	Focus	411	Conclude	191
Factor	925	Image	402	Core	187
Evaluate	832	Conduct	386	Feature	181
Volume	812	Correspond	386	Technique	174
Estimate	791	Interact	382	Objective	160
Method	774	Indicate	376	Significant	158
Element	736	Assume	353	Visible	148
Occur	679	Research	299	Appropriate	145
Similar	672	Shift	297	Device	134
Process	666	Available	291	Cycle	121
Identify	654	Previous	290	Unique	118
Period	641	Instance	286	Outcome	113
Obtain	538	Proportion	285	Aspect	104
Sequence	535	Complex	271	Distinct	104
Range	518	Distribute	271	Sufficient	93
Illustrate	489	Symbol	264	Input	60
Locate	472	Technology	260		

Additionally, many of these words are generally found across a variety of academic disciplines (*require, section, affect, factor, elements,* and *create*). They are not technical words, or words restricted to certain disciplines. In fact, their utility extends to a variety of disciplines. Gardner and Davies (2016), in reference to this usefulness across disciplines, refer to core academic words as being "saturated with academic sense" (p. 66). One important limitation of this core academic vocabulary is that some of these words may change meaning across different disciplines (e.g., *factor of a given number, factor influencing a political revolution,* and *factor causing a physical change*). These cases, however, appear to be exceptions, rather than the rule.

The presence of these shared word families in the ATC provides information about the AWL and the AVL. Since the ATC was created using authentic academic materials, a subset of high frequency words "saturated with academic sense" should appear across the grade levels and disciplines. These shared AWL and AVL word families, which appear across a variety of disciplines with high frequency in this authentic context, are core academic vocabulary—words that are foundational to academic discourse. The high frequency and range of these word families suggest that both pedagogical word lists capture some of the crucial word families for ELLs to know in order to understand authentic academic texts.

Words Unique to the AWL

Beyond the shared AWL and AVL word families, there are few words that are unique to the AWL with a range of nine (see Table 5). One crucial difference between the shared AWL and AVL word families found in the ATC and these unique AWL word families found in the ATC is the frequency. The frequency of the words unique to the AWL that appear in the ATC is

substantially lower. For example, the top unique AWL word family, *chapter,* appears in the ATC 2,396 times, whereas the top shared AWL and AVL word family, *function*, appears 4,441 times in the ATC. Similar to the shared word families, however, these words also tend to be core academic words (*area, major, expand, sphere, final,* etc.). These words cannot be confined to one specific academic discipline; their utility extends to a variety of disciplines.

Table 5
Exclusive AWL Word Families with a Range of 9

Chapter	2396
Energy	2306
Area	1803
Remove	1583
Constant	958
Compute	834
Analyse	757
Negate	658
Major	605
Expand	599
Percent	463
Final	430
Sphere	383
Respond	350
Consume	281
Eventual	253
Investigate	247
Reverse	212
Label	140
Text	139
Infer	136
Expose	93
Credit	80
Foundation	65

Words Unique to the AVL

In contrast to the relatively low number of unique AWL word families that appear in the ATC, the AVL has many unique word families with a range of nine in the ATC (see Table 6). These word families also have higher frequencies than the words unique to the AWL. For example, the top unique AVL word family *use* appears 6,858 times in the ATC, compared to the top unique AWL word family *chapter*, which only appeared 2,396 times in the ATC. Many of the words unique to the AVL in the ATC appear to be general academic words (*system, form, state, means, equal, examples,* etc.), like the words unique to the AWL.

Unlike words unique to the AWL, however, the number of unique AVL word families is fairly high, with 177 unique AVL word families compared to 24 unique AWL word families. Some unique AVL word families not found in the AWL seem to be general high frequency vocabulary (e.g., *use*, *add,* etc.), or vocabulary found in all contexts, rather than specifically academic contexts. Despite their seemingly general nature, they also have high frequency and high range in the ATC indicating their importance in academic contexts.

These core academic words should not be omitted from a pedagogical word list because ELLs in authentic academic contexts would be likely to encounter these word families early in their academic career. If the target of these ELLs is academic English rather than general English, they need to understand these words regardless of their seemingly general nature.

Table 6
Exclusive AVL Word Families in the ATC with a Range of 9

Use	6858	Measure	1173	Develop	698	Difference	560	Exchange	314	Essential	108
Find	5612	Mean	1171	Low	695	Practice	554	Success	309	Distinguish	64
Figure	4637	Need	1152	Test	693	Level	542	Current	295		
State	4433	Grow	1144	Common	681	Total	536	Extend	269		
I.e.	4227	Center	1117	Contain	661	Inform	517	Improve	266		
Content	4141	Model	1101	Negative	657	Direct	501	Associate	259		
Example	3901	Express	1096	Provide	657	Plan	498	Accept	254		
Part	3440	Result	1087	Exist	647	Organize	496	Basic	246		
Give	3399	Calculate	1080	Probably	603	Member	489	Reduce	246		
Solution	3073	High	1067	Shape	562	Discuss	484	Recognize	245		
Change	2680	Product	1061	Difference	560	Condition	479	Various	238		
Add	2543	Therefore	1047	Practice	554	Introduce	476	Necessary	236		
Term	2178	Help	1014	Level	542	Combine	470	Scale	229		
Form	2096	Group	1010	Total	536	Particular	467	Report	222		
Follow	2084	Table	971	Inform	517	Understand	464	Attempt	215		
Move	2030	Science	965	Direct	501	Experiment	460	Knowledge	214		
Material	1981	Relate	928	Plan	498	Reflect	459	Collect	205		
Experience	1728	Describe	909	Organize	496	Standard	457	Likely	188		
Both	1694	Support	876	Member	489	Century	456	Tend	188		
Review	1632	Note	870	Discuss	484	Assumption	442	Future	185		
Increase	1618	Apply	844	Condition	479	Study	441	Subject	185		
Act	1612	Include	838	Introduce	476	Refer	436	Account	183		
System	1597	Unit	835	Combine	470	Discover	427	Typical	181		
Whole	1559	Nature	804	Particular	467	Connect	412	Advance	179		
Equal	1519	Thus	781	Understand	464	Protect	407	Rapid	175		
Large	1460	Type	775	Experiment	460	Gain	389	Differ	173		
Explain	1401	Above	765	Reflect	459	Depend	372	Purpose	168		
Represent	1379	Effect	765	Test	693	Human	365	Enjoy	165		
Continue	1274	Observe	755	Common	681	Separate	365	Perform	165		
Main	1225	Base	744	Contain	661	Present	364	Manage	159		
Determine	1202	Important	740	Negative	657	Difficult	359	Examine	148		
General	1191	Analyze	738	Provide	657	Wide	354	Desire	146		
Limit	1190	Produce	738	Exist	647	Actual	349	Characteristic	134		
However	1181	Consider	724	Probably	603	Pattern	344	Description	133		
Rate	1180	Compare	701	Shape	562	Relative	339	Variety	123		

Top AWL Word Families in the ATC

It is also valuable to examine the AWL words that appear in the ATC most frequently, regardless of their range (see Table 7), to determine characteristic differences between the two pedagogical word lists. First, it is worth noting that the top AWL word families have a high range overall—*function, require, area, section,* and *region* appear in all nine texts of the ATC. Other words also appear in many texts. Like many of the word families previously mentioned, all of these top AWL word families appear in a variety of disciplines and have a great deal of academic saliency (*equate, function, require, section, affect, restrict,* etc.). In addition to the high ranges, their frequency is also high, indicating that they are valuable words to be taught and considered in terms of English for academic purposes. The average frequency is 1,876. Furthermore, it is crucial to note that the number of shared top words between the AWL and the AVL is high, with 13 of them shared with the AVL.

There are also several words included in the top 20 AWL words that do not appear in the AVL (*chapter, energy, area, remove, edit, series, react*). These word families would not be confined to one or two disciplines. In fact, they are like other word families previously mentioned— words that are important to understand in academic contexts. Therefore, these core academic words likewise deserve the attention of students, teachers, and curriculum developers involved with English for academic purposes.

Table 7
Top 20 AWL Word Families in the ATC

Word Family	Range	Frequency
Equate*	8	4791
Function*	9	4441
Chapter	9	2396
Energy	9	2306
Require*	9	2119
Area	9	1803
Section*	9	1770
Affect*	9	1723
Remove	9	1583
Restrict*	7	1536
Edit	6	1474
Overall*	8	1440
Economy*	6	1437
Subsequent*	8	1392
Define*	9	1335
Series	8	1304
Formula*	8	1247
Region*	9	1232
React	8	1107
Vary*	9	1087

Note: Words with an asterisk are also found in the AVL.

Top AVL Word Families

The top 20 AVL word families in the ATC reveal a great deal about the AVL (see Table 8). These words have a consistently high range, and they appear frequently throughout the ATC. In fact, their ranges and frequencies are substantially higher than the top AWL word families. The average frequency of these top 20 AVL word families is 3,601, almost twice as high as the average frequency of the top AWL word families. In terms of their characteristics, many of these words appear to be core academic words (*use, find, figure, function, state,* and *content*), like the top AWL word families. Of the top twenty words, however, only two of them appear in the AWL. Some of these words might not appear in the AWL because they are also general high

frequency vocabulary, and would therefore not have been considered in the Coxhead (2000) methodology. If the aim of an academic word list is to create a core set of academic vocabulary, these words that appear with high frequency and across the range of academic disciplines and grade levels cannot be excluded (e.g., *use, find, example,* and *add).* Other words also carry a great deal of academic weight and should likewise not be excluded from core academic vocabulary (*IE, example, part,* etc.).

Table 8
Top 20 AVL Word Families in the ATC

Word	Range	Frequency
Use	9	6858
Find	9	5612
Equation	6	4772
Figure	9	4637
Function*	9	4441
State	9	4433
IE	9	4227
Content	9	4141
Example	9	3901
Part	9	3440
Give	9	3399
Solution	9	3073
Value	8	2844
Change	9	2680
Govern	5	2551
Add	9	2543
Term	9	2178
Require*	9	2119
Form	9	2096
Follow	9	2084

Note: Words with an asterisk are also found in the AWL.

Word Families beyond the Top 570 AVL Word Families

After examining the AWL and the initial 570 word families of the AVL, it is important to determine if any AWL word families in the ATC also appear in the expanded AVL—i.e., beyond the top 570 AVL families already examined (see Table 9). First, the words with the greatest range need to be examined. There are a handful of word families in the ATC that also appear in the expanded AVL, and they seem to have general academic senses (*energy, sphere, expand*, and *infer*).

Additionally, there are many high frequency core academic words that are included in the expanded AVL word families that are not found in the AWL (*point, number, problem,* etc.). Some of these unique AVL word families also seem to be general high frequency words (*learn, more, work,* and *know*). While these words seem more general in nature, their absence in the AWL might bolster the claim that the AWL is less representative of a core academic vocabulary because it omits academic words in higher frequency tiers of the language.

These findings, along with the other observations about the AVL word families, indicate that academic vocabulary and general high frequency vocabulary might not be mutually exclusive, as is presupposed by the methodology employed by Coxhead when she did not consider words from the General Service List (GSL). This is additional evidence that the line between general high frequency vocabulary and core academic vocabulary is blurred. In fact, several researchers in recent years have emphasized this same limitation in the approach that generated the AWL (c.f. Gardner & Davies, 2014; Hancioğlu, Neufeld., & Eldridge 2008; Schmitt & Schmitt, 2012). /

Table 9
AVL Word Families with a Range of 9 Beyond the Top 570 Word Families

Word Family	Frequency	Word Family	Frequency	Word Family	Frequency
Learn	4794	Word	478	Infer*	136
Point	3801	Appear	473	Inward	134
More	3307	Last	450	Perfect	121
Number	3011	Complete	445	Regular	109
First	2853	Decrease	436	Familiar	107
Work	2417	Stand	421	Gradual	101
Energy*	2311	Bind	417	Ready	94
Know	1851	Double	404	Remainder	82
Power	1840	Certain	403	Weigh	72
Problem	1704	Strike	388	Mention	64
Think	1670	Sphere*	383	Fold	59
Great	1580	Correct	382		
Due	1543	Return	376		
Lead	1518	Expand*	362		
Copy	1449	Send	360		
Second	1347	Short	346		
Set	1213	Single	338		
Land	1141	Better	337		
Name	1078	Educate	332		
Simple	1065	Special	319		
Cause	979	Ground	316		
Order	937	Join	315		
Direction	820	Accord	303		
Circle	810	Expect	296		
Operate	789	Oppose	296		
Question	753	Weight	282		
True	711	Arrange	278		
Position	667	Further	268		
Notice	661	Compose	262		
Reason	607	Quantity	257		
Less	606	Eventual*	253		
Open	559	Repeat	253		
Read	553	Consume*	239		
Age	536	Class	231		
Able	529	Imagine	230		
Within	518	Fix	216		
Per	515	Reverse*	213		
Allow	499	Label*	141		

Note: Words with an asterisk are also found in the AWL.

Highest Frequency AVL Families beyond Top 570

Finally, it is important to consider the highest frequency word families beyond the Top 570 AVL families (see Table 10). Many of these word families have high ranges, including many words with a range of nine (*learn, point, work,* etc.). Those that do not have a range of nine also tend to be strong academic words and carry a great deal of academic saliency (e.g., *graph, reserve,* etc.). Some of those beyond the top 570 word families are also absent from the AWL, possibly due to their "general" nature *(*e.g., *more, first, power,* and *rule)*. The absence of these words, however, again demonstrates how the AWL does not contain some core academic words that are frequent in a variety of disciplines and grade levels.

Table 10
Top 20 AVL Word Families beyond the Top 570 Word Families in the ATC

Word Family	Range	Frequency
Learn	9	4794
Point	9	3801
Graph	8	3607
More	9	3307
Number	9	3011
First	9	2853
Reserve	6	2820
Suppress	4	2729
Work	9	2417
Energy*	9	2311
Atom	7	2014
Nation	7	1943
Know	9	1851
Power	9	1840
Colony	5	1754
Problem	9	1704
Think	9	1670
Great	9	1580
Rule	8	1571
Due	9	1543

Note: Words with an asterisk are also found in the AWL.

CHAPTER FIVE

CONCLUSIONS

Summary of Findings

Both the AWL and the AVL purport to present a list of vocabulary items that ELLs are likely to encounter in academic settings, particularly written academic texts. The ATC, therefore, was designed to be representative of texts that ELLs would encounter in real-life situations. Analyzing this corpus for AWL and AVL word families allowed these claims to be more carefully examined.

First, the general descriptive statistics from this study determined the overall differences in coverage of the AWL and the AVL in an authentic context. Both lists provide significant coverage of the ATC. The AVL, however, provides greater coverage on the whole. The AVL as a list appeared more frequently in terms of types, tokens, and word families in the ATC. In addition to greater frequency, the range of AVL words appearing in the ATC was substantially higher than in the AWL.

The first goal of this study was to determine what words from both the AWL and the AVL are found in the ATC. In many regards, both lists provide a good estimation of the core set of academic vocabulary since many words that are "saturated with academic sense" (Gardner & Davies, 2016, p. 66) are shared between the lists. In other words, the shared AWL and the AVL words cannot be restricted to one academic discipline and their utility extends to a variety of contexts. These high frequency words also appeared consistently throughout the grade levels and disciplines of the ATC. Therefore, those involved with English for academic purposes, such as administrators, teachers, curriculum designers, and ELLs should consider these words.

A second purpose of this study was to determine what AWL words were unique to the ATC—in other words, not found in the AVL. The AWL had very few unique word families (24) but many of these words also appear to be core academic vocabulary. Like the shared AWL and AVL words in the ATC, the unique AWL word families appeared consistently throughout disciplines and grade levels in the ATC. Therefore, these word families should likewise be considered in terms of English for academic purposes.

Finally, this study also aimed to determine the unique words from the AVL in the ATC. The AVL had substantially more unique words appearing in the ATC (177) than the AWL. These words also appeared consistently in the grade levels and the various disciplines across the ATC. Characteristically, some words unique to the AVL appearing in the ATC seem to be more general in nature but are nonetheless high frequency in academic contexts. This finding suggest that, unlike the AWL, the AVL captures core academic vocabulary from the higher frequency tiers of English, such as the word families of *know* and *study.*

One primary finding from this study is that the AVL, while not necessarily a perfect list of core academic vocabulary, captures core academic vocabulary in substantially better ways than the AWL. The methodology employed by Coxhead (2000) asserts that the words in higher general frequency tiers of the language should not be considered in a list of academic words, although recent studies have shown that many AWL word families are indeed found in the highest tiers in modern corpora of English (e.g., Gardner & Davies, 2014; Schmitt & Schmitt, 2012). This presumption ignores the high frequency, high range, and saliency of these seemingly general high frequency words in academic contexts. The presence of a high number of unique

AVL word families in high frequency tiers of general English suggests that these word must be considered.

Limitations

In this study, there are several key limitations to consider. First, this study compared word families rather than any other definition of a word. This decision limited the analysis of the actual types (distinct words) that appear in the ATC, but it is necessary to provide a reasonably fair comparison of the two lists. The AWL was originally based on the principle of word families. Therefore, in order to allow for comparisons such as the one done in this study, the lemma-based AVL was converted into families. As a result, the high frequency members of some AVL word families (e.g., *find, study, use)* skewed some of the frequency counts. This same is also true of AWL families, where a certain word or words were frequent in academic texts, yet all members of the family were included, regardless of their academic saliency. However, this unavoidable limitation of the study is additional evidence that the field needs to move away from the word-family paradigm and move towards the more valid constructs of lemma, or even lexeme (individual meanings), when analyzing frequency of written and spoken texts. Word families simply do not provide the necessary granularity needed for valid and reliable measures of vocabulary.

In addition to the aforementioned limitations, the conclusions from this study should not be considered definitive. It is evident that each word list has some advantages and disadvantages. This study compares the similarities and differences of the AWL and the AVL in an ecologically valid corpus. Researchers should be extremely careful to draw absolute conclusions about the validity and reliability of word lists based on a single evaluation

Implications

This study sheds light on academic core vocabulary in authentic contexts. Teachers, curriculum developers, and ELLs need to consider which list of academic vocabulary best represents the vocabulary in academic materials. From this study, it appears that the AVL has better coverage across grade levels and disciplines. Basing curricula, instruction, and individual study on the AVL rather than the AWL allows these stakeholders to better deal with the challenges of academic vocabulary. The AWL has provided valuable insights into academic vocabulary for many years, but the differences in coverage as well as the high frequency core academic words found only in the AVL need to be considered by those involved with English for academic purposes.

Suggestions for Future Research

This research has demonstrated key similarities and differences between the AWL and the AVL in an authentic academic corpus, with the AVL appearing to have some important advantages over the AWL. Additional validation studies are needed, and could include larger corpora, more grade levels, and a broader range of academic disciplines for making comparison between the AWL and the AVL. Additionally, the AWL and the AVL have been limited to single word units rather than multi-word vocabulary items. Future research should also examine multi-word vocabulary items that are prevalent in authentic academic corpora.

Summary Statement

This study has demonstrated some important advantages of the AVL over the AWL. Therefore, individuals with a vested interest in English for academic purposes should consider how the AVL can help others to achieve in academic situations. For example, teachers of both

native English speakers and ELLs who are learning English for academic purposes should emphasize the importance of these words. Material developers, likewise, should consider how well the AVL represents authentic academic materials and consider altering their materials as necessary. Finally, researchers in applied linguistics should continue to examine the composition of academic texts using the AWL and the AVL to determine the advantages and disadvantages of both lists, which both seem to capture a set of core academic vocabulary to some extent. This core academic vocabulary appears to be an essential component of academic language. Command of such words has important ramifications for ELLs and native speakers in high-stakes academic settings.

APPENDIX A

Books included in the Academic Text Corpus

8th grade texts

Garcia, J. (2007). *Creating America :A history of the United States*. Evanston, IL: McDougal Littell. (178,538 words)

Utah State Office of Education. (2014). *8th grade mathematical foundations textbook.* Salt Lake City, UT: Utah State Office of Education. (64,945 words)

Utah State Office of Education. (2014). *8th grade integrated science.* Salt Lake City, UT: Utah State Office of Education. (29,887 words)

High school texts

Danzer, G. A. (2007). *The Americans*. Evanston, IL: McDougal Littell. (450,207 words)

Jordan, L., & Dirga, K. (2013). *CK-12 Algebra II with trigonometry concepts.* CK-12 Foundation. (156,087 words)

Utah State Office of Education. (2014). *Chemistry*. Salt Lake City, UT: Utah State of of Education. (64,945 words)

College texts

Dearden, D., & Lawler, M.J. (2012). *Physical science foundations*. Provo, UT: Brigham Young University. (238,269 words)

Fish, R., Latimer, B., & Souza, E. (2013). *City upon a hill.* Provo, UT: Brigham Young University. (110,598 words)

Stewart, J. (2012). *Calculus: Early transcendentals*. Belmont, Cal. Brooks/Cole, Cengage Learning.(630,971 words)

REFERENCES

Bauer, L. (1993). *Manual of information to accompany the Wellington corpus of written New Zealand English*. Department of Linguistics, Victoria University of Wellington.

Beck, I. L., Perfetti, C., and McKeown, M. G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology, 74,* 506-521.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing, 8*(4), 243-257.

Biemiller, A. (1999). *Language and reading success* (Vol. 5). Cambridge, MA: Brookline Books.

Biemiller, A. **(**2003). Vocabulary: Needed if more children are to read well. *Reading Psychology 24,* 323-35.

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology, 93,* 498–520.

Campion, M., & Elley, W. (1971). An academic vocabulary list. Wellington: New Zealand Council for Educational Research.

Carter, R. (2012). *Vocabulary: Applied linguistic perspectives*. Routledge.

Chall, J.S., Jacobs, V.A., & Baldwin, L.E. (1990). The Reading Crisis: Why Poor Children Fall Behind. Cambridge, Mass.: Harvard University Press.

Chen, Q. & Ge, G.C (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes, 26,* 502-514.

Cobb, T. (2010). Learning about language and learners from computer programs. *Reading in a Foreign Language*, *22*(1), 181-200.

Corson, D. **(**1997). The learning and use of academic English words. *Language Learning, 47,* 671-718.

Coxhead, A. (2011). The Academic Word List 10 years on: Research and teaching implications. *TESOL Quarterly*, *45*(2), 355-362.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism, 19,* 121–129.

Eldridge, J. (2008). "No, there isn't an 'academic vocabulary,' but…": A reader responds to K. Hyland and P. Tse's "Is there an 'academic vocabulary'?" *TESOL Quarterly, 42*(1), 109-113.

Engels, L. K. (1968). The fallacy of word-counts. *IRAL-International Review of Applied Linguistics in Language Teaching*, *6*(1-4), 213-232.

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, *28*(2), 241-265.

Gardner, D. **(**2013). *Exploring vocabulary: Language in action.* New York, NY: Routledge.

Gardner, D., & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics*, *35*, 305–327

Ghadessy, P. (1979). Frequency counts, words lists, and materials preparation: A new approach. *English Teaching Forum, 17*, 24-27.

Gilner, L. (2011). A primer on the general service list. *Reading in a Foreign Language*, *23*(1), 65.

Hancioğlu, N., Neufeld, S., & Eldridge, J. (2008). Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes*, *27*(4), 459-479.

Hatch, E., & Brown, C. (1995). *Vocabulary, Semantics, and Language Education*. New York, USA. Cambridge University Press.

Hart, B., & Risley, T. R. (2003). The early catastrophe. The 30-million-word gap. *American Educator, 27*, 4-9.

Heatley, A. Nation, I.S.P, & Coxhead, A. (2002) RANGE program. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure?. *Reading in a foreign language*, *8*, 689-689.

Hosenfeld, C. (1984). Case studies of ninth grade readers. In C. Alderson and A.H. Urquhart (Eds.): *Reading in a foreign language* (pp. 231-249). London, UK: Longman.

Hsueh-chao, M.H. & Nation, P. (2000). Unknown vocabulary density and reading comprehension. Reading in a Foreign Language, *13*(1), 403-430.

Lynn, R. W. (1973). Preparing word lists: a suggested method. *RELC Journal, 4(1),* 25-32.

Nagy, W. , & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition, *Reading Research Quarterly, 47,* 91-108.

Nation, P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press

Nation, P. (2004). A study of the most frequent word families in the British National Corpus. *Vocabulary in a second language: Selection, acquisition, and testing*, (pp. 3-13).

Nation, P. (2008). *Teaching vocabulary: Strategies and techniques*. Boston: Heinle, Cengage Learning.

Nation, P., & Kyongho, H. (1995). Where would general service vocabulary stop and special purposes vocabulary begin?. *System*, *23*(1), 35-41.

Nation, P., & Waring, R. **(**1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds): *Vocabulary: Description, acquisition and pedagogy (*pp. 6–19*).* Cambridge, UK: Cambridge University Press.

Neufeld, S., & Billuroğlu, A. (2005). In search of the critical lexical mass: How 'general'in the GSL? How 'academic'is the AWL? Retrieved from www.lextutor.ca/vp/tr/BNL_Rationale.doc

Neufeld, S., Hancioğlu, N., & Eldridge, J. (2011). Beware the range in RANGE, and the academic in AWL. *System*, *39*, 533-538.

Praninskas, J. (1972). *American university word list*. London: Longman.

Rayson, P. (1996). CLAWS 7 Tagger [Computer software]. Lancaster, U.K.

Richards, J. C. (1974). Word lists: problems and prospects. *RELC Journal*,*5*(2), 69-84.

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know?. *TESOL Quarterly*, (pp.145-171).

Schmitt, N., & Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(04), 484-503. Cambridge University Press.

Sutarsyah, C., Nation, P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based case study. *RELC Journal*, *25*(2), 34-50.

Wang, J., Liang, S.L. & Ge, G.C. (2008). Establishment of a medical academic wordlist. *English for Specific Purposes 27*, 442-458.

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes 28*, 170-182

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