Comparing Academic Vocabulary List (AVL) Frequency Bands to Leveled Biology and History Texts

Lynne Crandall
Brigham Young University

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

Part of the Arts and Humanities Commons

BYU ScholarsArchive Citation

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.
Comparing Academic Vocabulary List (AVL) Frequency Bands
to Leveled Biology and History Texts

Lynne Crandall

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

Master of Arts

Mark E. Davies, Chair
Troy L. Cox
K. James Hartshorn
Robert J. Reynolds

Department of Linguistics
Brigham Young University

Copyright © 2019 Lynne Crandall
All Rights Reserved
ABSTRACT

Comparing Academic Vocabulary List (AVL) Frequency Bands to Leveled Biology and History Texts

Lynne Crandall
Department of Linguistics, BYU
Master of Arts

For decades, teachers and language learners have been concerned about matching the difficulty level of texts to the proficiency level of learners in order to achieve comprehensible input, which leads to effective learning. Some leveling systems and research use word lists as part of their leveling processes, particularly the Academic Word List. The Academic Vocabulary List (AVL) has not been explored yet as a leveling tool, so this study aims to address this lack of research by examining how the AVL words vary in cumulative frequency bands and also in separate frequency bands with regard to level and topic. The AVL was divided into 5 frequency bands and compared against corpora of biology and U.S. history texts at the elementary, junior high, high school, and university levels. Results showed that the biology texts had a higher percentage of total AVL tokens than the history texts did, suggesting that the AVL may be more suitable for some disciplines than others. For the cumulative bands, Bands 1 through 3 proved to have the highest percent deltas, suggesting that words 1 to 800 are the most useful to learn. Looking at each separate band, Bands 1 and 2 had the highest percent of AVL tokens at the high school level, implying that the words of these bands are especially valuable for learners at this level. The university level had the highest percentage for Bands 3 through 5. There was no statistical significance for any band concerning the factor of the relationship between topic and level, but there was statistical significance for the factor of proficiency level at every level. For the factor of topic, there was significance for every band except Band 3. For each band, the elementary and junior high texts were generally similar to each other concerning the AVL tokens found in them. High school and university texts were similar to each other for Bands 1 and 2 but were not similar to each other for Bands 3 through 5.

Keywords: academic vocabulary, readability, AVL, leveling, text difficulty, frequency bands
ACKNOWLEDGEMENTS

I would like to thank my family, friends, and cohort for their endless support and encouragement. I would also like to thank my chair and committee, Dr. Davies, Dr. Cox, Dr. Hartshorn, and Dr. Reynolds, for their time, advice, mentoring, and kindness. Thank you for always being on my side. Finally, I would especially like to thank Dr. Dee Gardner, who inspired this thesis in the first place but passed away before its completion.
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE PAGE</td>
<td>i</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Literature Review</td>
<td>3</td>
</tr>
<tr>
<td>Readability</td>
<td>3</td>
</tr>
<tr>
<td>Ways to Level Texts</td>
<td>4</td>
</tr>
<tr>
<td>Readability formulas</td>
<td>4</td>
</tr>
<tr>
<td>Coh-Metrix</td>
<td>8</td>
</tr>
<tr>
<td>The Lexile Framework</td>
<td>9</td>
</tr>
<tr>
<td>The Fountas and Pinnell Text Level Gradient</td>
<td>10</td>
</tr>
<tr>
<td>Word Lists</td>
<td>11</td>
</tr>
<tr>
<td>Early word lists</td>
<td>12</td>
</tr>
<tr>
<td>The Academic Word List</td>
<td>13</td>
</tr>
<tr>
<td>The Academic Vocabulary List</td>
<td>16</td>
</tr>
<tr>
<td>Research Questions</td>
<td>21</td>
</tr>
<tr>
<td>Methods</td>
<td>23</td>
</tr>
<tr>
<td>Creation of the Corpora</td>
<td>23</td>
</tr>
<tr>
<td>Converting texts</td>
<td>25</td>
</tr>
<tr>
<td>Frequency Bands of the AVL</td>
<td>26</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>28</td>
</tr>
<tr>
<td>Results</td>
<td>29</td>
</tr>
<tr>
<td>Cumulative Vocabulary Coverage</td>
<td>29</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1 *Number of Texts and Word Counts of Each Corpus* .......................................................... 25
Table 2 *Percentage of AVL Tokens* ............................................................................................... 26
Table 3 *Descriptive Statistics for Cumulative Bands by Percentage* ............................................ 30
Table 4 *Biology Band 1 Descriptive Statistics by Percentage* ....................................................... 32
Table 5 *History Band 1 Descriptive Statistics by Percentage* ......................................................... 32
Table 6 *Biology Band 2 Descriptive Statistics by Percentage* ......................................................... 34
Table 7 *History Band 2 Descriptive Statistics by Percentage* ......................................................... 34
Table 8 *Biology Band 3 Descriptive Statistics by Percentage* ......................................................... 35
Table 9 *History Band 3 Descriptive Statistics by Percentage* .......................................................... 35
Table 10 *Biology Band 4 Descriptive Statistics by Percentage* ....................................................... 37
Table 11 *History Band 4 Descriptive Statistics by Percentage* ....................................................... 37
Table 12 *Biology Band 5 Descriptive Statistics by Percentage* ....................................................... 38
Table 13 *History Band 5 Descriptive Statistics by Percentage* ....................................................... 39
LIST OF FIGURES

Figure 1 Percentage means of cumulative bands ................................................................. 30

Figure 2 Biology and history band 1 descriptive statistics .................................................. 33

Figure 3 Biology and history band 2 descriptive statistics .................................................. 34

Figure 4 Biology and history band 3 descriptive statistics .................................................. 36

Figure 5 Biology and history band 4 descriptive statistics .................................................. 37

Figure 6 Biology and history band 5 descriptive statistics .................................................. 39
CHAPTER ONE

Introduction

For decades, teachers and language learners have been concerned about matching the difficulty level of language to the proficiency level of learners. In 1982, Stephen Krashen introduced his input hypothesis, which states that in order to acquire a language, learners need comprehensible input, or input slightly above their proficiency level, described as $i + 1$. Learners can receive input through listening to spoken language and reading written text. Regarding written texts, Brabham and Villaume (2002) explain that students become frustrated with texts that are too difficult and unmotivated with texts that are too easy. Specifically regarding difficulty, texts tend to be incomprehensible when readers understand less than 95% of the vocabulary (Laufer, 1989; Hsueh-chao & Nation, 2000), which often happens for learners with inadequate vocabulary in academic settings (Corson, 1997; Gardner, 2013; Hsueh-chao & Nation, 2000). Of course, the higher percentage of vocabulary readers understand, the more they are able to comprehend, and the lower percentage they understand, the less they are able to comprehend. For most effective learning that leads to acquisition, learners need comprehensible reading texts, which includes appropriate vocabulary.

After recognizing the need for comprehensible input, the natural next step is to determine what input, or text, matches learners’ levels. One way to do this is through leveling texts. Leveling refers to determining the difficulty of a text to facilitate systematic reading instruction, and many different techniques have evolved over time that have been used to determine the difficulty of a text. For example, basal readers have been widely used (Betts, 1946), a huge range of readability formulas exist, and currently, for grades K-12, Common Core Standards in English Language Arts require teachers to place emphasis on reader comprehension (National Governors Association Center for Best Practices, 2010). These standards state that text complexity has
dimensions of quantitative, qualitative, and reader and task considerations. There are different ways to level texts that incorporate various language aspects, but all leveling processes have the same goal: to identify the difficulty of a text in order to inform teachers’ and learners’ decisions about language learning.

Some research and leveling systems use word lists as part of their leveling processes, particularly the Academic Word List (AWL), a widely used word list. Some of these leveling processes are discussed later in the literature review. One word list that has not been explored yet as a leveling tool is the Academic Vocabulary List (AVL) (Gardner & Davies, 2014). Due to this lack of research, the current study explores the potential of the AVL to also be used in leveling processes. To do this, the AVL was divided into five frequency bands and compared against a corpus comprising biology and U.S. history textbooks at the following four levels: elementary school (4th to 5th grade), junior high school (6th to 8th grade), high school (9th to 12th) grade, and university (textbooks used in general education classes at a large university). Even though adult English language learners are unlikely to read textbooks, especially at the elementary level, the principles of text leveling used in this study can be applied to multiple contexts. The purpose of this study is to determine how accurately frequency bands of the AVL discriminate among leveled academic reading texts in the disciplines of biology and U.S. history disciplines. The following research questions are discussed:

1. To what extent do AVL words vary in cumulative frequency bands across biology and U.S. history texts?
2. To what extent do AVL words vary in each separate frequency band with regards to level and topic?
CHAPTER TWO

Literature Review

This literature review will first discuss factors of readability, or text difficulty. Four ways to level texts are then presented, and a history of word lists is reviewed with the most focus on the Academic Word List (AWL) and the Academic Vocabulary List (AVL), including how the AWL has been used as part of leveling processes. The research questions for this study are then presented again.

Readability

One of the earlier definitions of readability was proposed by Dale and Chall (1949):

In the broadest sense, then, readability is the sum total (including the interactions) of all those elements [referring to legibility, interest, ease of reading, or a combination of those] within a given piece of printed material that affects the success that a group of readers have with it. The success is the extent to which they understand it, read it at an optimum speed, and find it interesting. (p. 23)

A simple, more modern conceptualization of readability refers to the difficulty or ease with which a reader can read and understand a text (Oakland & Lane, 2004), and this is the definition by which this thesis refers to readability.

There are many variables that affect the difficulty of a text, including both reader factors and text factors. A variety of reader factors exist, such as motivation, background knowledge, and previous reading experience, but since these factors are primarily dependent on the reader, text factors generally receive the most attention regarding text difficulty (Fulcher, 1997). One of the most common text factors that affects difficulty is syntactic complexity (Berendes et al., 2018; Chen, 2016; Eslami, Fulcher, 1997; 2014; Liontou, 2012; Liu & Chiu, 2009; Vajjala & Meurers, 2014), sometimes quantified by readability formulas that primarily deal with sentence
length. Morphological features can also play a role in the difficulty of a text (Berendes et al., 2018) as well as vocabulary (Chen, 2016; Fulcher, 1997; Vajjala & Meueres, 2014). Other factors include the author’s voice (McTigue & Scott, 2010), information and new concept density (Liu & Chiu, 2009), psycholinguistic features such as word familiarity (Vajjala & Meueres, 2014), and coherence (Berendes et al., 2018; Halliday & Hasan, 1976; Liontou, 2012; McTigue & Scott, 2010). These factors are interrelated, and they contribute to the difficulty of a text and play a role in the decision-making process of material developers, teachers, and learners as they create or choose reading texts.

**Ways to Level Texts**

It is clear that evaluating text complexity is multi-dimensional, but the following section will focus on just a few aspects of this widely varied and diverse field. Six of the most commonly used readability tests are examined in the following section, and while these tests don’t take into account vocabulary frequency, it is still beneficial to understand them because much of the literature and research on readability refer to these tests. The Coh-Metrix, the Lexile Framework, and the Fountas & Pinnell text level gradient are then presented, all of which do involve word frequency as well as other factors to determine text readability.

**Readability formulas.** Readability formulas are often used to match L1 readers with an appropriate difficulty level of L1 reading material. These formulas assess the difficulty of a text quickly and objectively (Gunning, 2003; Kotula, 2003), often by taking into account sentence complexity and vocabulary difficulty. The formulas generally apply a mathematical equation to a portion of text and produce a numerical score, often defined on a grade-level continuum. The mathematical formulas typically measure sentence complexity by sentence length, and they measure vocabulary difficulty by word length or by comparing lists of easy or difficult words (McLaughlin, 1969). The following six readability tests are among the most commonly used.
**Flesch Reading Ease.** The Flesch Reading Ease formula is one of the oldest and most well-known readability tests. This formula was designed to practically and objectively measure the difficulty of a text. Rudolph Flesch (1948) developed this formula by revising a previous readability formula he had created. There are two components required to calculate the readability ease (RE): the average sentence length (ASL, the number of all words divided by the number of sentences) and the average number of syllables per word (ASW, the number of syllables divided by the number of words). \[ RE = 206.835 - (1.015 \times ASL) - (84.6 \times ASW). \] The result is a number between one and 100. A score between 60 and 70 is considered standard and operates under the assumption that readers understand 50% to 70% of text.

**Flesch-Kincaid Grade Level.** Because the results from the Flesch Reading Ease Formula need to be converted into grade levels, the Flesch-Kincaid Grade Level formula was created for an easier comprehension of results. This formula is based on the Flesch Reading Ease Formula, but it is adjusted so that the results can immediately be applied without consulting a conversion table. Rather than resulting in a number between one and 100, the results are based on the U.S. education system, with the number 12 referring to the senior year in high school. Rudolph Flesch co-authored this formula with John. P. Kincaid, and it is commonly used in the field of education (Readability Formulas, n.d.c). This readability index is calculated by taking the average sentence length and the average number of syllables per word, similar to the Flesch Reading Ease Formula. Then, the following formula is used: Flesch-Kincaid Reading Age = (0.39 x ASL) + (11.8 x ASW) – 15.59.

**Gunning Fog Index Readability Formula.** The FOG Index is similar to the Flesch-Kincaid Grade Level because its results are also related to the U.S. school system. One unique limitation is that it discounts the notion that not all multisyllabic words are equally difficult, and therefore the length of words does not always represent the difficulty of the word. It was created
by Robert Gunning, an American textbook publisher, (Readability Formulas, n.d.d). The FOG Index formula comprises the following steps. First, one takes a sample passage of at least 100 words and then counts the number of words and sentences. Afterwards, one would divide the total number of words by the number of sentences. This results in an average sentence length (ASL). The next step is to count the number of words with three or more syllables that are not proper nouns, combinations of easy words (hyphenated words), or two-syllable verbs made into three-syllable words by just adding –es and –ed endings. After this is completed, one divides this number by the number of words in the sample passage, which results in the percent of hard words (PHW). The Grade Level equals 0.4(ASL + PHW). The general rule is that short sentences with fewer hard words achieve a higher readability score than long sentences written with harder or longer words.

**Coleman-Liau Index.** This readability formula keeps the same assumption of reader comprehension but is different from other formulas because it uses characters as the smallest unit of measurement, not words. This is because its creators “believed that computerized assessments understand characters more easily and accurately than counting syllables and sentence length” (Readability Formulas, n.d.b). It was designed by linguists Meri Coleman and T. L. Liau to aid the U.S. Office of Education in calibrating the reading levels of textbooks in the public school system. To calculate the Coleman-Liau Index, one would find the average number of letters per 100 words (L) and the average number of sentences per 100 words (S) and use the following formula: CLI = 0.0588L – 0.296S – 15.8. The result is a number that correlates with U.S. grade levels.

**Automated Readability Index (ARI).** This readability test is similar to most other tests in that it is calculated using ratios of word difficulty and sentence difficulty. It is most similar to the Coleman-Liau Index because it also relies on characters rather than syllables per word. To
calculate this index, one would find the average word length (characters per word) and the average sentence length (words per sentence). Readability = 4.71(average word length) + 0.5(average sentence length) − 21.43. This formula also results in numbers that correlate with U.S. grade levels (Readability Formulas, n.d.a).

**SMOG Readability Formula.** All of the previous formulas assume that readers will comprehend 50% to 70% of the texts. The SMOG Readability Formula, created by G. Harry McLaughlin (1969), is unique because it is based on an assumption that 100% of the text will be comprehended by the reader. The formula uses 10 sentences from each the beginning, middle, and end of the text being assessed. After the sentences are chosen, the readability result is calculated by counting the number of words with three or more syllables in each group. Then the square root of that number is taken and rounded to the nearest 10. Finally, adding three to that number determines the SMOG grade. The result is a number that correlates with the U.S. education system beginning with the 5th grade.

All six readability tests use average word and sentence lengths to calculate their readability scores, but there are some unique factors. The ARI and the Coleman-Liau are different than the other tests because they use number of characters rather than syllables to calculate the average word length. The SMOG is the only test that assumes the reader comprehends 100% of the passage. Another unique factor is the contexts in which each readability formula is most appropriate. The Flesch Reading Ease formula is generally considered to be useful for any kind of text, while the Gunning Fog Index is ideal for business publications. The Automated Readability Index is ideal for technical documents and manuals, and the Flesch-Kincaid formula is also more suited for technical documents. Both the SMOG and the Coleman-Liau target proficiency levels ranging from the 4th grade to university levels (Readability Formulas, n.d.e).
Despite the wide use of readability formulas, some researchers have criticized them. McTigue and Slough (2010) point out that one weakness is readability formulas’ inability to accurately measure and quantify vocabulary complexity. Various formulas include measures of word length, number of syllables, and word frequency, but while these measures may correlate to vocabulary complexity, they are unable to demonstrate all of the aspects that are part of what makes vocabulary complex. Another limitation, due to the nature of the formulas, is their inability to address the communication of meaning, such as discriminating between discourse and nonsensical words (Rush, 1985). Also, several studies have demonstrated the weaknesses of readability formulas through their empirical results (Davison & Kantor, 1982; Fulcher, 1997; Leroy, Helmreich, & Cowie, 2010; Pitcher & Fang, 2007). Readability formulas did not explain why participants labeled certain sentence structures as difficult (Leroy, Helmreich, & Cowie, 2010) and ranked texts in order of perceived difficulty (Fulcher, 1997). Reading levels were not accurately reflected by readability formulas (Pitcher & Fang, 2007), and sentence length, a key component of readability formulas, was shown to not always contribute to text complexity (Davison & Kantor, 1982). These criticisms validly challenge the effectiveness of readability formulas, but the convenience of these formulas often weighs more heavily in researchers’ decisions to use them.

**Coh-Metrix.** One widely used alternative way to measure readability is the Coh-Metrix Second Language Reading Index (Crossley, Allen, & McNamara, 2011; Crossley, Greenfield, & McNamara, 2008; Green, Unaldi, & Weir, 2010; Plakans & Bilki, 2016). It was created in 2004, and it differs from previous readability metrics because it focuses on measuring the cohesion and coherence of a text (Graesser, McNamara, Louwere, & Cai, 2004). These measures are based in theory and findings from computational linguistics and psycholinguistics, and features of the Coh-Metrix tool stem from earlier research by McNamara, Kintsch, E., Songer, and Kintsch, W.
Their study found that readers with less background knowledge about the topic of the text benefited from more cohesion, while readers familiar with the topic benefited from cohesion gaps (McNamara et al., 1996). There are thirteen primary measures with 200 other optional measures for each inputted text. The primary measures range from word and sentence dimensions to paragraph and discourse dimensions. Several specific measures include syntactic complexity, co-reference cohesion, causal cohesion, type to token ratio, word frequency, and readability, measured with the Flesch-Reading Ease and Flesch-Kincaid Grade Level readability formulas (Graesser, McNamara, Louwere, & Cai, 2004).

Several studies have shown the Coh-Metrix Index to be a better readability measure than traditional readability formulas (Crossley, Allen, & McNamara, 2011; Crossley, Greenfield, & McNamara, 2008; McNamara, Louwere, McCarthy, & Graesser, 2010). Findings show that the Coh-Metrix Index is better at distinguishing texts with high versus low cohesion and predicting reading difficulty, and it is also more closely aligned with intuitive text processing used by L2 material writers.

The Lexile Framework. The Coh-Metrix Index and the six readability formulas discussed earlier focus on analyzing the text itself to determine the difficulty level, but other tools have been created that measure both text complexity and reader ability, such as the Lexile Framework (Stenner, 1996). Texts are analyzed with an algorithm to compute text complexity, and readers are given assessments that determine their reading proficiency level, or Lexile. The goal behind this is that by measuring both text and reader levels, books can be much more accurately matched with the students’ reading ability, allowing readers to be challenged sufficiently but not overburdened by a lack of comprehension.

The Lexile Framework determines levels by comparing reading items with a Rasch scale (Stenner, Burdick, H., Sanford, & Burdick, D., 2006); the main components in determining the
Lexile are sentence length and vocabulary use. Lexile ranges can include any number from 100L to 1700L, the L referring to Lexile, and anything below the Lexile is simply labeled as beginning reader (BR). The students’ ability to read is also given a Lexile score based on standardized state tests, reading programs, or other assessments.

Since their inception, Lexile scores have become a widely used tool in K–12 educational settings that inform and assist teachers, parents, and students about what reading material would be most appropriate. More specifically, the Lexile Framework is utilized by teachers in an L1 classroom setting to pair students with level-appropriate texts. After students receive a score based on their reading, a text’s Lexile level must be looked at, which can be found in the book or by searching a database of books online that have been given Lexile measures. It is recommended that L1 readers choose a text that is within a certain range of their Lexile reader measurement, not choosing a book that is too easy (100L below their range) or too difficult (50L above their range) (MetaMetrics, n.d.b).

**The Fountas and Pinnell Text Level Gradient.** Just as the Lexile Framework determines readability by looking at the text and reader, so does Fountas and Pinnell’s (1996) F&P Text Level Gradient. This system was created with the purpose to help young readers face challenging texts successfully by making sure the text elicits reading strategies, vocabulary, and syntax that the readers have already mastered or are familiar with, while also requiring readers to expand their current processing strategies (Fountas & Pinnell, 1999).

The F&P Text Level Gradient, similar to some readability tests, classifies books according to the following parameters: word count, variety of vocabulary, frequency of vocabulary, sentence length, and sentence complexity (Fountas & Pinnell, 1996). These first five parameters help account for the level of linguistic difficulty. Going beyond readability and Lexile measures, the F&P Text Level Gradient also considers the genre, text structure, content,
themes and ideas, literary features, illustration support, and book and print features (Fountas & Pinnell Literacy, 2018). The combinations of all these parameters becomes a level guide for the selection of books. The resulting reading levels are specified as A–Z+, and they are divided into A–D for kindergarten, E–J for first grade, K–M for second grade, N–P for third grade, Q–S for fourth grade, T–V for fifth grade, W–Y for sixth grade, Z for seventh and eighth grade, and Z+ for high school and above. In fact, Fountas and Pinnell themselves have published book lists that follow the specifications of their level gradient. Teachers can use these lists to appropriately select material that will help their developing readers.

Word Lists

Many of the aforementioned leveling systems use vocabulary as one of their components. As mentioned earlier, readability formulas often use the number of syllables to determine the difficulty of a word with more syllables correlating to a higher word difficulty. This is not always an accurate depiction of word difficulty because the difficulty of words can stem from the fact that they have a low frequency and readers have not encountered them before. However, the fact that readability formulas and other leveling systems include vocabulary as one of their components helps show the importance of vocabulary. Gardner (2013) refers to vocabulary as the “fuel of language, without which nothing meaningful can be understood or communicated” (p. 2). This is especially true in academic situations, where academic vocabulary is essential to success (Beck, Perfetti, & McKeown, 1982; Biemiller, 2003; Corson, 1997; Nagy, W., Townsend, D., Lesaux, N., Schmitt, N., 2012; Townsend, Filippini, Collins, & Biancarosa, 2012). Some researchers have asked what vocabulary is most valuable for ELLs to learn (Gardner, 2013; Webb & Chang, 2012). Gardner (2013) also describes how not all vocabulary is the same because some words are used in different situations and may have different meanings. To tailor teaching to specific sets of learners, researchers have created word lists, organized by
frequency, to facilitate learning useful vocabulary for different learners. These word lists can play an important role in text leveling because they represent the difficulty of words more accurately than readability formulas do.

**Early word lists.** In 1953, West created the General Service List (GSL), a list of 2,000 core word families created from a corpus that had approximately five million words. His decisions about vocabulary were mostly based on word frequency. The GSL became widely used despite later criticism about its size (Engels, 1968) and age (Richards, 1974). Sixty years after its creation, the New General Service List was published, which adapted the GSL using a larger, more modern corpus (Browne, 2014). However, because GSL did not address academic vocabulary, several academic-specific word lists have been created (Campion & Elley, 1971; Coxhead, 2000; Gardner & Davies, 2014; Ghadessy, 1979; Lynn, 1973; Praninskas, 1972). One researcher compared the GSL to two graded-reader wordlists (Wan, 2008). The results showed that there were 1,497 words that occur in the series but not on the GSL, which suggests that using the GSL as a placement test is problematic. However, for the first most frequent 1,000 word families in the GSL, 93.03% and 96.16% were represented in the two respective graded-reader series. For the second 1,000 word families, the coverage dropped from 76.07% to 82.34%. These data suggest that studying the words on the GSL is beneficial for learners to be able to understand high-frequency vocabulary included in the graded-reader books (Wan, 2008).

The University Word List (WSL) (Xue & Nation, 1984) was the first extensive academic word lists and was created by editing and combining four other academic word lists created by Campion & Elley (1971), Praninskas (1972), Lynn, (1973), and Ghadessy (1979). The UWL also became widely used, but because it was built from four other studies, it lacked consistency and inherited limitations from its constituent lists.
**The Academic Word List.** In 2000, Coxhead created the AWL after noting that the UWL was created from several other limited lists. One of the main limitations was the fact that the lists were created from small corpora, so Coxhead addressed the issue of size in her corpus in addition to the issues of representation, organization, and word selection (Coxhead, 2000).

A key issue Coxhead addressed was word selection, or what counts as a word. The AWL is organized by word family, or base words with all inflections and transparent derivations (Coxhead, 2000), based on research by Bauer and Nation (1993), which suggests that comprehending derivational morphemes of a word family does not require much more effort by learners if they know the base word and understand basic word-building processes. Coxhead specified three other criteria in word selection for the AWL:

1) word families were not included if they were already part of the GSL,
2) a member in the word family had to occur at least 10 times in each of the 4 disciplines and in 15 or more of the 28 subject areas, and
3) members in the word family had to occur at least 100 times in the corpus (Coxhead, 2000).

To address the issue of size, Coxhead created a corpus of about 3.5 million words, based on research that suggested a corpus needed to be that size in order for each word family to appear at least 100 times in the corpus (Coxhead, 2000; Francis & Kucera, 1982). This word count would account for a word family appearing around 25 times in each of the four disciplines included in the corpus, thus providing a large enough sample size for a sufficient frequency of academic words. The four disciplines in the corpus, composed of 28 smaller subjects, address the issues of representation and organization. Coxhead included academic texts from a variety of sources in the art, commerce, law, and science disciplines in order to provide an accurate representation of academic vocabulary (2000). However, the majority of these texts were
published in New Zealand, perhaps weakening the claim of wide representation, as noted by Gardner and Davies (2014). The AWL comprises the 570 most frequent academic word families across college-level texts. This word list is a seminal publication of scholarship of this area and has become the standard vocabulary list for students preparing for academic study (Davies & Gardner, 2014; Nagy et al., 2012).

Part of the AWL’s wide usage is in the field of text difficulty. A variety of studies related to text difficulty use the AWL as part of their methods. Two studies used the AWL to determine how well their newly created language materials, a corpus and a vocabulary listening test respectively, differentiated between different levels of text (Mclean, Kramer, & Beglar, 2015; Pendar & Chapelle, 2008). Other studies involved leveling texts and used the AWL as part of their leveling process (Green, Unaldi, & Weir, 2010; McAlister, 2010; Miller, 2011; Sabet & Minaei, 2017; Vajjala & Meurers, 2012). Specifically, Vajjala & Meurers (2012) studied the impact of a range of lexical and syntactic features on predicting text complexity of a corpus of leveled educational articles. One of the lexical features they included was the proportion of words in a text which were found in the AWL. The AWL feature “turned out to be one of the most predictive features” and was in the top ten best predictive features from both the lexical and syntactic features (Vajjala & Meurers, 2012, p. 167). Also, some studies have used the AWL when choosing which words to use (Poole, 2012), measuring the difficulty of vocabulary that students chose to use in a task (Green & Hawkey, 2011), and analyzing the difficulty of words that students looked up in dictionaries (Pritchard, 2008). The variety of studies that have utilized the AWL in processes related to text leveling demonstrates just how much the AWL is used in these processes.

In addition to being used in research, the AWL is used in several online text leveling websites and programs. One widely used text readability program is Paul Nation’s Range
program, a downloadable software program that analyzes text based on word frequency. The first 2,000 words of the GSL are used as well as the AWL. Some online websites are similar to the Range program. Extensive Reading Central (https://www.er-central.com) offers users a choice between various word lists. It divides the lists into levels by frequency order and compares texts to those levels. The AWL is one of these lists, and it is paired with the GSL, which comprises levels 1 and 2 while the AWL is classified as level 3. Similarly, Frequency Level Checker (http://language.tiu.ac.jp/flc/index.html), a more rudimentary website, also uses the GSL to designate levels 1 and 2 and the AWL to designate level 3.

There are several other websites that use the AWL in their leveling process. Text Inspector (https://textinspector.com) offers the AWL as one way to analyze a text. It looks at percentages of types and tokens of all AWL words in a text and gives a correlated CEFR level. Another tool was developed by Educational Testing Service and is called TextEvaluator. This system uses ten components to level a test, the first being academic vocabulary. Within that component, there were ten features that were considered, two of which were based on the AWL (Sheenan, 2016). Finally, Lextutor (https://lextutor.ca/) uses the Lexical Frequency Profiler (Laufer, & Nation, 1995), which uses AWL as one of its frequency levels. Although these online leveling tools, as well as the previously mentioned research, use the AWL in different ways, they are all similar in that they use the AWL as part of a text leveling process.

The AWL has proved to be a great improvement over previous word lists and has been widely used throughout the world. However, despite the AWL’s popularity and wide usage, some researchers have questioned whether or not this list adequately represents core academic vocabulary (Chen & Ge, 2007; Gardner & Davies, 2014), and Gardner and Davies (2014) identified two specific limitations of the AWL. The first limitation is using word families to organize the list because it can be difficult for learners to learn all of the related words in a word
family. For example, the base word *react* means *respond* while the derivations of react have varied meanings: *reactionary* (strongly opposed to social or political change), *reactivation* (to make something happen again), and *reactor* (a device or apparatus) (Gardner & Davies, 2014).

Also, word families do not address how word meanings may change from one discipline to another, such as *major*, which refers to a military rank, musical chord, or area of study depending on the discipline in which it is used. Another consideration is how word meanings may change depending on part of speech. One example is the word *proceeds*. As a verb, it is the third person singular form of *proceed*, which means *to continue*, but as a noun, *proceeds* means *profits*. The second limitation is the fact that the AWL was built on top of the GSL, which means that some academic vocabulary in the GSL is not included in the AWL. It is possible that core academic vocabulary was included in the GSL and therefore excluded in the AWL. Because of these limitations, Davies and Gardner (2014) created a new list: the Academic Vocabulary List (AVL), which aims to improve upon the AWL and other previous word lists.

**The Academic Vocabulary List.** In order for the AVL to avoid similar limitations inherent in the AWL, it was created with the following key considerations (Gardner & Davies, 2014):

1. The new list must initially be determined by using lemmas, not word families.
   
   Subsequent groupings of the list into families may be warranted for certain instructional and research purposes.

2. The new list must be based on a large and representative corpus of academic English, covering many important academic disciplines.

3. The new list must be statistically derived (using both frequency and dispersion statistics) from a large and balanced corpus consisting of both academic and non-academic materials. The corpus must be large enough and the statistics powerful
enough to be able to separate academic core words (those that appear in the vast majority of the various academic disciplines) from general high-frequency words (those that appear with roughly equal and high frequency across all major registers of the larger corpus, including the academic register), as well as from academic technical words (those that appear in a narrow range of academic disciplines).

4. The academic materials in the larger corpus, as well as the non-academic materials to which it will be compared, must represent contemporary English, not dated materials from 20 to 100 years ago. Otherwise, the validity of the new list could be questioned.

5. The new list must be tested against both academic and non-academic corpora, or corpus-derived lists, to determine its validity and reliability as a list of core academic words. (p. 312)

These considerations guided the creation of the AVL. The first consideration was fulfilled by using lemmas (base forms of words with inflectional affixes) instead of word families (base forms of words and all affixes, including derivational morphemes). The corpus used to create the AVL, the Corpus of Contemporary American English (COCA), tags words for grammatical parts of speech using the CLAWS 7 tagger from Lancaster University, which aided in the lemmatization process (Gardner & Davies, 2014). A later list of word families was created to fulfill certain research and instructional purposes, but that list was created from the lemmas of the AVL. As previously mentioned, the AVL was created from COCA, a 425 million word modern English corpus, specifically its 120 million-word academic subcorpus, which fulfills the second and fourth considerations. The academic subcorpus was created from academic journal articles, and the following nine disciplines were included: education; humanities; history; social
science; law and political science; science and technology; medicine and health; business and finance; and philosophy, religion, and psychology. The variety of disciplines used in the corpus promotes an accurate representation of core academic vocabulary in the corpus.

The third consideration was fulfilled by using several statistical tests to separate academic words from other words in the corpus. First, academic words were required to appear 50% more often in academic materials than in the general materials, thus eliminating non-academic words. Second, core academic words were required to appear in at least seven disciplines with at least 20% of the expected frequency, based on size of the discipline. This requirement helped eliminate technical, discipline-specific words from the list of core academic vocabulary. Third, lemmas in the core were required to have a dispersion of at least 0.80. The dispersion measure shows how a word is spread across the corpus, with 0.01 meaning the word only occurs in a small part of the corpus and 1.00 meaning the word is perfectly dispersed throughout the whole corpus. For example, the following words had a dispersion of 0.95: employ, definition, frequently, emphasis, and primarily. They are broader than these words with dispersions of 0.84 to 0.87: psychological, flow, climate, experiment, and waste. These words with a lower dispersion are more specific to certain domains than the words with a higher dispersion are. There is no particular dispersion measure recommended by research; 0.80 was chosen after multiple tests showed it eliminated technical vocabulary while keeping more core academic vocabulary (Gardner & Davies, 2014). The final measure was also designed to help eliminate discipline-specific words. This measure stated that a “word cannot occur more than three times the expected frequency (per million words) in any of the nine disciplines” (Gardner & Davies, 2014, p. 316). These four criteria helped exclude general high-frequency words and also technical, discipline-specific words.
The fifth and final consideration was fulfilled by comparing coverage of the AVL with coverage of the AWL in COCA and the British National Corpus. In order to make a direct comparison between the two lists, the AVL needed to be converted from a list of lemmas into a list of word families because the AWL is made from word families. The comparisons were made with case studies of the top 570 AVL word families, which also helped in making direct comparisons with the 570 word families of the AWL. The first case study showed that the AVL consistently covered a higher percentage of words in academic materials (COCA, 13.8%; BNC, 13.7%) than in newspapers (COCA, 8.0%; BNC, 7.0%) or fiction (COCA, 3.4%; BNC, 3.4%). The second case study showed that the AVL consistently covered a higher percentage of words than the AWL in COCA academic subcorpus (AVL, 13.8%; AWL, 7.2%) and BNC academic subcorpus (AVL, 13.7%; AWL, 6.0%). The data from the case studies demonstrate a significant difference between the AVL and the AWL with regards to coverage. However, it is important to note that the AWL excludes some higher frequency words in the GSL while the AVL does not, so a direct comparison of coverage, though compared with word families, may still not be fair. It is logical to assume that the AVL will naturally have a higher coverage because it contains higher frequent words that are also on the GSL.

As with any list, there are limitations to the AVL. The authors recognize that more research needs to be done to identify core multiword academic vocabulary and core spoken academic vocabulary (Gardner & Davies, 2014). Despite these limitations, the core considerations of the AVL successfully avoid the limitations of the AWL, and the AVL is a current, accurate, and comprehensive academic word list intended to aid people learning English for academic purposes.

The creators of the AVL expressed that their hope that the AVL would be used “to improve the learning, teaching, and research of English academic vocabulary in its many
contexts” (Gardner & Davies, 2014, p. 325). The AVL has been examined and used in several ways. One way the AVL has been used is inspiration for the creation of other word lists. Several researchers noted that existing discipline-specific word lists are usually developed using Coxhead’s (2000) method, which excludes general high-frequency words (Lei & Liu, 2016). After examining the work Gardner and Davies (2014) did with the AVL and the challenges they made to the AWL, Lei and Liu (2016) created a new medical academic vocabulary list (MAVL). They followed the same method of creation and testing as the creation of the AVL used in addition to one principle of frequency used in the AWL. Their MAVL had a better coverage of medical English and was 53% shorter than the existing medical academic word list created by Wang, Lian, and Ge (2008).

Three other word lists utilize the same methodology behind the creation of the AVL (Brezina & Gablasova, 2015; Dang, Coxhead, & Webb, 2017; Green & Lambert, 2018), although one does not directly cite the AVL as their foundation of decision making (Brezina & Gablasova, 2015). The new-GSL, created in 2015, is based on the GSL but departs from the GSL’s organization by word family and adopts an organization by lemma, similar to the organization of the AVL (Brezina & Gablasova, 2015). Second, the Academic Spoken Wordlist (ASWL), created in 2017, followed Gardner and Davies’s (2014) approach by creating an entirely new list, thus avoiding limitations that stem from building word lists on other general high-frequency vocabulary lists (Dang, Coxhead, & Webb, 2017). The ASWL also used comparable statistical measures in its creation used in the AVL, particularly dispersion. Finally, the Secondary School Vocabulary Lists were created following the AVL’s creation with regards to range, dispersion, range ratio, and frequency ratio (Green & Lambert, 2018). These lists are discipline-specific lists for eight core subjects. The formation of all these wordlists with features either inspired by or similar to the AVL shows one aspect of the AVL’s utility.
The AVL has also been used as part of other research studies (Frankenberg-Garcia, 2018; Mirzaei, Meshgi, Akita, & Kawahara, 2017; Wingrove, 2017). Frankenberg-Garcia explored collocational choices of academic English users at a British university. She chose ten nouns from among the most 50 frequent nouns in the AVL to serve as the bases to elicit collocation. Another researcher used the AVL to quantify the amount of academic vocabulary used in TED talks and academic lectures (Wingrove, 2017). He compared the percentage of TED talk and academic lecture text found in the academic core of the AVL (the first 500 lemmas) and the percentage of text found in the rest of lemmas of the AVL. More academic text was found in the academic lectures than in the TED talk texts. Mirazaei et al. (2017) also used TED talks in their study. The videos of the talks were used to examine the effectiveness of partial and synchronized captioning (PSC) compared to full captioning. In this study, both the AWL and the AVL were used to achieve high accuracy of the academic word specificity, and results showed no statistical significance between the PSC and full captioning. These three studies show the potential for the AVL to be used in research studies as a reference and as a standard to compare vocabulary coverage.

**Research Questions**

In a later article about the AVL, Davies and Gardner (2016) stated that “the AVL…has a useful place in academic training and research, and we hope that others will continue to investigate these possibilities” (p. 68). The AVL has been used in research in various ways, but so far it has not been used as a leveling tool. Since vocabulary plays an important role in most leveling systems and many research studies as well as online text leveling websites have used the AWL and GSL as part of their leveling processes, it is worth investigating the possibility of using the AVL as a leveling tool as well. Evaluating how the AVL discriminates among accepted leveled texts will show if it accurately reflects accepted leveling processes. This will improve its
validity and help fulfill its creators’ desire for others to investigate the possibilities of the AVL’s potential role in academic training and research (Gardner & Davies, 2016). As mentioned in the introduction, the study aims to answer the following questions:

1. To what extent do AVL words vary in cumulative frequency bands across biology and U.S. history texts?

2. To what extent do AVL words vary in each separate frequency band with regards to level and discipline?
CHAPTER THREE

Methods

This chapter describes the process used for comparing the AVL to the corpora of biology and U.S. history textbooks created for this study. It includes a description of the textbooks chosen for the corpora, how the corpora were compiled, how the AVL was divided into frequency bands, justification for that division, and how the data were analyzed.

Creation of the Corpora

To answer the research questions and determine the AVL’s potential as a leveling tool, corpora were created of four proficiency levels (elementary, junior high, high school, and university) for two different disciplines: biology and U.S. history. These two disciplines were chosen because they are inherently different from each other and would be likely to have little overlap of vocabulary. They are also subjects that are taught at every proficiency level. The textbooks used in the corpora were independent from the texts used to create the AVL because the AVL was created from academic journal articles. The corpora comprise textbooks in those disciplines or sections of textbooks in those disciplines. When only sections of textbooks were used, those sections corresponded to similar topics found in other textbooks in the disciplines, and the rest of the textbooks were not used because the topics were not related. For example, some elementary and junior high school science textbooks contained sections on biology, chemistry, physics, and earth science, so only the biology sections were included in the corpus. The decision to use only the related sections of textbooks kept the integrity of the discipline and eliminated a potential undesirable variable to consider when analyzing the comparison.

Other variables considered were types of text, publishers, dates of publication, authors, and number of textbooks to include per corpora. All of the textbooks used were expository textbooks found in the extensive curriculum section of the Harold B. Lee Library located at
Brigham Young University or on the Fall 2018 BYU Biology 100 and History 220 and 221 (U.S. History) class booklist. Because of the expository nature of the texts, there was a lack of personal pronouns found in the texts, although there were several quotations in certain textbooks that contained personal pronouns. Some researchers have shown support for the hypothesis that different publishers show variances of text difficulties (Berendes, et al., 2018). In their study, Berendes, et al. (2018) examined the linguistic complexity of German textbooks at different grade levels and from different publishers. They found meaningful differences between publishers in the way they classified texts. These results suggest that different publishers level their texts differently. To help avoid the variable of publisher differences, there was a total of 15 different publishers included in the corpora, and the dates of publication ranged from 1997 to 2018. These differences added variety to the corpora and decreased the possible effect of certain publishers or publication dates favoring certain vocabulary or leveling certain vocabulary differently. Most textbooks were written by multiple authors with PhD degrees; however, there was one middle school biology textbook written by only one author. All of the textbooks had multiple consultants or reviewers, which decreases the likelihood of one author’s voice overly affecting the vocabulary of a textbook. It’s possible that when sections of textbooks were used, a single author wrote that section because that was his or her area of expertise. However, the textbooks made no mention of that, and the multiple authors listed were listed as authors of the whole book with often many consultants listed too. Regarding the number of textbooks used per corpora, at least three textbooks per section were used, except for biology high school. Some sections, particularly the elementary levels, contained more texts because those texts were significantly shorter than the textbooks used for higher proficiency levels. Table 1 shows the number of texts and total word count per discipline per level. A careful consideration of what
textbooks to use in the corpora decreased the number of external variables and their potential to weaken the integrity of the comparisons of the corpora to the AVL.

Table 1

*Number of Texts and Word Counts of Each Corpus*

<table>
<thead>
<tr>
<th>Proficiency</th>
<th>Biology</th>
<th>U.S. History</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Texts</td>
<td>Word Count</td>
</tr>
<tr>
<td>Elementary</td>
<td>5</td>
<td>323,166</td>
</tr>
<tr>
<td>Junior High</td>
<td>3</td>
<td>1,890,234</td>
</tr>
<tr>
<td>High School</td>
<td>2</td>
<td>806,799</td>
</tr>
<tr>
<td>University</td>
<td>3</td>
<td>681,396</td>
</tr>
</tbody>
</table>

*Note.* These numbers are the total word count of the texts. Only the function words were considered in the analyses.

Converting texts. All content of the textbooks was scanned and saved in a .pdf format. Then the files were converted into Word documents and saved in a .docx format using ABBYY FineReader, an optical character recognition program. The .docx files were then concatenated into one file per textbook and saved as plain text files. These file format conversions were necessary in order to count the number of words in the corpora and also to compare the vocabulary in the corpora to the AVL.

The plain text files were then tagged for part of speech using the CLAWS 7 tagger (http://ucrel.lancs.ac.uk/claws/), which resulted in files that had one word and one part of speech per line. These were imported into a relational database (MS SQL Server 2012 R2), where lemma information was added for each { word form / PoS tag } combination. We then ran a simple SQL command to import into a new table the frequency of each { lemma / PoS } pair in each text. For example { glucose / noun } occurred 8 times in the elementary biology text B1d, 27 times in the junior high school text B2b, 99 times in the high school text B3b, and 108 times in the university text B4c (see Appendix A for a list of textbooks used in the corpora).
We then assigned to each “word” (a lemma / PoS pair) a number that corresponded to each of the five frequency levels in the AVL: 1 for words 1–200, 2 for 201–400, 3 for 401–800, 4 for 801–1500, and 5 for 1501–3100. (Obviously, words that are not in the AVL did not receive a number.) For each of the 28 texts in the corpus, we then counted the number of words from each AVL level as well as the percentage of all nouns, verbs, adjectives, and adverbs in that text that were from the different AVL levels. For example, in the elementary school history text H1c, there are 132,040 nouns, verbs, adjectives, and adverbs. The number of words for each AVL level and the corresponding percentage of all tokens in each level are shown in the Table 2.

Table 2

*Percentage of AVL Tokens*

<table>
<thead>
<tr>
<th>AVL level</th>
<th>1 (1–200)</th>
<th>2 (201–400)</th>
<th>3 (401–800)</th>
<th>4 (801–1500)</th>
<th>5 (1501–3000)</th>
</tr>
</thead>
<tbody>
<tr>
<td># tokens</td>
<td>10539</td>
<td>2550</td>
<td>2310</td>
<td>939</td>
<td>247</td>
</tr>
<tr>
<td>% all tokens</td>
<td>0.08</td>
<td>0.10</td>
<td>0.12</td>
<td>0.12</td>
<td>0.13</td>
</tr>
</tbody>
</table>

With the token count for each of the 28 texts, we were then able to create aggregate totals by text level as well (elementary, junior high, high school, and university), although most of the statistical tests used the frequency as input in each of the 28 individual texts.

**Frequency Bands of the AVL**

Frequency is generally accepted as an indicator of word difficulty, and the more frequently a word occurs, the more important it is to learn (Kremmel, 2016). For example, in the AVL, word 1, *study*, is more frequent and therefore more beneficial to learn than word 3,000, *privation*. Any system with frequency bands or fixed categories will inherently possess limitations, though. For example, if there were a division of bands at word 200, one could question the significance of the difference between word 199 and word 201. However, Kremmel posits that “although vocabulary learning does not follow a strict frequency order for each
individual word, it certainly does in larger clusters (bands), at least at the higher frequency levels. Therefore, a system of fixed categories makes sense both from an assessment as well as a theoretical and pedagogical perspective” (2016, p. 980). Due to this reasoning, this study used frequency bands of the AVL as a fixed category system.

Kremmel (2016) also challenges the assumption that frequency bands should be divided into bands of 1,000 words, the traditional method of dividing frequency bands. In his study, he examined the percentage of coverage in COCA provided by the lemmas of the COCA frequency list in increments of 500. The first 500 most frequent content words provide 26.73% of the coverage with the following 5 bands providing an additional 5.82%, 3.51%, 2.43%, 1.80%, and 1.4%. These data show a slow plateau of coverage provided by the frequency bands. Based on percentage of coverage, the high-frequency lemmas are more useful for language learners, and “It would thus make sense to sample more, and in more detail, at this end of the frequency continuum….“ and “it might also make more sense to cluster lemmas together in bigger bands toward the lower frequency end…” (Kremmel, 2016, p. 981). This supports the idea that natural language frequency distributions are Zipfian (Zipf, 1935). Therefore, equally-sized bands would illogically divide a language continuum, while smaller bands at the more frequent end and larger bands at the less frequent end is a more logical division. Following in line with this reasoning, the AVL was divided into the following five bands based on word frequency: Band 1 (1–200), Band 2 (201–400), Band 3 (401–800), Band 4 (801–1500), and Band 5 (1501–3000). A logarithmic division was chosen because it fulfilled Kremmel’s justification and fit a Zipfian curve, and there were 5 bands chosen to best fit the logarithmic division of the AVL. The bands were divided at word 1500 because there are a total number of 3012 words in the AVL, and 1500 fairly divided the remainder of the AVL.
**Data Analysis**

The first research question in this study is, “To what extent do AVL words vary in cumulative frequency bands across biology and U.S. history texts?” To answer this research question, the percentage of AVL tokens in each band was calculated, with the frequency bands cumulatively increasing until Band 5 represented words 1 to 3,000 of the AVL. Looking at the size of percent changes between bands informs which frequency bands of the AVL add the most tokens to the coverage and therefore contain the most beneficial words to learn. No inferential statistical analyses were performed on these data because the assumption of local independence was not met. Since the bands are cumulative, they are not independent from each other.

The second research question in this study is, “To what extent do AVL words vary in each separate frequency band with regards to level and discipline?” In other words, how do proficiency level and discipline effect the amount of AVL words in each separate frequency band? In this case, the assumption of local independence is met because the frequency bands are analyzed individually. This allows us to see if differences between the bands are significant. An ANOVA was run on these data with 5 dependent variables (the 5 frequency bands of the AVL) and factors of discipline (biology and U.S. history), level (elementary, junior high, high school, and university), and the relationship between proficiency level and discipline. Descriptive statistics for each band are also examined, namely means and standard deviations.
CHAPTER FOUR

Results

Cumulative Vocabulary Coverage

In order to answer the first research question, it is useful to look at the descriptive statistics for the cumulative bands. As mentioned earlier, the bands were divided into increasing increments (e.g. Band 1 included words 1–200 and Band 5 included words 1501–3000). For the cumulative bands, each band added all of the words in the previous bands (e.g. Band 1 includes words 1–200 and Band 5 includes words 1–3000). The focus of the cumulative descriptive statistics is on the bands, so the textbooks from all of the proficiency levels are included in these data. There were thirteen texts in the biology corpus and fifteen texts in the U.S. history corpus. There were 3,008,402 total tokens used composed of content words only, i.e. nouns, verbs, adjectives, and adverbs. Since some levels contained more tokens than others, the data were normalized and expressed as percentages.

The descriptive statistics are reported in Table 3. The mean describes the proportion of tokens in the corpus that are found in the respective bands of the AVL. The means increase as the bands increase with Band 5 containing 17.38% of tokens in the biology corpus found in the AVL and 13.60% of tokens in the U.S. history corpus found in the AVL. The largest percent delta of means for biology was from Band 1 to Band 2 with a delta of 2.77 as seen in Table 3. For history, the largest difference was from Band 2 to Band 3 with a percent delta of 2.66. The smallest difference in means for each corpus is from Band 4 to Band 5 with percent deltas of 0.46 and 0.40, respectively. This shows a slight plateau in the percentage of tokens found in the AVL and is indicative of which areas of the AVL are most valuable to learners. The largest standard deviation was found in Band 5 for both corpora with 5.1% for biology and 4.4% for
U.S. history. Looking at the means of the bands shown in Figure 1, it is clear that the biology texts contained a higher percentage of tokens from the AVL than the history texts did.

Table 3

Descriptive Statistics for Cumulative Bands by Percentage

<table>
<thead>
<tr>
<th>Band</th>
<th>Biology (N=13)</th>
<th>U.S. History (N=15)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>1 (1–200)</td>
<td>9.69</td>
<td>2.4</td>
</tr>
<tr>
<td>2 (1–400)</td>
<td>12.46</td>
<td>3.2</td>
</tr>
<tr>
<td>3 (1–800)</td>
<td>15.08</td>
<td>4.0</td>
</tr>
<tr>
<td>5 (1–3000)</td>
<td>17.38</td>
<td>5.1</td>
</tr>
</tbody>
</table>

Figure 1. Percentage means of cumulative bands.
Coverage per Band

In order to answer the second research question, it is useful to examine each band in the AVL by itself, not cumulatively. For each band, the descriptive statistics are presented followed by the results of an ANOVA performed on these data with the frequency bands as the dependent variables compared to factors of proficiency level (elementary school, junior high school, high school, and university), discipline (biology and U.S. history), and the relationship between proficiency level and discipline. The descriptive statistics are presented for biology and history texts separately and are indicative for which proficiency levels the bands of the AVL are most appropriate. The ANOVA results are reported from an analysis of the biology and history texts combined. As reported earlier, there were 13 biology texts used and 15 history texts used. There were 5 biology elementary texts and 6 history elementary texts, 3 junior high texts for both disciplines, 2 biology high school texts and 3 history high school texts, and 3 university texts for both disciplines. See Appendix B for a list of samples from each of the corpora with the AVL words highlighted in different colors for each band.

Band 1. As seen in Tables 4 and 5 and visually represented in Figure 3, biology texts had higher percentage means than the history texts did, meaning there was a higher percentage of AVL tokens compared to total tokens in the text. The biology means increased with the proficiency level up to 12.53% at the high school level. The university level slightly decreased to a mean of 11.06%. The history texts showed a similar pattern with the highest mean at 8.76% for high school and the university level mean decreasing to 7.33%. The biology levels were above the grand mean at the junior high, high school, and university levels, and only the high school level was above the grand mean for history. For both biology and history texts, the elementary level had the highest standard deviation at 1.2% and 2.55%, respectively.
There was a statistically significant difference for discipline \( F(1, 20) = 22.702, \ p < .001 \) and proficiency level \( F(3, 20) = 10.646, \ p < .001 \). There was not a statistically significant difference for the relationship between discipline and level. A Tukey post hoc test showed that for Band 1, the elementary level was significantly different (noted by \( p \)-values) with large effect sizes (noted with Cohen’s \( d \)) from the levels observed for junior high (\( p = .023, \ d = 4.35 \)), high school (\( p < .001, \ d = 7.23 \)), and university (\( p = 0.10, \ d = 4.89 \)). The junior high, high school, and university levels were not significantly different from each other.

Table 4

**Biology Band 1 Descriptive Statistics by Percentage**

<table>
<thead>
<tr>
<th>Proficiency</th>
<th>( N )</th>
<th>( M )</th>
<th>( SD )</th>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>5</td>
<td>7.20</td>
<td>1.20</td>
<td>6.15</td>
<td>8.25</td>
</tr>
<tr>
<td>Junior High</td>
<td>3</td>
<td>10.62</td>
<td>0.52</td>
<td>10.03</td>
<td>11.21</td>
</tr>
<tr>
<td>High School</td>
<td>2</td>
<td>12.53</td>
<td>0.47</td>
<td>11.88</td>
<td>13.18</td>
</tr>
<tr>
<td>University</td>
<td>3</td>
<td>11.06</td>
<td>0.94</td>
<td>10.00</td>
<td>12.12</td>
</tr>
</tbody>
</table>

Table 5

**History Band 1 Descriptive Statistics by Percentage**

<table>
<thead>
<tr>
<th>Proficiency</th>
<th>( N )</th>
<th>( M )</th>
<th>( SD )</th>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>6</td>
<td>5.78</td>
<td>2.55</td>
<td>3.74</td>
<td>7.82</td>
</tr>
<tr>
<td>Junior High</td>
<td>3</td>
<td>7.21</td>
<td>1.34</td>
<td>5.69</td>
<td>8.73</td>
</tr>
<tr>
<td>High School</td>
<td>3</td>
<td>9.36</td>
<td>0.53</td>
<td>8.76</td>
<td>9.96</td>
</tr>
<tr>
<td>University</td>
<td>3</td>
<td>7.33</td>
<td>1.20</td>
<td>5.97</td>
<td>8.69</td>
</tr>
</tbody>
</table>
Band 2. Band 2 is similar to Band 1 because the highest means for both biology and history are at the high school level, 3.96% and 3.32%, shown in Tables 6 and 7. The means slightly decrease at the university level for both disciplines. As shown in Figure 3, the high school and university levels are above the grand mean for both disciplines, and the biology junior high level is also above the grand mean. The largest standard deviation is at the elementary history level at 0.86%, and the lowest standard deviation is at the biology high school level at 0.10%.

There was a statistical significance for discipline \(F(1, 20) = 8.107, p = .010\), and level \(F(3, 20) = 23.947, p = <.001\). There were no statistically significant differences between the discipline and proficiency level. The Tukey post hoc test showed that the elementary and junior high levels were grouped together, and the high school and university levels were grouped together. The elementary level was significantly different from the levels observed for high school \((p = <.001, d = 11.38)\) and university \((p = <.001, d = 10.12)\). The junior high level was also significantly different from the levels observed for high school \((p = .003, d = 6.0)\) and university \((p = .007, d = 5.0)\).
Table 6

**Biology Band 2 Descriptive Statistics by Percentage**

<table>
<thead>
<tr>
<th>Proficiency</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>95% CI lower</th>
<th>95% CI upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>5</td>
<td>1.83</td>
<td>0.32</td>
<td>1.55</td>
<td>2.11</td>
</tr>
<tr>
<td>Junior High</td>
<td>3</td>
<td>2.83</td>
<td>0.32</td>
<td>2.46</td>
<td>3.20</td>
</tr>
<tr>
<td>High School</td>
<td>2</td>
<td>3.96</td>
<td>0.10</td>
<td>3.83</td>
<td>4.09</td>
</tr>
<tr>
<td>University</td>
<td>3</td>
<td>3.66</td>
<td>0.09</td>
<td>3.56</td>
<td>3.76</td>
</tr>
</tbody>
</table>

Table 7

**History Band 2 Descriptive Statistics by Percentage**

<table>
<thead>
<tr>
<th>Proficiency</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>95% CI lower</th>
<th>95% CI upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>6</td>
<td>1.68</td>
<td>0.86</td>
<td>0.99</td>
<td>2.37</td>
</tr>
<tr>
<td>Junior High</td>
<td>3</td>
<td>1.87</td>
<td>0.23</td>
<td>1.61</td>
<td>2.13</td>
</tr>
<tr>
<td>High School</td>
<td>3</td>
<td>3.32</td>
<td>0.21</td>
<td>3.09</td>
<td>3.55</td>
</tr>
<tr>
<td>University</td>
<td>3</td>
<td>3.18</td>
<td>0.40</td>
<td>2.73</td>
<td>3.63</td>
</tr>
</tbody>
</table>

![Estimated Marginal Means of Band 2 Ratio](image)

**Figure 3.** Biology and history band 2 descriptive statistics.

**Band 3.** The pattern of means is slightly different in Band 3 than it is in Bands 1 and 2, shown in Tables 8 and 9 and visually shown in Figure 4. For Band 3, the means increase with the highest means at the university level. The high school and university levels are both above the
grand mean. This is true for both biology and history. The largest standard deviation for biology is at the junior high level, 0.52%, and the largest standard deviation for history is at the elementary level, 0.61%.

The only statistical significance was found with regards to level \([F(3, 20) = 32.438, p < .001]\). Discipline and the relationship of discipline and level were not significant. The Tukey post hoc test showed that the elementary and junior high levels were grouped together, and the high school and university levels were not grouped with any other levels. This means that the elementary level was significantly different from the high school level \((p = .001, d = 8.22)\) and university level \((p < .001, d = 12.65)\). The junior high level was significantly different from the high school \((p = .020, d = 4.5)\) and university \((p < .001, d = 8)\) levels. Also, the high school and university levels were significantly different from each other \((p = .043, d = 3.5)\).

Table 8

*Biology Band 3 Descriptive Statistics by Percentage*

<table>
<thead>
<tr>
<th>Proficiency</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>95% CI lower</th>
<th>95% CI upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>5</td>
<td>1.85</td>
<td>0.32</td>
<td>1.57</td>
<td>2.13</td>
</tr>
<tr>
<td>Junior High</td>
<td>3</td>
<td>0.52</td>
<td>0.52</td>
<td>1.78</td>
<td>2.96</td>
</tr>
<tr>
<td>High School</td>
<td>2</td>
<td>3.44</td>
<td>0.19</td>
<td>3.18</td>
<td>3.70</td>
</tr>
<tr>
<td>University</td>
<td>3</td>
<td>0.24</td>
<td>0.24</td>
<td>3.72</td>
<td>4.26</td>
</tr>
</tbody>
</table>

Table 9

*History Band 3 Descriptive Statistics by Percentage*

<table>
<thead>
<tr>
<th>Proficiency</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>95% CI lower</th>
<th>95% CI upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>6</td>
<td>1.95</td>
<td>0.61</td>
<td>1.46</td>
<td>2.44</td>
</tr>
<tr>
<td>Junior High</td>
<td>3</td>
<td>2.27</td>
<td>0.09</td>
<td>2.17</td>
<td>2.37</td>
</tr>
<tr>
<td>High School</td>
<td>3</td>
<td>2.95</td>
<td>0.28</td>
<td>2.63</td>
<td>3.27</td>
</tr>
<tr>
<td>University</td>
<td>3</td>
<td>0.46</td>
<td>0.46</td>
<td>3.25</td>
<td>4.29</td>
</tr>
</tbody>
</table>
Figure 4. Biology and history band 3 descriptive statistics.

**Band 4.** The pattern of percentage means for Band 4 is similar to Band 3 because the means increase with the highest means found at the university level. The highest biology mean is 2.55% and the highest university mean is 2.03%, shown in Tables 10 and 11 and visually represented in Figure 5. Biology and history high school and university levels are above the grand mean in addition to the biology junior high level. The largest standard deviation for biology is 0.48% at the high school level, and the largest for history is 0.30% at the high school level.

Similar to Bands 1 through 3, there was a statistical difference for the factors of discipline \( F(1, 20) = 21.261, p < .001 \) and level \( F(3, 20) = 37.098, p < .001 \) with no statistical significance for the relationship between discipline and level. The Tukey post hoc test showed a similar pattern to Band 3: Elementary and junior high levels were grouped together, and the high school and university levels were given their own groups because they were different from all of the other levels. The significant differences between levels were as follows: elementary and high school \( p < .001, d = 9.26 \), elementary and university \( p < .001, d = 13.97 \), junior high and high
school ($p = 0.008, d = 5.80$), and junior high and university ($p < 0.001, d = 9.70$), high school and university ($p = 0.056, d = 3.31$).

Table 10

**Biology Band 4 Descriptive Statistics by Percentage**

<table>
<thead>
<tr>
<th>Proficiency</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>5</td>
<td>1.05</td>
<td>0.29</td>
<td>0.79 - 1.31</td>
</tr>
<tr>
<td>Junior High</td>
<td>3</td>
<td>1.57</td>
<td>0.15</td>
<td>1.41 - 1.73</td>
</tr>
<tr>
<td>High School</td>
<td>2</td>
<td>2.14</td>
<td>0.48</td>
<td>1.48 - 2.80</td>
</tr>
<tr>
<td>University</td>
<td>3</td>
<td>2.55</td>
<td>0.32</td>
<td>2.19 - 2.91</td>
</tr>
</tbody>
</table>

Table 11

**History Band 4 Descriptive Statistics by Percentage**

<table>
<thead>
<tr>
<th>Proficiency</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>6</td>
<td>0.77</td>
<td>0.32</td>
<td>0.52 - 1.02</td>
</tr>
<tr>
<td>Junior High</td>
<td>3</td>
<td>0.85</td>
<td>0.05</td>
<td>0.79 - 0.91</td>
</tr>
<tr>
<td>High School</td>
<td>3</td>
<td>1.62</td>
<td>0.30</td>
<td>1.29 - 1.95</td>
</tr>
<tr>
<td>University</td>
<td>3</td>
<td>2.03</td>
<td>0.19</td>
<td>1.81 - 2.25</td>
</tr>
</tbody>
</table>

Figure 5. Biology and history band 4 descriptive statistics.
**Band 5.** As shown in Tables 12 and 13, the biology texts have higher percentage means than the history texts, except at the elementary level, where they are the same, and the university level, where history is slightly higher than biology. All of the levels increase in percentage means with the highest means at the university levels. The highest biology mean is 1.02%, and the highest history mean is 1.04%. This can be seen in Figure 5. Only the high school and university levels are above the grand mean. The largest standard deviation for biology is at the high school level, 0.15%, and the largest for history is at the university level, 0.18%.

There were statistically significant differences for discipline and level, but not for the relationship between discipline and level. The reported ANOVA results were $F(1, 20) = 4.441, p = .048$ for discipline and $F(3, 20) = 73.868, p < .001$ for level. The post hoc test showed similar groupings as Bands 3 and 4. Elementary and junior high levels were similar to each other, and the high school and university levels were different from all the other levels. The elementary level was significantly different from the levels observed for high school ($p < .001, d = 9.0$) and university ($p < .001, d = 19.39$). The junior high level was significantly different from the levels observed for high school ($p = 0.002, d = 6.03$) and university ($p < .001, d = 16.95$), and the high school level was significantly different from the university level ($p < .001, d = 8.04$).

Table 12

**Biology Band 5 Descriptive Statistics by Percentage**

<table>
<thead>
<tr>
<th>Proficiency</th>
<th>$N$</th>
<th>$M$</th>
<th>$SD$</th>
<th>lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>5</td>
<td>0.21</td>
<td>0.11</td>
<td>0.11</td>
<td>0.30</td>
</tr>
<tr>
<td>Junior High</td>
<td>3</td>
<td>0.38</td>
<td>0.02</td>
<td>0.36</td>
<td>0.40</td>
</tr>
<tr>
<td>High School</td>
<td>2</td>
<td>0.15</td>
<td>0.15</td>
<td>0.54</td>
<td>0.95</td>
</tr>
<tr>
<td>University</td>
<td>3</td>
<td>1.02</td>
<td>0.10</td>
<td>0.91</td>
<td>1.14</td>
</tr>
</tbody>
</table>
Table 13

*History Band 5 Descriptive Statistics by Percentage*

<table>
<thead>
<tr>
<th>Proficiency</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>95% CI lower</th>
<th>95% CI Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>6</td>
<td>0.21</td>
<td>0.14</td>
<td>0.09</td>
<td>0.32</td>
</tr>
<tr>
<td>Junior High</td>
<td>3</td>
<td>0.22</td>
<td>0.04</td>
<td>0.18</td>
<td>0.26</td>
</tr>
<tr>
<td>High School</td>
<td>3</td>
<td>0.51</td>
<td>0.05</td>
<td>0.45</td>
<td>0.57</td>
</tr>
<tr>
<td>University</td>
<td>3</td>
<td>1.04</td>
<td>0.18</td>
<td>0.83</td>
<td>1.24</td>
</tr>
</tbody>
</table>

*Figure 6. Biology and history band 4 descriptive statistics.*

**Summary**

For the cumulative coverage of AVL tokens, the percentage of tokens in the AVL to the rest of the tokens of the texts was calculated for cumulative frequency bands. Naturally, the percentage of tokens increased with each proficiency level. The largest percent change from one band to another was from Band 1 to Band 2 for biology and Band 2 to Band 3 for history. The smallest percent change was from Band 4 to Band 5 for both biology and history. Also of note, the biology texts consistently had a higher percentage of AVL tokens than the history texts did.

Looking at each band separately, there are more AVL tokens found in biology texts than in history texts. This was seen for every band at every proficiency level except for a few
exceptions where there were slightly more or the same number of tokens in the history texts as the biology texts. In general, the high school and university levels for biology and history and the junior high biology level were above the grand mean while the other levels were below the grand mean. The results of the ANOVA show that the factor of discipline was statistically significant for all bands except Band 3, and the factor of level was significant for all bands. There was no statistical significance for the relationship between discipline and level. A Tukey post hoc test was performed for the factor of level but not for the factor of discipline because there are only two disciplines, which is not enough variables for a post hoc test. For all bands except Band 1, the elementary and junior high levels were not significantly different from each other, but they were significantly different from all the other levels. The high school and university levels were not different from each other in Bands 1 and 2, but they were significantly different from each other for Bands 3 through 5. The effect sizes of all these differences were large.
CHAPTER FIVE

Discussion

The first research question of this study addressed the variance of AVL words in cumulative frequency bands for biology and U.S. history texts. For this analysis, all four proficiency levels were examined together. As reported previously, the biology texts consistently had higher percentages of AVL tokens than the history texts. 17.38% of tokens in the total biology texts are part of the AVL while 13.60% of tokens in the total history texts are part of the AVL. These data suggest that the AVL may be slightly more useful for some disciplines than other disciplines, despite the AVL’s goal to provide core academic vocabulary. It is also possible that some disciplines include more core vocabulary than other disciplines do, and some disciplines include more technical vocabulary than others do. In this study, it was clear that biology texts contained more tokens that were part of the AVL than history texts did.

One possible concern about how accurately these data reflect AVL coverage is the fact that tokens (total number of word occurrences) were analyzed instead of types (only one occurrence counted per word). At first glance, it could hypothetically be possible that a number of words specific to the biology domain and included on the AVL were highly frequent in the biology texts and therefore the cause of the higher mean percentage. This would challenge the belief of core academic vocabulary and the AVL’s claim to provide this core vocabulary. There are some words on the AVL that are naturally more frequently found in biology texts than in history texts. These are words such as *biological* (frequency of 482), *biomedical* (frequency of 1,767) and *symbiotic* (frequency of 2,324). There are also some words that are naturally more frequently found in history texts than in biology texts, such as *history* (frequency of 26), *society* (frequency of 39), and *European* (frequency of 224). However, the constraints with which the AVL was created successfully exclude too technical of words from being included in the AVL.
These constraints, described in Chapter 2, include requiring words to occur in at least seven of the nine different disciplines with at least 20% of the expected frequency and to be evenly spread across the corpus with a dispersion of at least 0.80 (1.0 meaning perfectly distributed and 0.01 meaning the word only occurs in a small portion of the corpus.) For example, organism is not included in the AVL because it did not comply with these constraints. Biological, biomedical, and symbiotic, history, society, and European did comply with these constraints. For example, symbiotic had a dispersion of 0.86 and was found in all nine of the subdisciplines in the corpus used to create the AVL. Naturally, it occurred more frequently in the science discipline (146 times) than it did in the history discipline (48 times). It is possible that a word like symbiosis occurs in different disciplines due to its different meanings. For example, the discipline of history may use a more metaphorical meaning of symbiosis. The word history had a dispersion of 0.93, and it occurred 14,348 times in history texts and 6,224 times in science texts. However, the strict constraints with which the AVL was created prevented technical, domain-specific words from being included, so it is unlikely that frequent technical words were the cause of the higher percentage means in the biology corpus than the history corpus.

Another area of discussion the first research question presents is the percent delta between bands. Although there could be differences depending on the proficiency levels, this analysis examined the proficiency levels together in order to provide a larger picture of the data. The bands of the AVL were examined cumulatively, so Band 1 (words 1–200) had a lower percentage mean than Band 5 (words 1–3000) did, as expected. The percent delta between bands can provide some insight about the usefulness of the AVL words in each frequency band. The larger the percentage delta between bands, the more AVL tokens were included and therefore the more useful the AVL band is. The opposite is true for the smaller the percentage delta between bands. For biology, the largest percentage delta is from Band 1 to Band 2, suggesting that words
201 to 400, the words added to Band 1 to create Band 2, are frequent and useful to learn. The percent change is similar from Band 2 to Band 3, suggesting words 401 to 800 are also useful. The percent delta is not as large from Band 3 to Band 4 and from Band 4 to Band 5, suggesting that words 801 to 3000 are not as frequent and it would not be as productive for students to focus on learning the words in those bands. These data are similar to the data for history with one small change. The percent delta was largest from Band 2 to Band 3, and the percent delta from Band 1 to Band 2 was slightly lower. Of course, one would expect the higher frequent words of the AVL to be more valuable to learn than the less frequent words are—this is a fundamental reason of the creation of word lists. However, examining the percent deltas between bands more closely identifies at which point the words of the AVL taper off in frequency.

It is of note that the decision to use logarithmic frequency bands to model a Zipfian natural language frequency distribution supports Kremmel’s (2016) ideas about frequency band divisions. Even though the lower bands represented a fewer number of words and the higher bands represented a larger number of words, the lower bands added more vocabulary coverage than the higher bands did.

The second research question addressed the variances of AVL words in each separate frequency band with regards to discipline (biology and U.S. history) and level (elementary, junior high, high school, and university). For Bands 1 and 2, the high school proficiency level had the highest percentage of AVL tokens for both biology and history. This suggests that knowledge of the words in Bands 1 and 2 is especially valuable for learners at the high school level. For Bands 3 through 5, the university proficiency level had the highest percentage of AVL tokens for both biology and history, showing that the AVL words in the higher frequency bands are found more frequently at the university proficiency level than at lower levels. Learning the AVL words in these bands is more useful for learners at the university levels and not as useful
for learners at the elementary and junior high levels. Logically, this supports the notion that university texts are more difficult to read partly due to vocabulary. Overall, the percentage of AVL tokens found in the texts suggests that learning AVL words is more useful for learners preparing for or at high school and university proficiency levels and is not as useful for learners at elementary or junior high levels.

Regarding the results of the ANOVA, there was no statistical significance for any band comparing the relationship between discipline and level. For the factor of discipline, there was a statistical significance at every band except for Band 3. This suggests that in general, vocabulary of biology and history texts are significantly different from each other, no matter the frequency level of the vocabulary being compared. For the factor of level, there was a statistical significance for every band, suggesting that every frequency band was able to discriminate between at least two different proficiency levels. For Band 1, the elementary level was significantly different from all the other levels. For all the other bands of the AVL, elementary and junior high texts were similar to each other. Perhaps for the first 200 words of the AVL, elementary texts had a minimal number of tokens while the junior high texts contained some AVL tokens, but for the rest of the AVL words, both elementary and junior high texts did not contain many tokens. Because the AVL is a word list of academic words, it is logical that the words on the AVL would not be frequent at low proficiency levels. For Band 2, high school and university texts are not significantly different from each other, but they are for Bands 3 through 5. These results show that the AVL vocabulary of high school and university texts are more similar to each other the higher the frequency of the AVL words. The less frequent the AVL words are, the less similar high school and university texts are to each other. The effect sizes of all the statistically significant differences were large, showing that based on the size of the samples, the effect of the significance is large.
In traditional readability tests, the length of words is a key indicator of vocabulary difficulty and therefore part of the overall text difficulty. Today, there are more precise measurements of vocabulary difficulty that involve word frequency. A number of text leveling systems use word frequencies in their systems, and many use the AWL as part of their processes. As of yet, the AVL has not been used in leveling processes. Because the AVL improves upon the AWL and based on the statistically significant differences found comparing the AVL to the corpora in this study, it is worth further investigating the AVL’s utility to be used as a text leveling tool. Using the AVL as a leveling tool could also help students at various proficiency levels know which words to focus on, based on the expected AVL coverage at their particular proficiency level. Also, it would be meaningful to perform similar studies with the AWL in the future. Similar studies performed with the AWL and the AVL with regards to text leveling may shed more light on the differences between these two lists.

Limitations

As mentioned earlier, using tokens instead of types could be considered a limitation because a high number of tokens for a low number of types could potentially skew the representativeness of the data. However, if learners know a type that is frequently repeated, their overall comprehension is positively affected because they know a larger percentage of the text. It’s also possible that some disciplines may tend to have more types than other disciplines do. For example, the discipline of history may include common types from other disciplines such as philosophy and religion, while the discipline of biology is more constricted and would probably not include as many types from additional disciplines. Performing the same analyses using types instead of tokens to see the differences of AVL coverage could be enlightening.

Additionally, a greater number of texts per corpora would have been useful in order to eliminate outside variables, such as author’s voice and publisher’s preferences. It is possible that
these variables may have a large influence with fewer text samples. A minimum of three texts were chosen per corpora, except for the biology high school corpus. Only two texts were used, and the study would have benefited from a greater number of texts used for each of the corpora.

Another limitation mentioned earlier and inherent to this study is the use of frequency bands. Dividing a word list into frequency bands subjectively separates words into fixed categories. For example, there may be a boundary at word 200, but words 199 and 201 may not be that different from each other. However, previous research has shown that vocabulary learning follows a frequency order of clusters (bands), so using frequency bands in assessment makes sense (Kremmel, 2016).
CHAPTER SIX

Conclusion

In order for language students to effectively learn, they need comprehensible reading texts, partly decided by the vocabulary of the texts. There are many ways to level texts, including using word lists such as the GSL and the AWL in the process. This study explored the usefulness of the AVL to be used in the text leveling process. The AVL was divided into five frequency bands and compared against the vocabulary of biology and U.S. history texts at four proficiency levels. Overall, there were more AVL tokens found in biology texts than in U.S. history texts, suggesting that the AVL may be better suited for some disciplines over other disciplines. Also, in analyzing the percent deltas of the cumulative bands, the most beneficial bands of the AVL to learn are Bands 1 through 3. After Band 3, the percent of AVL tokens found in the texts plateaus.

Looking at each band separately, the high school proficiency level had the most AVL coverage for Bands 1 and 2, and the university proficiency level had the most coverage for Bands 3 through 5, suggesting that the AVL is better suited for students preparing for or at the high school and university levels. For most of the bands, the elementary and junior high texts were statistically grouped together, and the high school and university levels were grouped separately. These results suggest that using the AVL in the text leveling process may be more appropriate in certain disciplines and at different proficiency levels. If the level of the text is too low, there are too few AVL words in the texts to accurately assign various text levels.

Although word lists are commonly used in text leveling processes, there is little research about doing so. Future research would wisely investigate the validity of using word lists in text leveling processes. Additionally, future research would benefit from investigating the differences between analyzing corpora with types in addition to tokens, including more sample texts, and considering additional disciplines.
References


discourse: The quantitative and qualitative academic papers in the field of the TEFL.

unbiased estimates of text complexity. In Proceedings of the Workshop on Natural
Language Processing for Improving Textual Accessibility (p. 49–58).

presented at the Fourth North American Conference on Adolescent/Adult Literacy,
Washington, D.C.


of academic word knowledge for the academic achievement of diverse middle school

Vajjala, S., & Meurers, D. (2014). Assessing the relative reading level of sentence pairs for text
simplification. In Proceedings of the 14th Conference of the European Chapter of the
Association for Computational Linguistics (p. 288–297).

using insights from second language acquisition. In Proceedings of the seventh workshop
on building educational applications using NLP (p. 163–173). Association for
Computational Linguistics.


Appendix A

References of Biology Textbooks Used in the Corpora

Selections used from books of the same series were combined into one text document for analysis.

Elementary


**Junior High**


**High School**


**University**


References of U.S. History Textbooks Used in the Corpora

**Elementary**


**Junior High**


**High School**


University


Appendix B

Extracts of Biology Texts

AVL words are highlighted in the extracts as follows: Band 1, blue; Band 2, green; Band 3, orange; Band 4, pink; Band 5, red.

**Elementary**


What do you think of when you hear the word *bacteria*? Many people think of germs. But not all bacteria are harmful. Many of these single-celled organisms are helpful. Some enrich the soil by breaking down dead plants and animals. Others help animals digest food. Still others help make food. Cheese and yogurt are foods that form when certain bacteria mix with milk.

Bacterial cells are different from plant and animal cells. Like plant cells, bacterial cells are surrounded by cell walls. But bacteria do not have a nucleus or membrane-bound organelles. Instead, their chromosomes and other materials float in the cytoplasm.

Another kind of single-celled organism makes up most of the group called protists. A protist...is a simple organism, usually a single cell, with a nucleus and organelles. Some protists have cell walls and chloroplasts. These protists are plantlike. Other protists have no cell walls or chloroplasts and are animal-like.

The diatoms...shown at the beginning of this lesson are plantlike—they have chloroplasts and make their own food. They are part of a group of protists called *algae*.... Algae produce a lot of Earth’s oxygen. They also produce a lot of food for ocean life. Animal-like protists are called *protozoa*. They get food by “eating” other small organisms, such as algae and bacteria. (p. 58)
Around the same time that Pasteur was conducting his experiments, a German physician named Rudolf Virchow was studying cells. Like Pasteur, he did not support the idea of spontaneous generation. In 1855, Virchow presented a theory that cells are produced only by other cells. This last piece of information on cells helped complete what is now known as the cell theory. The cell theory states:

- All living things are made up of cells.
- Cells are the basic units of structure and function in living things.
- New cells are produced from existing cells.

You pull a box of cereal from the kitchen cabinet and head to the table. As your stomach rumbles, you pour some cereal into a bowl. The cereal you’re about to eat was processed in a factory. In a factory, many different people and machines work together to manufacture a product. In a cereal factory, ingredients such as grains, sugar, and flour are taken into the factory. They are mixed together and processed to produce cereal. The finished cereal is then placed into packages that are sent out to stores.

In many ways, living cells are like factories that produce goods. They take in raw materials, use them to build products such as proteins, package the products, and transport them to different parts of the cell or to other cells. The different jobs are performed by structures within the cell called organelles. The cells of animals and plants share most of the same kinds of organelles and other cell parts. (p. 38–39)
Hooke observed small, box-shaped structures, such as those shown in Figure 2. He called them cellulae (the Latin word meaning *small rooms*) because the boxlike cells of cork reminded him of the cells in which monks live at a monastery. It is from Hooke’s work that we have the term *cell*. A cell is the basic structural and functional unit of all living organisms.

During the late 1600s, Dutch scientist Anton Van Leeuwenhoek…designed his own microscope after he was inspired by a book written by Hooke. To his surprise, he saw living organism in pond water, milk, and various other substances. The work of these scientists and others led to new branches of science and many new and exciting discoveries.

Scientists continued observing the living microscopic world using glass lenses. In 1838, German scientist Matthias Schleiden carefully studied plant tissues and concluded that all plants are composed of cells. A year later, another German scientist, Theodor Schwann, reported that animal tissues also consisted of individual cells. Prussian physician Rudolph Virchow proposed in 1855 that all cells are produced from the division of existing cells. The observations and conclusions of these scientists and others are summarized as the cell theory. The cell theory is one of the fundamental ideas of modern biology and includes the following three principles:

1. All living organisms are composed of one or more cells.
2. Cells are the basic unit of structure and organization of all living organisms.
3. Cells arise only from previously existing cells, with cells passing copies of their genetic material on to their daughter cells. (p, 182–183)
The word biology comes from the Greek roots "bio" meaning "life" and "logy" meaning "the study of" (see Appendix I for more word roots). But what is life? If you look up "life" in a dictionary, you will find definitions such as "the quality that distinguishes a vital and functioning being from a dead body," but you won't discover what that "quality" is. Life is intangible and defies simple definition, even by biologists. However, most agree that living things, or organisms, all share certain characteristics that, taken together, define life:

- Organisms acquire and use materials and energy.
- Organisms actively maintain organized complexity.
- Organisms sense and respond to stimuli.
- Organisms grow.
- Organisms reproduce.
- Organisms, collectively, evolve.

Nonliving objects may possess some of these attributes. Crystals can grow, and a desk lamp acquires energy from electricity and converts it to heat and light, but only living things can do them all.

The cell is the basic unit of life. A plasma membrane separates each cell from its surroundings, enclosing a huge variety of structures and chemicals in a fluid environment. The plasma membranes of many types of cells, including those of microorganisms and plants, are enclosed in a protective cell wall (FIG. 1-1).
In England, many Puritans had been jailed because of their religious beliefs. In 1629, a group of wealthy Puritans in England formed the Massachusetts Bay Company. King Charles I of England gave the company a charter to settle in North America. The charter was like a permit allowing the Puritans to settle areas claimed by England.

The Puritans brought more than 700 people to North America. They also came with different kinds of equipment, a herd of cows, and about 60 horses. The Puritans founded Massachusetts Bay Colony along the Charles River. The word Massachusetts means “at or near the great hill” in Algonkian. They named their first settlement Boston.

John Winthrop was chosen to be the governor of the colony. He wanted to create “a city upon a hill” that would be a model for how God wanted people to live. In the Puritan community, each “free man,” as a male colonist was called, signed a covenant. A covenant is a special promise between a person and God. In his covenant, a Puritan promised that his family would live by certain rules of the Puritan Church.

Everyone worked hard to build the community. Men, women, and children worked in the fields growing crops and raising cattle, pigs, and sheep. They also helped each other.

In the center of each village was the village common. A common is a grassy area that is shared by everyone in the community. Near the common was the meeting house, which also served as the church. The Puritans also built schools. Free education was unheard of in Europe at the time, but Puritan schools were free. (p. 181)
In 1628, a small group of settlers arrived in North America. They had a charter, or official paper, from the king of England. It gave the settlers permission to start a colony in New England. There, they built a village called Salem.

Like the Pilgrims, these settlers came to New England to practice their religious beliefs. They also came to start farms and businesses. Unlike the Pilgrims, the new settlers did not want to break away from the Church of England. They wanted to change some religious practices to make the church more “pure.” For this reason, they were called Puritans.

In 1630, John Winthrop led a second group of Puritans to settle the Massachusetts Bay Colony. Winthrop hoped their settlement would be an example of Christian living. In a sermon, he said, “…We shall be as a city upon a hill. The eyes of all people are upon us….?” Winthrop’s group chose to build their “city upon a hill” south of Salem, near the mouth of the Charles River. They named their settlement Boston, after a town in England. Most early settlements in New England were built along the Atlantic coast to make it easier for colonists to get supplies from English trading ships.

In 1630, John Winthrop was elected governor of the Massachusetts Bay Colony. He and the other Puritan leaders kept a strict control over life in the colony. They did not welcome people whose beliefs were different from their own. They thought that dissent, or disagreement, might hurt their society. (p. 179)
Religious tensions in England remained high after the Protestant Reformation. A Protestant group called the Puritans wanted to purify, or reform, the Anglican Church. The Puritans thought that bishops and priests had too much power over church members.

The most extreme English Protestants wanted to separate from the Church of England. These Separatists formed their own churches and cut all ties with the Church of England. In response, Anglican leaders began to punish Separatists.

The Pilgrims were one Separatist group that left England in the early 1600s to escape persecution. The Pilgrims moved to the Netherlands in 1608. The Pilgrims were immigrants—people who have left the country of their birth to live in another country.

The Pilgrims were glad to be able to practice their religion freely. They were not happy, however, that their children were learning the Dutch language and culture. The Pilgrims feared that their children would forget their English traditions. The Pilgrims decided to leave Europe altogether. They formed a joint-stock company with some merchants and then received permission from England to settle in Virginia.

On September 16, 1620, a ship called the Mayflower left England with more than 100 men, women, and children aboard. Not all of these colonists were Pilgrims. However, Pilgrim leaders such as William Bradford sailed with the group. (p. 78–79)
Puritan magistrate John Winthrop, in a speech to his fellow Puritans while their ship, the Arabella, was still making its way across the Atlantic, invoked a remarkable image. “We shall be as a city upon a hill,” he said. “The eyes of all people are upon us.” This may have been the first iteration of what was to become the idea of America.

What Winthrop meant was that he and his fellow Puritans were going to show the world what God could write upon that tabula rasa. His city upon a hill would be nothing less than a vision of the world as God had intended it to be—the world recast according to holy principles. In a later speech, for example, Winthrop delved into the nature of liberty, explaining the difference between natural and civil liberty. The difference would be crucial to that city upon a hill. Given natural liberty, men were free to do precisely what they pleased, Winthrop argued, and the sad state of the world reflected the choices that most of them made. In the Puritan commonwealth, by contrast, men would enjoy civil liberty, where one was free to do only that which was good, just, and honest.

Winthrop’s “city upon a hill” was a little like Plato’s ideal republic. It was to be as near to perfection as a flawed and sinful world would allow. It would include many of the attributes that political philosophers had imagined of the Good Society:

• Reasonable order, created by the people themselves.
• Reasonable prosperity for everyone.
• A strong, vibrant culture, prizing science and literature.
• Peaceful toward others, yet strong and well respected. (p. 24)