New England Classical Journal

Volume 49
Issue 1 *Special Issue In Honor of Jacqui Carlon*

Pages 80-101

8-12-2022

Latin Vocabulary Knowledge and the Readability of Latin Texts: A Preliminary Study

John Gruber-Miller
Cornell College, jgruber-miller@cornellcollege.edu

Bret Mulligan
Haverford College, bmulliga@haverford.edu

Latin Vocabulary Knowledge and the Readability of Latin Texts: A Preliminary Study

John Gruber-Miller
Bret Mulligan

Abstract: Studies have found a strong correlation between vocabulary knowledge and L2 reading comprehension. This preliminary study of the readability of Latin texts considers how common measures of lexical complexity (word length, word frequency, lexical sophistication, lexical density, and lexical variation) can inform instructors about what texts have the least (and most) lexical complexity. By defining several key measurements of Latin lexical complexity, we establish a provisional account of the lexical difficulty of some familiar Latin texts that are frequently taught in elementary, intermediate, and advanced levels, and propose *LexR*, a single, informative, integrated score that provides a sense of the comparative lexical complexity of Latin texts.

Keywords: lexical complexity, readability, Latin, word length, word frequency, lexical sophistication, lexical density, lexical variation

1. Introduction

For students of Latin, acquiring a vocabulary large enough to read texts assigned in class poses a significant — and often intimidating — challenge. Imagine these all-too-familiar scenarios. In the first scenario, students, who may have been expected to learn 450–1000 words in their first year(s) of studying Latin, are tasked to read with fluency a historical Latin text. These texts contain many words they cannot recall in the new context and many hundreds of additional words they have never before encountered.[1] In the second, instructors use Latin novellas with "sheltered" vocabulary, in which the number of vocabulary words is limited, to attempt to align class readings with the lexical knowledge of their students. But they then find themselves at odds with their institution's curriculum, which prioritizes historical Latin texts. They may also regret that they are not offering their students the experience of reading the authors who inspired them to learn Latin. Intrepid Latin instructors who recognize these difficulties still confront a double bind: we remain uncertain about how unknown vocabulary affects our students' ability to read and we lack fundamental tools that would help us assess how well texts, whether historical or contemporary, align with their students' level of reading proficiency.[2]

A confluence of developments in different disciplines over the past decade makes this an opportune moment to assay the role that vocabulary knowledge exerts on the readability of Latin texts. Readability studies have matured and are now beginning to be applied not just to corpora of English texts for first and second language learners, but also to foreign languages (Xia, Kochmar, and Briscoe, 2016). Computer analysis of texts has progressed so that it is no longer necessary to analyze texts by hand. Curated datasets — such as those of *The Bridge* (Mulligan), the *Ancient Greek and Latin Dependency Treebank* (AGLDT),

1 The typical number of core vocabulary words in an introduction sequence was calculated from a survey of the required vocabulary in 18 elementary Latin textbooks, which ranged from 433 words in *De Romanis* to 1468 words in *Latin for Americans*; the mean of the sample was 947 words.

2 Digital projects to help instructors discover readable texts are on the horizon: e.g., *Bridge/Oracle* identifies texts and passages that contain the highest percentage of user-defined vocabulary (Mulligan, bridge.haverford.edu/oracle); *(Meletē)ToPān* (Koentges, github.com/ThomasK81/ToPan) facilitates topic modeling of ancient texts, allowing instructors to identify passages that are thematically (and so likely lexically) similar.

 https://doi.org/10.52284/NECJ.49.1.article.gruber-millerandmulligan

and *Opera Latina* by the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) — can expedite the statistical analysis of Latin texts.[3] Meanwhile, recent trends in Latin pedagogy have refocused attention on the role of vocabulary acquisition in language learning. Instructors who employ an Active Latin approach have begun to publish Latin novellas, fictional texts that intentionally limit the amount of new vocabulary in a text without necessarily limiting grammatical structures (Piazza 2017, Venditti 2021; cf. Ramsby, this volume). Moreover, many Latin teachers are developing "embedded" texts for classical authors. These simplified versions of Latin authors (or textbook readings) embed the main ideas and many of the words and structures of the original, but substitute more frequent and familiar Latin vocabulary for uncommon words and simplify longer, more complex sentences.[4] These developments reveal an interest in assessing the readability of Latin texts and a newfound capacity to do so.

Yet measuring the readability of a text is a complex task. Since reading is an interactive process that combines author-driven, reader-driven, and text-driven factors (cf. Shrum and Glisan 172–88), textual features alone are insufficient to measure fully the readability of a text. According to Turk and Kirkman, the writer affects readability by "careful selection of the material, by organization, signposting and variation of emphasis" (1989). Likewise, certain reader-driven factors cannot be measured easily in text readability scores, such as motivation, working memory, and general linguistic proficiency, along with bottom-up (e.g., recognizing words and endings) and top-down processes (e.g., background knowledge, topic familiarity, and cultural knowledge). These significant factors in the readability of a text are highly subjective or exceedingly difficult to assay. Even limiting analysis to a full account of text-driven factors would require a formidable set of data. DuBay (2004), for example, identifies four textual features critical for readability: (1) *content* (propositions, organization, coherence); (2) *style* (semantic and syntactical features); (3) *design* (layout, typography, and illustrations); (4) and *discourse structure* (genre, chapters, headings, and navigation). The factors that affect the reading task, therefore, are complex and can be difficult to quantify. Yet, despite the complexity of fully describing the reading task, two text-driven factors have proven to be good predictors of text readability in English: lexical content and syntactic complexity (DuBay 2004). This preliminary study of readability in Latin texts explores the first and most important of these text-driven factors: lexical content. We consider how five common measures of lexical complexity (word length, word frequency, lexical sophistication, lexical density, and lexical variation) can inform instructors about what texts have the least (and most) lexical complexity. By defining several key measurements of Latin lexical complexity, we will establish a provisional account of the lexical difficulty of some familiar Latin texts that are frequently taught in elementary, intermediate, and advanced levels, and establish whether statistical measures of lexical complexity match the perceived difficulty of texts at different levels.

2. Vocabulary Knowledge and Reading Comprehension

Vocabulary knowledge has been recognized as a crucial component of language proficiency and as the best predictor of reading comprehension. How often has a student expressed frustration that they had to look up dozens of words in order to begin to understand a brief Latin text? Frequent consultation of the dictionary or on-line parsers is sure to demoralize even the most assiduous student. According to Chall, vocabulary difficulty explains as much as 80% of the total variability of readability scores for traditional print texts, with sentence

---

3 *The Bridge* (bridge.haverford.edu) is a free web tool that allows users to generate customized vocabulary lists from a human-curated database of Greek and Latin lists, textbooks, and texts; the AGLDT is part of the Perseus Digital library (https://perseusdl.github.io/treebank_data/); LASLA's *Opera Latina* enables the analysis of lexical, morphological, and stylometric data (web.philo.ulg.ac.be/lasla/opera-latina/).

4 For example, *Project Arkhaia* has produced sets of "AP Tiered Readings" that provide passages of Caesar's *Bellum Gallicum* and Vergil's *Aeneid* rewritten at several levels of complexity (http://lapis.practomime.com/index.php/operation-caesar-reading-list).

structure possessing a small additional amount of predictive power (Chall 1958, 156–58). Studies have found a strong correlation between vocabulary knowledge and reading comprehension in one's first language (Grabe and Stoller 2011) and even more so for reading in a second language (Laufer 1992, 1997; Qian 1999, 2002). In fact, vocabulary knowledge emerges as the strongest predictor of reading ability. A recent study by Staehr, which found that vocabulary knowledge predicted up to 72% of variance in reading comprehension, supports Chall's findings (Staehr 2008).[5]

To fully comprehend a text, the reader must know a high percentage of the words in that text. Indeed, the current consensus holds that full comprehension requires that a reader must know between 95–98% of the words in that text.[6] For example, Hu and Nation (2000) divided readers into four groups, in which students receiving a text composed from a list of the 2000 most frequent words in English. The text for each group was differentially modified by replacing none, 5%, 10%, or 20% of the words in the passage with nonsense words. No students could make sense of the passage comprising 80% known vocabulary, and only a few the passage in which one word out of ten words was unknown (90%). At 95% coverage (one word out of twenty unknown), fewer than half (35–41%) of the readers could adequately comprehend the passage. As a result, they conclude that 98% coverage (one word out of 50 unknown) was necessary to insure comprehension.

There has been some debate about the number of words needed to comprehend a typical text in English. Most estimates range from 2000–3000 word families (or 5000 individual words) to 8000–9000 word families.[7] The need to acquire such a high level of vocabulary knowledge is daunting for a second language learner.[8] Recent studies have suggested that a solid knowledge of high frequency vocabulary might be a reasonable goal for intermediate language learners. In examining English reading comprehension by Dutch high school students, Staehr (2008) found that knowing the most frequent 2000 word families was an important threshold and led to above average reading comprehension scores. Nation (2006) found that the 2000 highest-frequency word families provides 85% lexical coverage in English. Since the vocabulary of Latin is more limited than that of English, it is reasonable to assume that a shorter Latin list would provide analogous lexical coverage and preliminary work supports this postulate: e.g., Diederich's finding that 1471 words provided 83.6% coverage of a wide selection of Latin texts.[9]

3. Measures of Lexical Complexity

Lexical complexity or lexical richness is a multidimensional construct that must be measured in different ways. Researchers have been exploring how to measure lexical complexity in English for over a century (Zamanian and Heydari 2012). In early readability studies it was

---

5 In statistics these correlations are typically expressed using Pearson's *r* on a scale from -1 (variables are negatively correlated) to 1 (variables are linearly correlated), with 0 indicating that there is no correlation between the variables; the further Pearson's *r* is from 0, the greater the correlation; these studies found a positive correlation between a student's vocabulary knowledge and reading ability: Schoonen, Hulstijn, and Bossers 1998: $r = .60 – .76$; Pike 1979: $r = .88 – .94$; Qian 1999: $r = .78 – .82$; Qian 2002: $r = .73 – .77$),

6 Hu and Nation 2000; Laufer and Ravenhorst–Kalovsky 2010; Schmitt, Jiang, and Grabe 2011; cf. Schmitt, Cobb, Horst, and Schmitt 2015.

7 Since a word family "consists of a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately" (Bauer and Nation 1993, 253), the size of a word family for a particular word will grow as the learner's proficiency in the language increases.

8 2000–3000 word families: Laufer 1992, Van Zeeland and Schmitt 2012; 5000 individual words: Hirsh and Nation 1992, Laufer 1989; 8000–9000 word families: Nation 2006; Schmitt et al. 2011.

9 Diederich 1939; see below for a fuller discussion of Diederich's list and dataset. It is impossible to provide an exact accounting of a language's total vocabulary; but comparison of the 171,476 words in current use words in the *Oxford English Dictionary* (2nd ed., 1989) with the 39,589 words in the *Oxford Latin Dictionary* (1968) suggests the different scales of English and Latin vocabulary.

measured with two basic tests: word length and word frequency. More recent research has identified additional lexical features that can make a text easier or harder to comprehend. In addition to word length and word frequency, John Read identifies three other commonly used measures of lexical complexity: lexical sophistication, lexical density, and lexical variation (2000). Thus, the five most common measures of lexical complexity in recent studies are: (1) **word length**; (2) **word frequency**, or the percentage of high frequency words in a text; (3) **lexical sophistication**, or the percentage of low frequency words that provide more precise and more nuanced meanings; (4) **lexical density**, or the ratio of content words to function words; and (5) **lexical variation**, or the variety of different words used in the text.

3.1 Word length. Since English words of greater length are generally considered more difficult to process than words of short (one syllable) or medium (two syllables) length, it has been hypothesized that shorter words are easier to comprehend. Thus, many popular readability measures — such as the *Flesch Reading Ease Test*, the *Flesch-Kincaid Grade Level Test*, and the *Gunning Fog formula* — assess texts by analyzing word length and the number of syllables per word. There have to date been no studies about the effect of word length on comprehension in inflected historical languages.

3.2 Word frequency. Vocabulary frequency is the second most common vocabulary measure in readability studies. The more frequently a word appears, it is reasoned, the more likely that the reader will see it and learn it, as learners are more likely to acquire more frequently seen words than those less frequently seen. In English, West's *General Service List of English Words* and the *British National Corpus* of high-, mid-, and low-frequency words are examples of attempts to characterize word frequency. In Latin, two lists from the first half of the twentieth century, those by Lodge and Diederich, have had significant impact (Muccigrosso 2004). In 1907, Gonzales Lodge compiled a list based on the most commonly read portions of Caesar (*Gallic Wars,* Books 1–5), Cicero (*Catilinarians*, *On Pompey's Command*, and *For Archias*), and Vergil (*Aeneid, Books* 1–6). His list of the most frequent 2000 words was based primarily on words that appeared five or more times in his data. In 1939, Paul Diederich sampled texts of more than two hundred authors "from Ennius to Erasmus" for his dissertation, *The Frequency of Latin Words and their Endings*. From this corpus he produced two core lists that are more representative of Latin literature through the ages than that of Lodge: the *Diederich 300* and the *Diederich 1500*.[10] The *Diederich 300* comprises the 307 words that occur 100 times or more within his sample corpus, representing 58% of all the words in his data. For the longer list, the *Diederich 1500*, he began with the 1556 words that occur twenty times or more in his sample but then removed 85 words that are "important chiefly in medieval Latin." The resulting list of 1471 words account for 83.6% of the vocabulary in his corpus. In 2013, a team at Dickinson College led by Christopher Francese build on the work of Lodge and Diederich to create a core list of almost 1000 words for the *Dickinson College Commentaries* series; this list represents approximately 70% of the words in a typical Latin text. Unlike the two older lists, the DCC team took advantage of the digitized database of Latin words from the Dictionnaire fréquentiel et index inverse de la langue latine. This database, compiled by the team at the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA), is nearly four times larger than that used by Diederich, which in turn is approximately two and a half times larger than that of Lodge (Francese 2018; Muccigrosso 2004). Thus, the Diederich 300 and DCC Latin Core Vocabulary provide sufficient proxies for high frequency words across a range of Latin texts and genres. Texts that contain a high percentage of words in these lists can reasonably be considered more likely to be accessible than lower scoring texts.

10 To create his sample, Diederich used the texts that appeared in *The Oxford Book of Latin Verse* (1912), Avery's *Latin Prose Literature* (1931), and Beeson's *Primer of Medieval Latin* (1925).

3.3 Lexical sophistication. While the frequency of commonly occurring words is a standard measure of text difficulty, researchers also endeavor to find other ways to measure lexical sophistication and the effect that the number of unusual or advanced words has on the readability of a text (Kyle and Crossley 2015). In ESL lexical studies, sophisticated words might be defined as words introduced in Grade 9 or later (Linnarud 1986). Another approach is to rely on the *Academic Word List* (Coxhead 2000). Yet another approach defines lexical sophistication as words that are above a certain frequency threshold, such as words "beyond the basic 2000" (Laufer and Nation 1995; Lu 2012). Similarly, the Dale-Chall formula computes the number of words that do not appear in the 3000 easy words understood by 80% of fourth graders. It is argued that these rarer words reveal more precision and sensitivity to register and tone and therefore are more difficult to understand. In studies measuring L2 learners' lexical sophistication in their writing, researchers define this as the percentage of advanced words in a text divided by the total number of words. Since there are no academic word lists for Latin, this study will consider words not in the *Diederich 1500* as words that are rarer and hence more sophisticated.

3.4 Lexical density. Lexical density can be defined as the number of lexical or content words divided by the total number of words in a text (Ure 1971; Read 2000). Lexical words — nouns, adjectives, verbs, and adverbs — are the primary carriers of meaning and provide readers with more content than do function words such as pronouns, prepositions, conjunctions, and interjections. For example, a sentence such as "The quick brown fox jumped swiftly over the lazy dog" contains seven (underlined) content words out of ten total words for a 70% lexical density score. In contrast, a sentence such as "She told him that she loved him" yields only two content words out of seven total for a lexical density score of 29%. In general, a text with a lower lexical density is easier to understand, and spoken discourse tends to have a lower lexical density than written texts (Belinda 2007; Halliday 1985; Johansson 2008; Yu 2010). Ure (1971) found that the lexical density of written texts was usually more than 40% while the lexical density of spoken texts was generally below 40%. Although function words may be high frequency words, in L1 studies, children acquire function words more slowly than content words, perhaps because function words provide redundant information (Goodman et al. 2008).

3.5 Lexical variation. How often individual words are repeated within a text can also contribute to its lexical complexity of a text. The traditional way to measure lexical variation is to find the Type-Token Ratio (TTR), which is the ratio of the number of unique words (referred to as types) to the total number of words (or tokens) in a text. Traditional TTR, however, exhibits a bias against shorter texts. In general, the longer a text at the same level becomes, the lower its traditional TTR value will fall, since it is often necessary to re-use a number of function words before using a new lexical word. For example, in comparing the lexical richness of *Moby Dick* and *Sense and Sensibility*, Matthew Jockers determined that *Moby Dick*, with 16,872 unique words (types), exhibited more than 2.6 times as much lexical variation as *Sense and Sensibility,* which has 6325 unique words. But once the different lengths of the texts were taken into consideration, Melville's adjusted TTR (29%) surpassed Austen's (19%), by only a factor of 1.5. Three common variants on TTR attempt to compensate for this bias in traditional TTR: Root TTR (Guiraud 1960), Log TTR (Herdan 1964), and Corrected TTR (Carroll 1964).[11] With this bias in mind, we selected for analysis texts there were at least 2,000 words long and, for those texts that were considerably longer (e.g., the *Gospel of John* at 14,058 words or Petronius at 30,958), we analyzed excerpts of comparable length.

11 Lu (2012: 205) reports that "among the seven TTR measures applied to the total vocabulary, the three transformations of the original TTR, namely, CTTR, RTTR, and the D measure, performed the best." We chose not to use the D measure (Malvern et al. 2004) for this study because analysis has shown that D is most effective for shorter texts of 100–400 word tokens (McCarthy and Jarvis 2007); the shortest text in our sample was 2149 tokens (Catullus 69–116).

4. Textual Corpus

For this preliminary study, we deemed it reasonable to choose a few works that are commonly taught either in the first year of learning the language, at the intermediate stage, and finally at the more advanced level. To facilitate this study, we included in our sample mostly texts that had previously been lemmatized — i.e., the texts had been tokenized (divided into constituent words) and then each of these tokens had been matched to a dictionary headword or lemma. We did, however, lemmatize three new texts: *Via Periculosa* and two texts from the *Vulgate*.

Table 1. Sample Texts (by category and alphabetical)

Title	Category	Total Word Count	Unique Words or Types
38 Latin Stories	Beginner	5054	1046
Olimpi, *Via Periculosa*	Beginner	2647	291
Oxford Latin Course, College Ed. Chs. 1–10	Beginner	4181	735
Oxford Latin Course, College Ed. Chs. 24–31	Beginner	4224	1181
Eutropius, *Breviarium* Books 1 & 3	Intermediate	3367	891
Vulgate, Genesis 37, 39–43 (Story of Joseph)	Intermediate	3472	807
Vulgate, Gospel of John 1–8	Intermediate	6159	650
Ritchie, *Fabulae Faciles*, "Hercules" (12–56)	Intermediate	4541	1011
Apuleius, Metamorphosis (Finkelpearl)	Advanced	4643	1934
Caesar, *Bellum Gallicum* Book 1	Advanced	8259	1400
Catullus 1–60	Advanced	4514	1525
Catullus 61–68	Advanced	6428	2041
Catullus 69–116	Advanced	2149	833
Petronius 26–38, 41–49	Advanced	4453	1525
Vergil, *Aeneid* Book 1	Advanced	5147	1660

We chose four texts intended for students in beginning Latin: Groton and May, *38 Latin Stories*; Olimpi, *Via Periculosa*; and two selections from the narrative stories in *Oxford Latin Course, College Edition* (Chapters 1–10 and 24–31). For example, the subtitle for *38 Latin Stories* – "Designed to accompany Frederic M. Wheelock's Latin: An Introductory Course Based on Latin Authors" – makes explicit the target audience. The "Foreword" further indicates that the first eighteen stories are the author's compositions, "often inspired by Ovid," and the last twenty are adaptations of passages "from Caesar, Catullus, Cicero, Horace, Livy" (1). Olimpi's novella is based on Petronius' *Satyrica* 62–63 and the Author's "Preface" sets out his principles in adapting Petronius: he has deliberately "sheltered" the vocabulary, limiting the number of words a reader was expected to recognize to 88 common Latin words and glossing other words as they appear. Finally, the *Oxford Latin Course, College Edition* is a well-known first-year textbook that employs the reading approach, developing a continuous narrative that roughly follows the life of Horace. Chapters 1–10 recount a young Horace's life with his family; his travails at school, which include a simplified account of the Trojan War and Books 3 and 4 of the *Aeneid*; and his father's decision to send him to Rome to continue his studies. Chapters 24–31, the concluding chapters of the textbook, begin with Horace experimenting as a poet, being invited into the circle of

Maecenas, journeying to Brundisium, meeting Propertius and Tibullus, receiving a country villa from Maecenas, learning of the battle of Actium, becoming a friend of the *princeps*, and feeling that death was imminent. Nearly each of these later chapters include an excerpt from the poetry of Horace or his contemporaries. All four first-year texts provide generous help with vocabulary by glossing unfamiliar words.

For our intermediate sample texts we selected a modern work, an ancient historical work, and two biblical texts that did not offer complex syntactical challenges: the story of Hercules in *Ritchie's Fabulae Faciles* (Chapters 12–56); Books 1 and 3 of Eutropius' *Breviarium Historiae Romanae* or *Epitome of Roman History*; the story of Joseph and his brothers in *Genesis*, Chapters 37 and 39–43; and the *Gospel of John*.[12] Unlike some of the first-year texts, the vocabulary is not sheltered in any of these texts. Ritchie in his "Preface" to the 1884 edition emphasizes the difficulties of accidence, syntax, and idiom, but does not mention anything about the principles of vocabulary in retelling the stories. Neither does Kirtland in his 1903 revised edition. Eutropius' *Breviarium* served as a popular school text from the eighteenth to mid-twentieth century and at the turn of the twentieth century was available in fourteen different editions, suggesting its perceived usefulness as "a bridge between the textbook and the authors normally read during the second year (Beyer 2009, xi). The *Breviarium* traces Roman history in ten books from the founding by Romulus to the accession of Valens as emperor in 364 CE. Book 1 summarizes the seven kings of Rome, the overthrow of Tarquin the Proud by Brutus, and Camillus' defeat of the Gauls; Book 3 focuses on the second Punic War. Both are available in recent editions designed to be taught at the advanced-beginner-to-intermediate levels.[13] Jerome's translation of Genesis was a logical choice as an intermediate text since "the texts are considerably easier to read than mainstream Classical authors because the underlying Hebrew/Aramaic models have a simple, paratactic structure with highly repetitive vocabulary" (Clauss 2018). The story of Joseph was selected because it formed a coherent narrative of a length similar to those in the other texts being analyzed. Finally, the *Gospel of John* was selected because it offers a coherent, easy-to-follow narrative with a high degree of repetition that seemed appropriate for intermediate readers. Indeed, it is frequently listed as the book of the New Testament that is easiest to read (e.g., Wallace 2013).

At the advanced level, five authors were chosen: three prose (Caesar's *Gallic War* 1; Petronius' *Satyrica*; the Finkelpearl selections of Apuleius' *Metamorphoses*) and two poets (Vergil, *Aeneid* 1 and Catullus, whose *liber* was sectioned into its three traditional sub-divisions). A variety of styles and genres (historical *commentarii*, novel, lyric, elegy, epigram, and epic) were sought to offer a reasonable sampling of the kinds of texts student regularly encounter at this level. Three of these texts have been included on syllabi for Advanced Placement Latin, while the *Satyrica* and *Metamorphoses*—because of their subject matter, uncanonical (i.e., non-Ciceronian) style, and densely textured literary imitation and sophisticated parody of traditional texts—have rarely been taught at the high school level but in the past few decades have begun to be included in the college curriculum with increasing frequency.[14]

5. Method

Calculations for this study were made using the lemmatized textual data created for *The Bridge* (bridge.haverford.edu), a free web-based tool that allows users to produce customized

12 *Genesis* 38, the digression on Judah and Tamar, was excluded from the sample.

13 Beyer's editions of Eutropius (2009, 2015) are both subtitled "authentic Latin Prose for the Beginning Student." Vanderpool and Katsenes (2016), moreover, have produced an annotated, online version of Eutropius, Books 4–6 that is intended to be "a digital bridge to authentic Latin."

14 Finkelpearl states that her edition of Apuleius "is designed for upper intermediate to advanced students" (2012, ix); Severy-Hoven's recent student edition of Petronius (*The Satyrica of Petronius: An Intermediate Reader with Commentary and Guided Review*) claims to be "appropriate for third- or fourth-semester undergraduate Latin courses or advanced high school Latin" (2014, xiii).

Greek and Latin vocabulary lists. *The Bridge* generates vocabulary lists from a database of fully lemmatized texts—i.e., texts in which every word has been matched to its dictionary headword. For most *Bridge* texts, a combination of machine tools and human effort created these data, producing a level of accuracy beyond what is currently possible using machine tools alone.[15] To lemmatize a text, a digital version of the text (e.g., Olimpi's *Via Periculosa* or the *Gospel of John*) was first pre-processed using a Python-based application known as *Bridge/Lemmatizer*.[16] *Bridge/Lemmatizer* employs the *Classical Language Toolkit* tokenizer and a lemmatization table to identify unambiguous word-forms that match a single lemma.[17] The lemmata of unambiguous words, usually around 60% in a typical Latin text, are then converted into a format that allows the lemmata to be matched to data in the *Bridge Dictionary* and added to a lemmatization spreadsheet. The lemmatization sheets are then completed by Latin readers, who verify the automatically-identified unambiguous lemmata and manually match each ambiguous form with its proper Bridge lemma.

Using these data as a foundation, we created an Excel spreadsheet for each sample text. For every word in a sample text, we imported its part of speech from the *Bridge Dictionary*; noted whether the word was a proper noun or adjective; marked the first appearance of the word in a text, and indicated whether the word appeared in any of the sample Core Lists (Diederich and DCC).

From this spreadsheet we generated the following variables:

The *Word Count* is a simple sum of all words or tokens in each sample text.[18]

Average Word Length was calculated using Excel's native "LEN" function.

Unique Words (or types) is the sum of a column that flagged the first time a word (token) appeared in the text.

Word Frequency was calculated as the percentage of core words in a text (i.e., those also found in the *Diederich 300* and *DCC Latin Core Vocabulary*; but excluding proper names) divided by the total number of words.[19]

Lexical Sophistication was calculated as the percentage of advanced words in a text (i.e., those not in the *Diederich 1500*; but excluding proper names) divided by the total number of words.

To calculate *Lexical Density*, the sum of content words (i.e., nouns, adjectives, verbs, adverbs, numbers, or idioms) was divided by the *Word Count* and multiplied by 100 to generate the number of content words per 100 words.

Lexical Variation was calculated using four standard methods. TTR or Type-Token Ratio is the ratio of the number of unique word tokens or types (V) to the total number of word tokens in a text (N). We also calculated three standard methods to compensate for the susceptibility of TTR $\left(\frac{V}{N}\right)$ to be distorted by the length of the

15 Data for all sample texts were created using this process with the exception of Caesar, Catullus, and Vergil, which were lemmatized by scholars of the *Laboratoire d'Analyse Statistique des Langues Anciennes* (LASLA). Excluding those texts lemmatized by LASLA, the raw lemmatization sheets for all treated texts are available at github.com/GitClassical/Bridge/data/Readability_NECJ_Data. Detailed information about the lemmatized texts used for this study is available at: https://bridge.haverford.edu/about/texts.

16 *Bridge/Lemmatizer* is available on-line at https://bridge.haverford.edu/lemmatize/Lemmatizer.

17 The *Classical Language Toolkit* or CLTK (cltk.org) offers open-source code to support the natural language processing of the languages of Ancient, Classical, and Medieval Eurasia. The texts for this study were pre-processed using the CLTK lemmatizer version 0.1.91.

18 Our data for this study cleaved enclitics, creating two separate words that were then used as the basis for the statistical analysis; for example, *virumque* was treated as *virum* and *-que* in our sample. As a result, the *Word Count* was increased and the *Average Word Length* similarly decreased compared to the raw text. Given the general infrequency of enclitics in Latin, the effect of this decision on our results was minimal.

19 When proper names are excluded, they were not assumed to be known but were removed from the calculation entirely (i.e., from the measure count and the count of total tokens).

treated text: *R* or Guiraud's Root TTR $\left(\frac{V}{\sqrt{2N}}\right)$; *CTTR* or Carroll's Corrected TTR $\left(\frac{V}{\sqrt{2N}}\right)$; and *Log TTR* or Herdan's C $\left(\frac{logV}{LogN}\right)$. A *Composite Score* was calculated for each sample text using an average of each text's rank in (1) *Average Word Length*; (2) the percentage of words in the *Diederich 300* and DCC lists; (3) *Lexical Sophistication*; (4) *Lexical Density;* and (5) an average rank of the four calculations of *Lexical Variation*.

Finally, from these variables a new measure of lexical readability — *LexR* — was created. *LexR* factors several relevant measures of lexical complexity to produce a single score of a Latin text's lexical complexity. Unlike the *Composite Score*, which can only be used to compare the relative difficulties the fifteen texts within our sample corpus*, LexR* is derived from a fixed formula into which data from any Latin text could be entered. *LexR,* therefore, offers a reproducible method that can be applied to any lemmatized Latin text.

6. Results and Discussion

6.1 Word length. In early English readability studies, word length was considered a good gauge of a text's difficulty: the shorter words being easier to read and quickly comprehend than longer ones.

Table 2. Average Word Length in Sample Texts

Rank	Title	Average Word Length
1	Olimpi, *Via Periculosa*	4.91
2	*Vulgate, Gospel of John* 1–8	4.92
3	Catullus 69–116	5.35
4	*38 Latin Stories*	5.39
5	Catullus 1–60	5.47
6	*Oxford Latin Course, College Ed.* 1–10	5.48
7	*Vulgate, Genesis* 37, 39–43 (Story of Joseph)	5.55
8	Petronius 26–38, 41–49	5.61
9	Catullus 61–68	5.73
10	Vergil, *Aeneid* Book 1	5.78
11	*Oxford Latin Course, College Ed.* 24–31	5.80
12	Eutropius, *Breviarium* Books 1 & 3	5.88
13	Caesar, *Bellum Gallicum* Book 1	6.21
14	Apuleius, *Metamorphoses* (Finkelpearl)	6.27
15	Ritchie, *Fabulae Faciles*, "Hercules" (12–56)	6.36

Appropriately, Olimpi's *Via Periculosa* exhibited the shortest average word length (4.9) but the *Gospel of John* nearly matched it with the same rounded length. Catullus' polymetrics (5.3) and epigrams (5.5) exhibit word length similar to *38 Latin Stories* (5.4). *OLC* 1–10 (5.5) and the story of Joseph in *Genesis* (5.5), followed by Petronius (5.6), Catullus' long poems (5.7), and Vergil (5.8). Interestingly, the texts with the longest average word length — *OLC* 24–31 (5.8), Eutropius (5.9), and Caesar (6.2), Apuleius (6.3), and *Fabulae Faciles*

(6.4) — appear across our presumed beginner, intermediate, and advanced levels. The only poetic texts that approaches this group are Catullus 61–68 (5.7) and Vergil (5.8).

Poetic texts — with their large number of adjectives, present tenses, enclitics; and fewer compound verbs — tend to have a shorter word length. Prose texts, conversely, also tend to emphasize narratives using past tenses (imperfect, perfect, and pluperfect tenses) and thus tend to have longer verbs because the stems and endings for these verbs increases their length, thereby increasing the number of long words. Conversely, while there is some variability in mean word length (between 4.9 and 6.4 letters long), it is unclear whether this amount of variation is meaningful for comprehension.[20] The elimination of proper nouns did not significantly affect the rankings of mean word length, except for Eutropius, which fell from one of the longer average lengths to close to the median.

6.2 Word frequency. We also examined the extent to which the sample texts comprise high frequency vocabulary. For this measurement, we removed proper nouns and adjectives from consideration, since their semantic requirements, while not trivial, are different than other words and their capitalization in edited texts effectively signals to students that they belong to familiar class of words independent of previous knowledge of the word. Two frequency lists were used for this measurement: The *Diederich 300* list (58% coverage), the *DCC Latin Core Vocabulary* (70% coverage).[21]

Table 3. Word Frequency (% of total words in select core lists, excluding proper nouns)

Rank (average)	Title	% in *Diederich 300*	% in *DCC Latin Core*
1	*Vulgate, Gospel of John* 1–8	76.57	86.81
2	*38 Latin Stories*	69.39	86.39
3	Eutropius, *Breviarium* Books 1 & 3	59.36	81.13
4	*Oxford Latin Course, College Ed.* 1–10	57.97	82.18
5	Olimpi, *Via Periculosa*	68.24	79.99
6	Caesar, *Bellum Gallicum* Book 1	61.56	80.38
7	Ritchie, *Fabulae Faciles*, "Hercules" (12–56)	57.82	80.17
8	*Genesis* 37, 39–43 (Story of Joseph)	65.98	78.41
9	*Oxford Latin Course, College Ed.* 24–31	56.30	77.47
10	Catullus 69–116	58.16	74.70
11	Vergil, *Aeneid* Book 1	48.81	71.39
12	Catullus 1–60	56.69	71.00
13	Petronius 26–38, 41–49	54.32	67.79
14	Catullus 61–68	46.02	66.02
15	Apuleius, *Metamorphoses* (Finkelpearl)	43.12	57.52

Four texts exhibited a high proportion of words from the *Diederich 300* list: *Gospel of John* (76.6%), *38 Latin Stories* (69.4%), *Via Periculosa* (68.2%), and *Genesis* 37, 39–43

20 There is little correlation between word length and the other measures of lexical difficulty — e.g., percentage of words outside the Diederich 1500 ($r = 0.106$) and Log TTR ($r = 0.219$); since the effect of longer word length on Latin comprehension has not been studied, it would be premature to place too much weight on this measure.

21 We use the Diederich 1500 (84% coverage) for the measure of *Lexical Sophistication* (see below).

(66.0%), while three others contained a far lower percentage of words from the most common Latin vocabulary: Vergil (48.8%), Catullus' long poems (46.0%), and Apuleius (43.1%). When examining texts with a high percentage of words from the DCC list, seven texts cleared the threshold of having more than 80% of their vocabularies drawn from high frequency words. The *Gospel of John* (86.8%) and *38 Latin Stories* (86.4%) again led the way. Five other texts surpassed the 80% threshold: *OLC* 1–10 (82.2%), Eutropius (81.1%), Caesar (80.4%), *Fabulae Faciles* (80.2%), and *Via Periculosa* (80.0%). The poetic texts and Roman novels contained a below average percentage of the high-frequency words on the DCC list: Vergil (71.4%), Catullus 1–60 (71.0%), Petronius (67.8%), and Catullus' long poems (66.0%). Apuleius was a further outlier in the category (57.5%).

6.3 Lexical Sophistication. As a measure of their lexical sophistication or rareness, we calculated the percentage of words (excluding proper names) in our fifteen Latin texts that fall outside the *Diederich 1500* list (84% coverage for the texts in his sample).

Table 4. Lexical Sophistication in Sample Texts

Rank	Title	% Total Words outside *Diederich 1500* (excluding Proper Names)
1	*38 Latin Stories*	7.41
2	*Vulgate, Gospel of John* 1–8	7.57
3	Eutropius, *Breviarium* Books 1 & 3	7.58
4	*Oxford Latin Course, College Ed.* 1–10	9.05
5	Ritchie, *Fabulae Faciles*, "Hercules" (12–56)	13.17
6	*Oxford Latin Course, College Ed.* 24–31	13.50
7	Olimpi, *Via Periculosa*	13.70
8	*Genesis* 37, 39–43 (Story of Joseph)	14.43
9	Caesar, *Bellum Gallicum* Book 1	14.75
10	Vergil, *Aeneid* Book 1	18.68
11	Catullus 69–116	20.30
12	Catullus 1–60	23.58
13	Catullus 61–68	23.72
14	Petronius 26–38, 41–49	26.32
15	Apuleius, *Metamorphoses* (Finkelpearl)	33.64

Unsurprisingly, the Latin novels and the poetic texts revealed the greatest lexical sophistication and included a large percentage of words not found in Diederich: Apuleius (33.6%), Petronius (26.3%), Catullus 61–68 (23.7%), Catullus 1–60 (23.6%), Catullus 69–116 (20.3%) and Vergil (18.7%). Appropriately, first-year and intermediate texts limited their use of "sophisticated" words "beyond Diederich" to 15% of their total words or less: *38 Latin Stories* (7.4%), the *Gospel of John* (7.6%), Eutropius (7.6%) and *OLC* 1–10 (9.1%) showed the least lexical sophistication, with two modern texts — *Fabulae Faciles* (13.2%), *OLC* 24–31 (13.5%) and *Via Periculosa* (13.7%) — in the next group, and two ancient works, *Genesis* (14.4%) and Caesar (14.8%), at the higher end of the range among the intermediate prose texts.

Because of the level of historical and cultural knowledge required to process frequent proper names, we compared our standard measure of lexical sophistication with one that included proper names. The absolute lexical sophistication of most texts remained similar, with the typical text gaining 4–6% of "sophisticated" vocabulary. Only Eutropius saw its ranking change appreciably. Because of its high number of unique proper names, Eutropius displayed a much higher degree of lexical sophistication when proper names were included (23.5% vs. 7.6%).

6.4 Lexical Density. Lexical density tracks the number of lexical (or content) words that are the primary carriers of meaning — nouns, adjectives, verbs, and adverbs — divided by the total number of words in a text (Ure 1971; Read 2000). A text with a high lexical density has higher propositional content and so is likely more difficulty to comprehend.

Table 5. Lexical Density in Sample Texts

Rank	Title	Density Score
1	*Gospel of John* 1–8	59.75
2	*Genesis* 37, 39–43 (Story of Joseph)	67.66
3	Caesar, *Bellum Gallicum* Book 1	69.56
4	Catullus 69–116	71.75
5	Catullus 1–60	71.78
6	Ritchie, *Fabulae Faciles*, "Hercules" (12–56)	72.08
7	Olimpi, *Via Periculosa*	73.10
8	*38 Latin Stories*	73.55
9	Petronius 26–38, 41–49	73.63
10	*Oxford Latin Course, College Ed.* 24–31	75.17
11	Eutropius, *Breviarium* Books 1 & 3	77.00
12	Vergil, *Aeneid* Book 1	77.17
13	Apuleius, *Metamorphoses* (Finkelpearl)	77.73
14	*Oxford Latin Course, College Ed.* 1–10	78.33
15	Catullus 61–68	79.00

All fifteen sample texts surpassed Ure's benchmark for written texts (40%), ranging from the two Vulgate texts (59.7% and 67.7%), Caesar (69.6%), and Catullus's polymetrics and epigrams (71.8%) at the lower (i.e., less dense and so more readable) end of the distribution compared to Eutropius (77.0%), Vergil (77.2%), Apuleius (77.7%), *OLC* 1–10 (78.3%) and Catullus 61–68 (79.0%) at the higher end. It is not surprising that Latin texts exhibit such high lexical density. First, Latin typically uses pronouns only for emphasis and indicates the subject in the verb ending. And second, Latin case usage, especially the genitive, dative, and ablative, reduces the need for prepositions.[22] Yet studies have not shown

[22] The minimization of pronouns and prepositions explains the unexpectedly high lexical density of the stories in the first ten chapters of the *Oxford Latin Course for College,* since prepositions and conjunctions are rare in the first few chapters and pronouns are only formally introduced in Chs. 6 and 7.

a correlation between the lexical density of written texts and their readability. In studies of lexical density in student writing, the results do not demonstrate a significant relationship between the proportion of content words and the quality of essays (Cheryl 1995; Kalantari and Gholami 2017; Linnarud 1986; Vidakovic and Barker 2009). Similarly, To, Fan, and Thomas (2013) reported no strong link between lexical density and readability. Thus, it is unclear whether the lexical density results in this study suggest a meaningful relationship between this metric and the readability of the Latin texts.

6.5 Lexical Variation. In attempting to measure the lexical variation or diversity of the fifteen texts in our sample corpus, this study employed four common methods for calculating the Type-Token Ratio: traditional TTR, Root TTR, Log TTR, and Corrected TTR. Traditional TTR displayed the expected bias towards longer texts.

Table 6. Lexical Variation of Sample Texts (ranked by average of methods)

Rank	Title	TTR	RootTTR	CTTR	LogTTR
1	Olimpi, *Via Periculosa*	0.11	5.66	4.00	0.72
2	*Gospel of John*	0.11	8.28	5.86	0.72
3	*Oxford Latin Course, College Ed.* 1–10	0.18	11.37	8.04	0.79
4	*38 Latin Stories*	0.21	14.71	10.40	0.82
5	Caesar, *Bellum Gallicum* Book 1	0.17	15.41	10.89	0.80
6	*Genesis* 37, 39–43 (Story of Joseph)	0.22	13.99	9.90	0.82
7	Ritchie, *Fabulae Faciles*, "Hercules" (12–56)	0.22	15.00	10.61	0.82
8	Eutropius, *Breviarium* Books 1 & 3	0.26	15.36	10.86	0.84
9	*Oxford Latin Course, College Ed.* 24–31	0.28	18.17	12.85	0.85
10	Catullus 69–116	0.39	17.97	12.71	0.88
11	Catullus 1–60	0.34	22.71	16.06	0.87
12	Vergil, *Aeneid* Book 1	0.32	23.14	16.36	0.87
13	Petronius 26–38, 41–49	0.34	22.85	16.16	0.87
14	Catullus 61–68	0.32	25.46	18.00	0.87
15	Apuleius, *Metamorphoses* (Finkelpearl)	0.42	28.38	20.07	0.90

According to the three measures that attempt to compensate for text length, the texts that exhibited the greatest lexical variation were the poetic texts, Petronius, and Apuleius, although Catullus 69–116 showed less lexical diversity than the other texts in this group according to Root TTR and Corrected TTR. At the other end of the scale, *Via Periculosa* and the *Gospel of John* far outstripped the other texts in how little lexical variation was present. In other words, *Via Periculosa* kept its promise to shelter vocabulary and provide students numerous repetitions of the words in the story, while the hymn-like opening of the *Gospel of John* produced a high percentage of repeated words. The various measurements showed some disagreement regarding the lexical diversity of *Genesis*, *Fabulae Faciles*, and Caesar. According to Root TTR and Corrected TTR, Genesis (3475 words) demonstrated less lexical diversity than *Fabulae Faciles* (4541 words) and Caesar (8262 words). Yet Log TTR places *Genesis* on par with *Fabulae Faciles* and as more lexically diverse than Caesar. Similarly, in Root TTR and Corrected TTR, Catullus 69–116—at just 2149 words, the shortest text in our sample corpus—exhibits less lexical diversity than the other Catullan texts as well as Vergil and Petronius. For the texts in our sample, Log TTR seems to provide

less information about degrees of lexical variation than the other measures. Aside from *Via Periculosa* and the *Gospel of John*, Log TTR clusters the introductory and intermediate texts into one narrow band, and Petronius and the poetic texts into a second, with Apuleius as a slight outlier beyond this second band.

6.6 Composite Average. Table 7 compiles the rankings of texts in our sample corpus and generates a Composite Average of each text's ranking across our five key measures of lexical complexity.

Table 7. Compiled Rankings for Sample Texts and Composite Average

Title	Word Length	Word Frequency	Lexical Sophistication	Lexical Density	Lexical Variation	Composite Average
Gospel of John	2	1	3	1	2	1.8
38 Latin Stories	4	2	1	8	5	3.8
Olimpi, *Via Periculosa*	1	5	7	7	1	4.2
Genesis 37, 39–43	7	7	8	2	4	5.6
Oxford Latin Course 1–10	6	4	4	14	3	6.2
Caesar, *Bellum Gallicum* Book 1	13	6	9	3	7	7.2
Eutropius, *Breviarium* Books 1 & 3	12	3	2	11	8	7.6
Ritchie, *Fabulae Faciles* 12–56	14	7	5	6	6	7.6
Catullus 69–116	3	10	11	4	10	7.6
Catullus 1–60	5	11	12	5	11	8.8
Oxford Latin Course 24–31	11	9	6	10	9	9
Vergil, *Aeneid* Book 1	10	11	10	12	12	11
Petronius 26–38, 41–49	8	13	14	9	13	11.4
Catullus 61–68	9	14	13	15	14	13
Apuleius, *Metamorphoses* (Finkelpearl)	15	15	15	13	15	14.6

The Compiled Rankings (Table 7) allows a reader to grasp the relative difficulties of the sample texts according to different lexical measures and to observe which of these texts consistently manifest scores towards the easier end of the distribution (e.g., the *Gospel of John*), which manifest scores towards the more difficult (e.g., Petronius, Apuleius, Vergil), as well those texts that received inconsistent scores and so are "easier" by some measures but more "difficult" by others (e.g., *Oxford Latin Course, College* Chs. 24–31 or Catullus 69–116). This Composite Average, however, has several limitations. Since these rankings are relative, the insights are limited to our sample of texts: it would not be possible to compare a text outside our sample with these data. Moreover, the rankings obscure the true differences in given measures between texts. For example, the difference in lexical sophistication between the first and second ranked texts is only 0.1% but between the fourth and fifth it is 4.1%. It also gives equal weight to each measure.

7. *LexR*: A New Lexical Readability Score for Latin Texts

To facilitate further research on the readability of these and other Latin texts, we sought to establish a new measurement of lexical readability that could be generated for any Latin text for which the requisite textual data were available. Using the lexical data from our sample corpus of fifteen texts it was possible to construct such a measure: *Mulligan's LexR*. To create *LexR* we took the significant measurements identified in this study and subjected them to *Principal Component Analysis* (PCA). PCA is a statistical procedure that takes multiple variables or characteristics and combines them into a single score on a 12-point scale (-6 to +6), while retaining as much information about the differences between the texts as possible.[23] For ease of comprehension, we then converted PCA's default 12-point scale to a more intuitive 10-point scale with 0 indicating minimal and 10 maximal lexical difficulty.[24]

LexR = ((*mean word length* x 0.457) + (*% outside Diederich 300* x 0.063) + (*% outside DCC* x 0.076) + (*% outside Diederich 1500* x 0.092) + (*Log TTR* x 0.312) + (*Root TTR* x 0.143) –11.7) x 0.833

Where, –11.7 is the constant to set base score at 0; the coefficient 0.833 converts the score to a 10-point scale.[25]

Table 8. *LexR* Scores for Sample Texts

Title	LexR Score
Vulgate, Gospel of John 1–8	1.1
Olimpi, *Via Periculosa*	2.0
38 Latin Stories	2.4
Oxford Latin Course, College Ed. 1–10	3.1
Vulgate, Genesis 37, 39–43 (Story of Joseph)	3.5
Ritchie, *Fabulae Faciles*, "Hercules" (12–56)	3.6
Eutropius, *Breviarium* Books 1 & 3	3.8
Caesar, *Bellum Gallicum* Book 1	4.2
Oxford Latin Course, College Ed. 24–31	4.7
Catullus 69–116	5.0
Catullus 1–60	6.2
Vergil, *Aeneid* Book 1	6.5
Petronius 26–38, 41–49	6.8
Catullus 61–68	7.6

[23] Because research on the relative impact of these different measures of lexical difficulty on the comprehension of *Latin* has not yet been undertaken, we did not deem it appropriate to generate the coefficients for *LexR* ourselves but instead allowed these to be created algorithmically via PCA.

[24] We believe that almost all real-world Latin texts will fall within *LexR*'s standard range of 0 to 10; but theoretically texts with greater or lesser scores are possible; for example, a long text consisting of a single, repeated high-frequency word (e.g., the preposition "a" (*ab*) would have a score of -9.5; a text comprising thousands of lengthy words, each used only once and none of which is found in Diederich 1500 would have a score of 25.5.

[25] The precise methods for determining these factors are described in Section 5; the Primary Component Analysis excluded Corrected TTR from the calculation because it is collinear to Root TTR (i.e., it yields same information but on a different scale) and so does not contribute to creating a differential scoring of the texts. We also excluded Lexical Density because of the lack of evidence that it is correlated with reading difficulty (see above) and traditional TTR because of the different lengths of our sample texts and the known bias of traditional TTR against shorter texts.

These results suggest that *LexR* offers a mechanism for distilling the significant measures of lexical complexity discussed in this study into a single, informative, integrated score that provides a sense of the comparative lexical complexity, and so difficulty, of texts. Nevertheless, it is important to reiterate that *LexR* quantifies only one axis of textual difficulty. A reader may not necessarily find a text with a lower *LexR* score easier to read, since the overall accessibility of a text depends on a matrix of factors including its syntax, pragmatics, and assumed cultural background knowledge. Nor is it possible, at this point, to fully characterize how the scale of *LexR* scores relates to the comparative difficulties of a text: that is, it seems well founded to conclude that Catullus 69–116 (5.0) is more lexically challenging than *38 Latin Stories* (2.4) and meaningfully so; but it would be premature to characterize the former as more than twice as lexically difficult as the latter.[26]

8. Conclusions

In general, first-year texts in this study reveal less lexical complexity and a high proportion of high frequency words. Both selections of the *Oxford Latin Course* incorporate fewer high-frequency words from the *Diederich 300* than might be expected, but balance those low scores by employing a high percentage of words from the *DCC Latin Core Vocabulary* and *Diederich 1500* in the earlier chapters and the *Diederich 1500* list by the later chapters. As a result, it reveals more lexical variation than the other first-year texts, especially in the more complex later chapters. This suggests that the lexical difficulty and so the reading level of the text may increase faster than would be ideal once unadapted selections of Horace are introduced in its later chapters.

Most intermediate texts perform nearly as well as first-year texts and the *Gospel of John*, with the lowest *LexR* score of 1.1, exceeds them in many measurements. The lexical variation scores (Root TTR and CTTR) of the biblical texts are comparable to or better than *38 Latin Stories* and their level of lexical sophistication is below the mean. Although there remain questions about the correlation between lexical density and readability, it is interesting that the *Gospel of John* and the story of Joseph in *Genesis* score first and second overall (i.e., lowest) in the measure, perhaps in part because of the higher proportion of prepositions and pronouns in post-classical Latin. They likewise are more similar to the first-year texts than other classical texts in lexical variation (second and fourth, respectively); and in high frequency vocabulary: first and fourth by the *Diederich 300*; first and eighth by the DCC list. These specific results and the *LexR* scores of the *Gospel of John* (1.1) and the story of Joseph (3.5) reinforce recent suggestions that instructors might look to the Latin Vulgate and other postclassical texts as suitable bridges to classical authors (Clauss 2018; Hendrickson 2018).

Eutropius, traditionally thought to be a good school text, exhibits a mixed lexical profile that may challenge second-year students. Its *LexR* score of 3.8 placed it in the middle of our sample corpus, more like Caesar than beginner texts. Moreover, with 596 proper nouns, Eutropius includes proportionately more than twice as many proper nouns as any other text. With those proper nouns removed, the text reveals good scores in high frequency vocabulary and lexical variation. But factoring in those proper nouns, its lexical density is the third highest after only Catullus 61–68 and Vergil. In short, the density of proper nouns and the highly compressed nature of the text may lead intermediate readers to find the propositional content too challenging to manage without a solid knowledge of the major players in Roman history.

Caesar, a canonical prose text, also ranked in the middle of our sample corpus by *LexR* (4.2). While it exhibited the third lowest lexical density after the Vulgate passages, it consistently placed in the middle of our sample corpus for use of high frequency words,

26 Future testing of reading comprehension and comparing with *LexR* could establish whether this scale of lexical difficulty is linear, geometric, exponential, or discontinuous.

lexical sophistication, and lexical variation. These results support its selection as one of the first advanced-level texts that a student might read. Finally, the two Roman novels and the poetic texts of Catullus and Vergil reveal the highest lexical complexity. Although they tend to have shorter words on average than narrative prose texts, they display higher lexical sophistication and lexical variation. In terms of lexical sophistication, Vergil and Catullus 69–116 exhibit slightly less lexical sophistication than the rest of Catullus, Petronius, and Apuleius. In light of their high lexical sophistication and lexical variation, as well as their *LexR* scores, this study would suggest that instructors recognize the lexical challenges of tackling Vergil (6.5), Petronius (6.8), Catullus 61–68 (7.6), and especially Apuleius (9.5)—and the extra vocabulary work that readers and instructors will have to undertake to make these texts readable.

8.1 Implications for teaching

These findings suggest the importance of teaching vocabulary intentionally at all levels of the curriculum. As discussed above, vocabulary knowledge is critical for reading comprehension: without knowing 95–98% of the words in a text, the reader will be challenged to comprehend its meaning. While incidental vocabulary learning may occur as students read these texts, explicit teaching would foreground new words and help students internalize them into their developing lexicon (Nation 2013, 97). Opportunities for involvement, clear focus, repetition, and retrieval can help students achieve higher rates of learning vocabulary. Sample activities might include classifying words (e.g., synonyms, antonyms, positive/negative connotation, living or non-living, cause and effect, cultural context), drawing and labeling pictures, producing word families from one word (e.g., *liberare: liber, liberalis, liberaliter, libertus*), creating compound words from prefixes or suffixes, completing a cloze passage with targeted words, peer teaching of new or important vocabulary, or collaborative reconstruction of a passage (see Nation 2013; Golonka et al. 2015 for more examples). The DCC list and *Diederich 1500* provide excellent targets for students as they progress from year to year.

Even the texts that reveal the least lexical complexity have words that do not appear on any list. Ideally, an instructor would try to identify the highest frequency words that do not appear on these lists for intentional vocabulary instruction. With new tools it would be possible for instructors to craft more intentional paths to vocabulary acquisition for specific texts. For the texts, textbooks, and lists in its database, *The Bridge* allows users to create vocabulary lists for a text or a selection, screening out known words or revealing the words in common between texts. Within the generated list, users can find morphological information, links to lexical resources like *Logeion* (logeion.uchicago.edu), and the frequency of the word within the text or the entire Bridge corpus to quickly identify high- (or low-) frequency words. These lists can then be filtered to focus on one or more parts of speech, among other options, and then printed or downloaded in a variety of formats. For texts not in *The Bridge*, the Perseus (www.perseus.tufts.edu/hopper) and the Alpheios Projects (https://alpheios.net/pages/tools/) will link words to definitions, and the *Perseus Vocabulary Tool* (www.perseus.tufts.edu/hopper/vocablist?lang=la) will automatically generate a list of linked vocabulary, although the process yields a fair number of false returns (e.g., *edo* 'to eat' is listed as the 16th most popular word in *Bellum Gallicum* because of a mislemmatization of *est*). Any list created with the Perseus Tool would have to be corrected before being used as a reliable resource for teaching and reading.

Finally, to help students build their mental lexicon and be better prepared to approach intermediate and advanced historical texts, provide opportunities for extensive reading with Latin novellas (Piazza 2017) and explore medieval texts such as the Latin Vulgate, the travels of Egeria, *Apollonius King of Tyre,* Hrotsvitha's *Dulcitius,* or *The Legend of Barlaam and Josaphat*. Likewise, simplify difficult texts by removing or replacing

some of the more difficult vocabulary or use embedded readings to help students guess the new vocabulary when they see it in the unadapted passage.[27]

8.2 Limitations of the Present Study and Future Directions

This study provides a surface measure of the vocabulary knowledge a reader might need to approach these texts. Future studies might examine additional measures of lexical complexity such as: (1) the range of texts and genres in which a word can be found; (2) n-grams and collocations; (3) concreteness and imageability; (4) discourse features such as lexical cohesion and topic continuity from sentence to sentence; (5) the density of rare words used infrequently (e.g., the presence of hapax legomena); and (6) the number or percent of words a reader needs to reach the 95 or 98% thresholds needed for unassisted comprehension.[28] Moreover, since the different length of the sample texts affects the calculation of Type-Token Ratios, more work will need to be done to ascertain the best method for measuring lexical variation. One possibility would be to take repeated same-length sections of every text under analysis (e.g., word 1 to word 1000, word 2 to word 1001, etc.) to capture an average lexical variation for the text as well as measures of internal variability and slope (i.e., does a text become appreciable easier or more difficult as it unfolds). Finally, this preliminary work on lexical complexity could be complemented by analysis of syntactic features such as clause length, the number of T-units, and the number of subordinate clauses.[29]

Since only a portion of many of the authors in our sample have been analyzed, larger samples from these authors might yield richer results. In addition to the fifteen texts analyzed here, future studies could expand the number, range, and size of texts subjected to similar analysis. Other commonly taught texts (by different authors, in different genres, and from different time-periods) should be analyzed. Ideally, these preliminary results could be the foundation for a variety of experimental studies that test how traditional measures of lexical complexity correlate to the reading comprehension of Latin — and eventually other historical languages.

Cornell College
jgruber-miller@cornellcollege.edu

Haverford College
bmulliga@haverford.edu

27 For sample embedded readings, see *Tiered Reading List for Operation CAESAR* (lapis.practomime.com/index.php/operation-caesar-reading-list).

28 See Vajjala and Meurers 2012; Xia et al. 2016; an "n-gram" refers to sequences of words of a given (*n*) length (e.g., *re vera*; *ei nomen est*).

29 A "T-unit" refers to a main clause and all its subordinate clauses.

Works Cited

Balme, Maurice, and James Morwood. 2012. *Oxford Latin Course, College Edition: Readings and Vocabulary*. New York: Oxford University Press.

Bauer, Laurie, and Paul Nation. 1993. "Word Families." *International Journal of Lexicography* 6.4: 253–279. https://doi.org/10.1093/ijl/6.4.253

Belinda, C. C. 2007. *The Language of Business Studies Lectures: A Corpus-Assisted \ Analysis*. The Netherlands: John Benjamins Publishing Company.

Beyer, Brian. 2008. "Yale University Press to Publish RU Master's Thesis." Rutgers Classics Department Blog. 2 July 2008.

Carroll, J. B. 1964. *Language and Thought*. Englewood Cliffs, NJ: Prentice-Hall.

Chall, J. S. 1958. *Readability: An Appraisal of Research and Application*. Columbus, OH: The Ohio State University Bureau of Educational Research.

Chall, J. S., and E. Dale. 1995. *Readability Revisited, the New Dale-Chall Readability Formula*. Cambridge, MA: Brookline Books.

Clauss, James J. 2018. "Teaching the Old and New Testaments to Students of Greek and Latin Simultaneously with Numerous and Fascinating Outcomes." *Teaching Classical Languages* 10.1: 99–125.

Coxhead, A. 2000. "A new academic word list." *TESOL Quarterly* 34.2: 213–38. https://doi.org/10.2307/3587951

Dale, E. and J. S. Chall, 1949. "The Concept of Readability." *Elementary English* 26: 23.

Diederich, Paul. 1939. *The Frequency of Latin Words and Their Endings*. Diss. Chicago.

DuBay, W. H. 2004. *The Principles of Readability*. Costa Mesa: Impact Information.

Finkelpearl, E. 2012. *An Apuleius Reader: Selections from the Metamorphoses*. Wauconda, IL: Bolchazy-Carducci Publishers.

Francese, Christopher. 2018. "Core Vocabulary." Dickinson College Commentaries.

Golonka, Ewa, Anita Bowles, Noah Silbert, Debra Kramasz, Charles Blake, and Tim Buckwalter. 2015. "The Role of Context and Cognitive Effort in Vocabulary Learning: A Study of Intermediate-Level Learners of Arabic." *Modern Language Journal* 99.1: 19–39. https://doi.org/10.1111/modl.12191

Goodman, J., P. Dale, P. Li. 2008. "Does Frequency Count? Parental Input and the Acquisition of Vocabulary." *Journal of Child Language* 35: 515–31. https://doi.org/10.1017/S0305000907008641

Grabe, William, and Fredricka Stoller. 2011. *Teaching and Researching Reading*. 2nd ed. Harlow, UK: Pearson/Longman.

Guiraud, P. 1960. *Problèmes et méthodes de la statistique linguistique*. Dordrecht, The Netherlands: D. Reidl.

Halliday, M. A. K. 1985. *Spoken and written language*. Melbourne, AUS: Deakin University Press.

Hendrickson, T. 2018. "Why So Few of Us Teach Neo-Latin and Why More Of Us Should." *Eidolon* 12 March 2018. <https://eidolon.pub/why-so-few-of-us-teach-neo-latin-3f85eb1984b6>

Herdan, G. 1964. *Quantitative Linguistics*. London: Butterworths.

Hirsch, D., and I. S. P. Nation. 1992. "What Vocabulary Size is Needed to Read Unsimplified Texts for Pleasure?" *Reading in a Foreign Language* 8: 689–96.

Hu, M., and I. S. P. Nation. 2000. "Unknown Vocabulary Density and Reading Comprehension." *Reading in a Foreign Language* 13.1: 403–30.

Jockers, Matthew. 2013. *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press. https://doi.org/10.5406/illinois/9780252037528.001.0001

Johansson, Victoria. 2008. "Lexical Diversity and Lexical Density in Speech and Writing: A Developmental Perspective." *Lund University. Department of Linguistics and Phonetics Working Papers* 53: 61–79.

Kalantari, Reza, and Javad Gholami. 2017. "Lexical Complexity Development from Dynamic Systems Theory Perspective: Lexical Density, Diversity, and Sophistication." *International Journal of Instruction* 10.4: 1–18. https://doi.org/10.12973/iji.2017.1041a

Kirtland, John Copeland, ed. 1903. *Ritchie's Fabulae Faciles: A First Latin Reader*. New York: Longmans, Green, and Co.

Kyle, Kristopher, and Scott Crossley. 2015. "Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application." *TESOL Quarterly* 49.4: 757–86. https://doi.org/10.1002/tesq.194

Laufer, Batia. 1989. "What Percentage of Text-Lexis is Essential for Comprehension?" In *Special Language: From Humans to Thinking Machines,* edited by C. Lauren and M. Nordman, 316–23. Clevedon, UK: Multilingual Matters.

Laufer, Batia. 1992. "How Much Lexis is Necessary for Reading Comprehension? In *Vocabulary and Applied Linguistics*, edited by P. J. L. Arnaud and H. Béjoint, 126–32. Basingstoke, UK: Macmillan. https://doi.org/10.1007/978-1-349-12396-4_12

Laufer, Batia. 1997. "The Lexical Plight in Second Language Reading: Words You Don't Know, Words You Think You Know and Words You Can't Guess." In *Second Language Vocabulary Acquisition: A Rationale for Pedagogy*, edited by J. Coady and T. Huckin, 20–33. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524643.004

Laufer, Batia, and Tami Aviad-Levitzky. 2017. "What Type of Vocabulary Knowledge Predicts Reading Comprehension: Word Meaning Recall or Word Meaning Recognition?" *Modern Language Journal* 101.4: 729-41. https://doi.org/10.1111/modl.12431

Laufer, Batia, and I. S. P. Nation. 1995. "Vocabulary Size and Use: Lexical Richness in L2 Written Production." *Applied Linguistics* 16.3: 307–22. https://doi.org/10.1093/applin/16.3.307

Laufer, Batia, and Geke Ravenhorst-Kalovski. 2010. "Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension." *Reading in a Foreign Language* 22.1: 15–30.

"Lexical Density." 2019. *Analyze My Writing*.

Linnarud, M. 1986. *Lexis in Composition: A Performance Analysis of Swedish Learners' Written English*. Lund: Sweden: CWK Gleerup.

Lodge, Gonzales. 1907. *The Vocabulary of High School Latin: Being the Vocabulary of Caesar's* Gallic War*, Books I–V; Cicero* Against Catiline, On Pompey's Command, For the Poet Archias*; Vergil's* Aeneid*, Books I–VI; Arranged Alphabetically and in the Order of Occurrence*. New York. https://doi.org/10.2307/4385719

Lu, Xiaofei. 2012. "The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives." *Modern Language Journal* 96.2: 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Malvern, David, Brian Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical diversity and language development: quantification and assessment*. New York: Palgrave Macmillan. https://doi.org/10.1057/9780230511804

McCarthy, Philip M., and Scott Jarvis. 2007. "Vocd: a theoretical and empirical evaluation." *Language Testing* 24.4: 59–88. https://doi.org/10.1177/0265532207080767

Muccigrosso, John. 2004. "Frequent Vocabulary in Latin Instruction." *Classical World* 97.4: 409–33. https://doi.org/10.2307/4352875

Nation, I. S. P. 2006. "How Large a Vocabulary is Needed for Reading and Listening? *The Canadian Modern Language Review* 63.1: 59–82. https://doi.org/10.3138/cmlr.63.1.59

_____. 2013. *Learning Vocabulary in Another Language*. 2nd ed. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139858656

Olimpi, Andrew. 2017. *Via Periculosa: A Latin Novella*. Dacula, GA: Comprehensible Classics Press.
Piazza, John. 2017. "Beginner Latin Novels: A General Overview." *Teaching Classical Languages* 8.2: 154–66.
Pike, L. 1979. *An Analysis of Alternative Item Formats for Testing English as a Foreign Language*. TOEFL Research Reports No. 2. Princeton: Educational Testing. https://doi.org/10.1002/j.2333-8504.1979.tb01174.x
Read, John. 2000. *Assessing Vocabulary*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511732942
Ritchie, Francis. 1884. *Fabulae Faciles: A First Latin Reader*. London: Rivingtons.
Schmitt, Norbert, Tom Cobb, Marlise Horst, and Diane Schmitt. 2015. "How Much Vocabulary is Needed to Use English? Replication of van Zeeland and Schmitt (2012), Nation (2006) and Cobb (2007)." *Language Teaching* 50.2: 212–26. https://doi.org/10.1017/S0261444815000075
Schmitt, Norbert, Xiangying Jiang, and William Grabe. 2011. "The Percentage of Words Known in a Text and Reading Comprehension." *Modern Language Journal* 95.1: 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x
Schoonen, Rob, Jan Hulstijn, and Bart Bossers. 1998. "Metacognitive and Language Specific Knowledge in Native and Foreign Language Reading Comprehension: An Empirical Study among Dutch Students in Grades 6, 8 and 10." *Language Learning* 48.1: 71–106. https://doi.org/10.1111/1467-9922.00033
Shrum, Judith L., and Eileen W. Glisan. 2015. *Teacher's Handbook: Contextualized Language Instruction*. 5th ed. Boston: Cengage Learning.
Staehr, L. S. 2008. "Vocabulary size and the skills of listening, reading and writing." *The Language Learning Journal* 36.2: 139-52. https://doi.org/10.1080/09571730802389975
To, Vinh, Si Fan, and Damon Thomas. 2013. "Lexical Density and Readability: A Case Study of English Textbooks." *Internet Journal of Language, Culture and Society* 37: 61–71.
Turk, C., and J. Kirkman. 1989. *Effective Writing: Improving Scientific, Technical, and Business Communication*. 2nd ed. London: E and F. N. Spon.
Ure, J. 1971. "Lexical density and register differentiation." In *Applications of Linguistics*, edited by G. E. Perren and J. L. M. Trim, 443–452. Cambridge: Cambridge University Press.
Vajjala, Sowmya, and Detmar Meurers. 2012. "On Improving the Accuracy of Readability Classification Using Insights from Second Language Acquisition." The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, June 3–8, 2012. Montreal: Association of Computational Linguistics. 163–173.
Van Zeeland, H. & N. Schmitt. 2013. "Lexical Coverage in L1 and L2 Listening Comprehension: The Same or Different from Reading Comprehension?" *Applied Linguistics* 34.4: 457-79. https://doi.org/10.1093/applin/ams074
Vanderpool, Emma, and Matthew Katsenes, eds. 2016. *Eutropius, Breviarium Historiae Romanae*. Monmouth College.
Venditti, E. 2021. "Using Comprehensible Input in the Latin Classroom to Enhance Language Proficiency." *Journal of Classics Teaching* 22.43: 22–28. https://doi.org/10.1017/S2058631021000039
Wallace, Daniel B. 2013. "Reading through the Greek New Testament." Daniel B. Wallace Blog. 29 Dec 2013.
Xia, Menglin, Ekaterina Kochmar, and Ted Briscoe. 2016. "Text Readability Assessment for Second Language Learners." *Proceedings of the 11th Workshop on Innovative*

Use of NLP for Building Educational Applications. San Diego. 16 June 2016. Association for Computational Linguistics. 12–22. https://doi.org/10.18653/v1/W16-0502

Yu, Guoxing. 2010. "Lexical Diversity in Writing and Speaking Task Performances." *Applied Linguistics* 31.2: 236–59. https://doi.org/10.1093/applin/amp024

Zamanian, Mostafa, and Pooneh Heydari. 2012. "Readability of Texts: State of the Art." *Theory and Practice in Language Studies* 2.1: 43–53. https://doi.org/10.4304/tpls.2.1.43-53