This is the version of an article accepted at *Reading Research Quarterly* on January 10, 2017. The manuscript is scheduled to be published in the *53*(1) issue of the journal. Please note that this is neither the copy-edited nor published version.

Core Vocabulary: Its Morphological Content and Presence in Exemplar Texts

Elfrieda H. Hiebert

TextProject & University of California, Santa Cruz

Amanda P. Goodwin

Vanderbilt University

Gina N. Cervetti

University of Michigan

Abstract

This study addresses the distribution of words in texts at different points of schooling. The first aim was to identify a core vocabulary that accounts for the majority of the words in texts through the lens of morphological families. Results showed that 2,451 morphological families, averaging 4.61 members, make up the core vocabulary of school texts. The 11,298 words in the 2,451 morphological families account for 58% of the approximately 19,500 most frequent words in written English. The majority of the morphological families appear by the end of the elementary school period (85%), but a small group of morphological families (15%) is added through the middle to high school period. Analyses of the ranks of words across grade bands indicated that late-appearing words gain in prominence in higher level texts as some elementary-level words become less frequent. The second aim of the study was to determine the degree to which the core vocabulary accounted for the words in an independent but critical set of texts: the exemplar texts identified within the Common Core State Standards. The 2,451 families accounted for 97.1% (grades K and 1) to 89.1% (grade 11 through college) of the total words in texts and 95.6% (grades K and 1) to 74.9% (grade 11 through college) of the unique words in texts. Implications of the findings on the nature and role of the core vocabulary in complex texts are suggested for researchers, curriculum developers, and publishers.

Core Vocabulary: Its Morphological Content and Presence in Exemplar Texts

Vocabulary is unarguably a critical contributor to the comprehension of texts (Rickets, Nation, & Bishop, 2007; Sénéchal, Ouellette, & Rodney, 2006; Thorndike, 1917). When readers do not know the meanings of words that represent key ideas, successful comprehension can be compromised. The question remains, though, as to precisely which words students need to know to ensure that their vocabularies are equal to the task of comprehending the texts they encounter in college, communities, and careers. This question is especially important when considering the substantial diversity of vocabulary in school texts. Zeno, Ivens, Millard, and Duvvuri (1995) identified 154,941 unique words in their analysis of 17.25 million words in school texts from Grade 1 through college. School vocabulary instruction does not—indeed, could not—purport to comprehensively cover the English lexicon. But presumably, school instruction could support students in having a sufficient grasp of some key portion of the lexicon to successfully negotiate texts they read in workplaces and communities. The challenge is establishing what part of the lexicon to teach at what time. Because the English written lexicon has a high level of redundancy, we argue that identifying a core vocabulary based on frequency and morphological relations can deepen understanding and provide instructional guidance regarding the word demands students face when reading school texts at different grade levels.

When we examine theory and research, the literature is remarkably silent about which words and which lexical features merit instructional attention at different points in literacy development. Words on vocabulary assessments, including such long-standing measures as the Peabody Picture Vocabulary Test (PPVT), are chosen for their psychometric qualities to ensure that performances of students are distributed on a normal curve (Pearson, Hiebert, & Kamil,

2007), rather than for their representation of particular word features or frequencies. In the core reading programs that teachers often rely upon for teaching vocabulary, the 8 to 10 new words selected for instruction along with each reading passage are often chosen without empirical or conceptual basis (Graves et al., 2014), as Gates (1962) found with reading programs in the late 1950s and Stallman et al. (1989) confirmed in the late 1980s.

The most popular perspective on word selection in teacher education textbooks and policy documents, including the Common Core State Standards (CCSS; NGACBP & CCSSO, 2010) and the National Assessment of Educational Progress (NAEP; National Center for Education Statistics, 2012), is the three-tier model (Beck, McKeown, & Kucan, 2013). In this model, the English lexicon is viewed from the perspective of three tiers loosely related to frequency: (a) words of everyday speech, (b) general academic words and/or subtler or more precise synonyms of common words, and (c) technical vocabulary. The authors of this model suggest that the focus of instruction should be on the second tier. The tiers, though, are notoriously difficult to differentiate. For example, of 13 words identified by CCSS writers in Appendix A of the Standards as exemplifying tier-two words in Grade 4-5 texts, 11 are among the 2,500 most-frequent words in written English (e.g., *early, poured*)—words that, presumably, middle-grade students would know. Only two of the 13 words are words that could be described as general academic words or more precise synonyms of common words—*eruption* and *spouted.* This model provides a heuristic for teachers but does not provide either theoretical or empirical documentation on what lexical knowledge is critical for students to learn at particular developmental levels or whether the volume of vocabulary addressed in the tiered approach matches the task that students face in reading school texts.

Another framework, offered by Biemiller (2010), uses an existing database (Dale & O'Rourke, 1981) to identify the "words worth teaching." This framework uses summary data based on the meanings of words recognized by North American students to generate a list of around 11,000 word meanings, 43% of which are identified as requiring instruction in the elementary grades. The list does not, however, provide information on how central (i.e., frequent) words are in the lexicons of texts at different levels. For example, words such as *faucet* and *snout* that are predicted to appear twice in Zeno et al.'s 800,000 word sample of second-grade texts are identified as worth teaching in second grade but words predicted to appear often in second-grade texts (140 times or more)—*angry* and *build*—are relegated to Biemiller's easy category (i.e., not worth teaching). The worthiness of a word for teaching in Biemiller's scheme is not informed by the frequency with which students are likely to encounter a word in text or other factors such as the importance of the word in conveying meaning within a text.

These descriptions show that current models of vocabulary pedagogy fail to consider the manner in which instructional recommendations address the overall lexicon or even particularly salient parts of the lexicon. One study that did consider the overall lexicon was that of Nagy and Anderson (1984), which estimated that approximately 415,000 words make up written English (Nagy & Anderson, 1984, Table 4). These words, Nagy and Anderson demonstrated, can be clustered into approximately 88,500 morphological families. The conclusion of that study was that, due to the extremely large volume of words in texts, students need instruction that moves beyond direct teaching of individual words to experiences that promote the use of morphology and context when reading. Even 88,500 families represent too large a corpus to be instructionally relevant. Scholars have not considered, let alone determined, which of these

families should be emphasized in students' development. Neither has the relative weight of different families in texts been considered, particularly across what the CCSS (NGACBP & CCSSO, 2010) calls the staircase of text complexity, which describes the expected levels of text students need to read at particular grade levels if they are to achieve the literacy levels required for participation in the workplaces and communities of the 21st century by high-school graduation.

A lack of understanding of the size of the vocabulary required for success with school texts persists, even though there is both the capability and knowledge to address the nature of the lexicon in school texts in a way that was not possible previously, even at the time of the Nagy and Anderson (1984) analysis. A vast amount of knowledge about the lexicon has been generated as a result of digitization; however, much of this work exists in databases and in journal articles within disciplines disconnected from educational research. Digitization also means that samples of texts, such as those exemplifying learning tasks at different grade levels, can be scanned and analyzed with relative ease to determine the nature of the lexicon of school texts at different developmental levels.

The purpose of this paper is to draw on available information and digital tools to identify a core vocabulary—the portion of the English lexicon that accounts for the majority of the words in the texts that students are expected to read at particular points in schooling. Similar to Nagy and Anderson (1984), we take into account word relatedness and frame this core vocabulary within morphological families. But the changes in digital capacity in the last decade permit us to extend Nagy and Anderson's study in a number of ways.

First, we included all of the words in the portion of the lexicon predicted to be most prominent in written texts. By contrast, Nagy and Anderson (1984) randomly sampled 7,260

words of the *Word Frequency* database of 86,741 (Carroll, Davies, & Richman, 1971) to draw conclusions about the entire lexicon. Second, we address a limitation that Nagy and Anderson recognized in their study: The lack of information as to when particular word families are predicted to occur in students' texts. As Nagy and Anderson (1984) noted: "To get an accurate picture of the vocabulary that students actually encounter in printed school materials would require…a reanalysis of our data by grade level" (p. 322). The present study identifies the word families as a function of their presence at different grade levels. Third, the database available to Nagy and Anderson was restricted to Grades 3-9, leaving their conclusions confined to a specific grade span. By contrast, the database used in the current study spans the vocabulary in texts from 1st grade to college. A fourth extension of this work is the inclusion of a proof of concept analysis (Levac, Colquhoun, & O'Brien, 2010; Yang, 2005) where the word family database is applied to a distinct and separate set of texts from those on which the database was derived: the texts identified by CCSS developers (NGACBP & CCSSO, 2010) as exemplifying complex texts at different grade bands.

In summary, this study aims to establish a sweet spot, an optimal zone, in the distribution of word frequency such that we can understand what words students are likely to confront in different levels of school texts. Armed with such information, we will be in a better position to create models of vocabulary learning, curriculum, and instruction.

Review of Research

Core Vocabulary in Texts

Inventories of the number and frequency of words in written American English have been conducted for almost a century, as evident in the studies summarized in Appendix A. Efforts to tabulate unique words in school texts in the United States began with Thorndike's

(1921) analyses of almost 5 million words in texts that were deemed appropriate for school instruction in the early 20th century (e.g., the Bible and texts on farming and sewing). Around 125 word lists for use in schools were derived from Thorndike's original list and its two expanded versions (Thorndike, 1932; Thorndike & Lorge, 1944) over a 50-year period (Johnson, Smith, & Jensen, 1972). For example, Dale and Chall (1948) generated a list of approximately 3,000 words on which they based their readability formula for middle-grade texts, and Spache (1953) identified approximately 1,000 words to serve as the basis for a readability formula for the primary grades. Data on word frequency from Thorndike's analyses were claimed to be the basis for the basal reading programs that are often referred to as "Dick-and-Jane texts"—a textbook style that was prominent for a substantial portion of the 20th century (Hiebert & Raphael, 1996).

In the 1960s, scholars began using computers to count and characterize the words in large sets of texts. The first effort—that of Kucera and Francis (1967)—consisted of a one-million-word corpus from texts aimed at adults in 1961. Shortly thereafter, Carroll et al. (1971) analyzed school texts from Grades 3–9. Carroll et al. introduced the *U* function, which predicts the frequency of a word's appearance per million words in texts. Zeno et al. (1995) extended this work, including the use of the *U* function, in an analysis of texts from beginning reading through college levels.

Zeno et al.'s (1995) analysis was the last systematic effort to describe the distribution of vocabulary in school texts, but it is by no means the last to rank the vocabulary in large corpora of text. The Internet has made many more texts available to researchers, and several databases based on as many as a half-billion words have been produced (see Appendix A). Digitization has also meant that numerous analytic techniques have been applied in which vocabulary is

tagged according to various features such as age of acquisition or concreteness (e.g., Brysbaert, Warriner, & Kuperman, 2014; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012).

Reading researchers have used words from some of these word-frequency lists—most frequently that of Dale and O'Rourke (1981)—in their investigations (Scott, Lubliner, & Hiebert, 2006), but large corpora and new techniques for coding word features have remained largely outside the purview of reading researchers, developers and publishers of curriculum, and teachers. A review of textbooks used in teacher education courses showed little attention to or interpretation of the information from the various corpora analyses identified in Appendix A (Graves, Juel, Graves, & Dewitz, 2010; Roe, Smith, & Burns, 2011). For example, Roe et al. (2011) devote three of 13 chapters to word recognition and vocabulary. This material emphasizes the need for students to develop a sight vocabulary such as the words on the Dolch list (Dolch, 1936) or Fry Instant Words (Fry, 1980). Morphological components are discussed but nothing is said about the size of the entire lexicon or which groups of words beyond the Dolch and Fry lists might be important at different developmental stages.

Large databases of words are mainly used in one prominent area of reading practice and policy—systems that assign readability scores, or complexity levels, to texts. In current text complexity systems—such as Advantage/TASA Open Standard (ATOS; Milone, 2009) and the Lexile Framework (MetaMetrics, 2000)—vocabulary complexity is based on the ranks of the words in a text according to their frequency in English. The rankings are derived from large, digitized databases proprietorial to the system's owner.

One area in which word-frequency databases have been used in the creation of curricula and texts is in English as a Second Language (ESL) programs that are geared to young adults learning English, often as preparation for attending English-speaking colleges (see, e.g.,

Hazenberg & Hulstun, 1996). The General Service List (GSL) has been used extensively in ESL programs since West (1953) developed this list, including in the creation of leveled texts for use with adult students (Hill, 2008). More recently, Nation and his colleagues (Hirsch & Nation, 1992; Nation, 2013; Nation & Waring, 1997) have used the Kucera-Francis (1967) corpus to describe the English learning task for ESL learners. According to Nation and Waring (1997), the first 6,000 unique words account for 89.9% of all words in the Kucera-Francis corpus. A similar analysis has not been conducted with texts of different grade levels used for instruction of schoolchildren.

With that said, a series of studies that have used digitized corpora and digital tools to analyze texts suggest that a core vocabulary may be identifiable. The initial study in this line of work was directed at establishing the prominent vocabulary in third-grade assessments (Hiebert, 2002). In an effort to establish clarity about what words students must know to be successful on third-grade assessments, Hiebert examined the vocabulary in third-grade texts of three types of assessments: (a) norm-referenced tests, (b) state tests, and (c) oral reading assessments. The texts were analyzed to establish the number of complex words per 100 words of text. Complex words were defined as those that were multi-syllabic or did not appear among the first 1,000 words on Zeno et al.'s (1995) *Educator's Word Frequency Guide* (*EWFG*). On average, the six assessments in the sample had 4.8 complex words per 100. The state assessments were the most challenging (6 complex words per 100) and the norm-referenced tests were the least challenging (3.5 complex words per 100).

The use of single-syllable words as a criterion for word complexity in the Hiebert (2002) study lacks nuance in describing students' familiarity and exposure to words. For example, third graders' experience with single-syllable words such as *juice* and *sneeze* likely differs from their

knowledge of single-syllable words such as *quay* and *hue*. Hence, in a subsequent study, Hiebert (2005) aimed to determine the vocabulary that accounted for 90% of the tokens (total words) in fourth-grade assessments of three states (Florida, New York, Texas) and the NAEP. Fourth grade was chosen because this grade-level is the first assessed by the NAEP. The 90% level was chosen because leading literacy scholars (Clay, 1991; Stahl & Heubach, 2005) have proposed this level of word recognition as sufficient for comprehension. The database consisted of the *EWFG* (Zeno et al., 1995), which was organized into seven zones based on the predicted appearances of a word in a million words of text. These zones ranged from the 107 words that are predicted to occur 1,000 or more times per million to the 135,473 words that are predicted to occur fewer than once per million. Ninety percent of the tokens on all assessments were accounted for by the words in zones 1 through 5: words that are predicted to occur from 1,000 or more times per million (Zone 1) to 10 appearances per million (Zone 5).

The next study in this line of work (Hiebert, 2013) consisted of an evaluation of the excerpts from 168 exemplar texts in Appendix B of the CCSS (NGACBP & CCSSO, 2010). Results showed that the corpus used in the 2005 study (Hiebert, 2005)—the 5,586 words that are predicted to occur 10 or more times per million words of text—accounted for 92.5% of the words in texts in Grades 2-3 and 88% of the words in texts at the Grades 11-College and Career Ready (CCR) level.

The consistency of the patterns reported in Hiebert's (2005, 2013) studies and Nation and Waring's (1997) analysis of the Kucera-Francis corpus suggests that further examination of a core vocabulary in texts across grade levels is warranted. A more intensive examination of the exemplar texts identified by CCSS writers (NGACBP & CCSSO, 2010) also merits attention. The corpus in the Hiebert (2013) study, which relied on the excerpts provided within Appendix

B of the Standards, consisted of approximately 80,000 words—a relatively small corpus of text. A more extensive and comprehensive sample of texts, including texts for Grades K-1, is required to make conclusive statements about the role of the core vocabulary in the CCSS exemplar texts at different grade levels. Furthermore, as the following discussion shows, a view of morphological families calls for a more in-depth examination of the core vocabulary.

Morphological Learning

English has a lexicon that, because of its history, is larger than that of most languages (Mugglestone, 2013); however, the two languages that have contributed heavily to the lexicon of English—German and French—are both morphological, or made up of meaningful word parts (Venezky, 1999). When words are categorized according to shared morphemes, the size of the lexicon decreases substantially. In their study of the size of the school lexicon, Nagy and Anderson (1984) called for considering—and teaching—words as morphological families. When reflecting on the concept of a 'word', they write "absolute vocabulary can only be discussed in terms of some theory of relatedness among words" (p. 306). Considering morphological families, they argued, takes into account the fact that knowledge of one member of a family very likely provides leverage in figuring out the meanings of other family members, at least those with relatively transparent semantic relationships to the known word.

Since Nagy and Anderson's (1984) conclusion that morphological connections are critical for students to negotiate the many rare words present in text, the research literature on students' ability to use morphological relationships and also to become more adept in using this knowledge through instruction has grown. The ability to infer the meanings of words that are morphologically related to known words has been used to explain the dramatic annual increases in children's word knowledge—increases estimated by Anglin (1993) at 20 words per day

between grade 1 and grade 5, far exceeding the number of words instructed in school. According to Anglin, the impressive growth in students' knowledge of derived words is attributable in part to their ability to infer word meanings through morphological problem solving. Similarly, Carlisle and Stone (2005) showed that second through sixth graders read morphologically complex words (*winner*) more accurately than matched morphologically simple words (*dinner*) of the same length, frequency, and spelling. Additional evidence of a facilitative effect of morphologically organized words comes from studies that have demonstrated that students have faster and more accurate recognition of words after they have been exposed to these words' morphological relatives (e.g., McCutchen, Logan, & Biangardi-Orpe, 2009; Rabin & Deacon, 2008).

Even young children are able to recognize and manipulate relatively transparent relationships among some morphologically complex words in order to determine word meanings. For example, Jones (1991) found that first-grade children understood the relationships between inflected words and their base morphemes. The children were able to delete morphemes from inflected and compound words to create new words and provide semantic information about the new words (e.g., *plants* to *plant* and *windshield* to *wind*).

Furthermore, a series of reviews in the last decade confirm the efficacy of instruction in supporting students' morphological knowledge and problem solving (Bowers, Kirby, & Deacon, 2010; Carlisle, 2010; Goodwin & Ahn, 2010; Reed, 2008). All reviewers describe the benefits of integrating morphological instruction with other strategies, such as using context, supporting Nagy and Anderson's (1984) focus on the application of morphological knowledge within the context of reading.

When considering morphological families as an organizing principle within a core vocabulary, frequency of root words as well as derived words becomes critical. All of the above studies rely on students being able to apply knowledge of at least one family member to deciphering new words. Students are more likely to know or have been exposed to higher-frequency root words or family members. For example, Carlisle and Katz (2006) showed that, for fourth and sixth graders, root word frequency, average family frequency, derived word frequency and family size contributed to being able to read morphologically complex words. Goodwin and colleagues (Goodwin, Gilbert, & Cho, 2013; Goodwin, Gilbert, Cho, & Kearns, 2014) have shown similar supports of root word and derived word frequency for adolescents in reading and building lexical representations. These findings support the framing of a core vocabulary within the context of frequent morphological units, related words, and families.

The Present Study

This study aims to determine the size of the core vocabulary when viewed from the perspective of morphological families rather than individual words. We begin by examining the core vocabulary that, in an exploratory study (Hiebert, 2013), accounted for a sizable portion of the vocabularies in excerpts of the exemplar texts across the CCSS staircase of text complexity. A particular interest in the current study is in determining at which developmental levels words within the morphological families become prominent. The second major aim of the present study is to apply the database of the core vocabulary of word families to a unique database—that is, a set of books that were not used to establish the original database of word families. The set of texts chosen for this proof of concept analysis (Levac et al., 2010; Wang, 2005) consist of the CCSS exemplar texts, which have been offered by CCSS writers as illustrating the kinds of

texts that students need to read if they are to attain literacy levels necessary for success in college and careers (NGACBP & CCSSO, 2010).

These goals translate into the following specific questions:

Morphological Families Within Core Vocabulary

1. How many morphological families are present among words predicted to have at least 10 appearances per million words of written English?

    1.a. How many additional members of lead words in the core vocabulary are among words with predicted appearances of 1 to 9 per million words?

    1.b. How are morphological families distributed across grade bands?

Core Vocabulary In CCSS Exemplar Texts

2. What percentages of the total and unique words within the CCSS exemplar texts at different grade bands are accounted for by the core vocabulary?

Method

In analyzing texts and corpora, numerous choices need to be made. We describe choices made in this study according to the two fundamental research foci: (a) establishing the size of the core vocabulary when viewed as morphological families and (b) establishing the number of total and unique words accounted for by the core vocabulary in the CCSS exemplar texts.

Establishing Size of Morphological Families in a Core Vocabulary

Selection of word corpus. The *EWFG* (Zeno et al., 1995) was chosen as the corpus for examining the presence of a core vocabulary and related morphological family members for a number of reasons. First, this database is the most recent analysis of school texts and is more comprehensive, covering Grades 1 to 13, than the only other available analysis of school texts—Carroll et al.'s (1971)—that sampled texts from Grades 3 to 9. Second, the *EWFG* provides

grade-level appearances of the 19,469 most-frequent words, permitting an investigation into when words can be expected to become prominent in school texts. Third, Brysbaert and New (2009) have concluded that a corpus of 1 to 3 million words is sufficient to obtain a reliable estimate of high-frequency words. The *EWFG* meets this criterion with 154,941 types (unique words) based on 17,272,580 tokens (total words).[1]

Criteria for inclusion in a core vocabulary. The *U*-function metric (a prediction of the number of appearances of a word per million words of text) is an indication of the prominence of a word in written English and the likelihood that students will encounter a word by particular points in their reading development. We chose 10 appearances per million words of text as the minimum criterion for inclusion in the core vocabulary primarily because this group of words had accounted for a majority of the total words in a preliminary study (Hiebert, 2013). We label this criterion as *U*10+, which refers to 10 or more predicted appearances per million words of text.

The 5,586 words with *U*10+ in the *EWFG* (*U*10+; Zeno et al., 1995) formed the database for identifying lead words and the initial morphological families. As is typical in the identification of vocabulary, proper names and individual letters (except *a* and *I*, which function as words) were eliminated from the database (a total of 312 words). The sample that became the basis for generation of morphological families consisted of 5,275[2] words.

---

[1] A comparison of placement of words in the *EWFG* and five additional corpora was conducted, validating Brysbaert and New's claim of similar rankings of high-frequency words across databases. This analysis can be found in the on-line materials (Appendix B).

[2] The word *miss* was added to the database because it became evident from numerous applications of the *EWFG*, interactions with colleagues using the *EWFG,* and examinations of databases such as the COCA, BNC, and Kucera-Francis that this word appears with a frequency of 10 or more times per million words.

The representative of a family—heretofore called the *lead word*—was the word with the highest *U* function within the *U*10+ group. We began to use the term "lead" rather than "head" (the conventional term used to describe root words) when our analyses made it evident that the most prominent member of at least a portion of the morphological families was an inflected, affixed, or compounded form, not the root or head word.

The source for categorizing the 5,275 words into morphological families was the database generated by Becker, Dixon, and Anderson-Inman (1980). Becker et al. assigned a root word to each word within a corpus of 25,782 words taken from Thorndike and Lorge (1943); they defined a root word as the smallest word from which a given word can be semantically derived. For example, *arrange* was judged to be the root word for *arranged, arrangement, arranging, disarrange, prearrange, rearrange, rearranged, rearrangement,* and *rearranging.*

We departed from the procedures used by Becker et al. (1980) in the coding of compound words. Becker et al.'s decision to treat compound words as distinct root words fails to recognize the semantic and morphological connections among words. Nagy and Anderson (1984), who conducted their analysis after that of Becker et al., placed compound words with root words. Consequently, we developed guidelines for placing compound words within morphological families rather than treating them as unique words.

The interest of the present analysis was in endocentric compound words, where the two most typical patterns are: (a) modifier is either descriptive of the head word (e.g., *blackboard,* a type of board that is black) or (b) determinative of the head word (e.g., *playground,* a ground where games are played). The decision was made to place a compound word with the primary word (e.g., *basketball* with *ball* rather than *basket*), except for cases of pronouns with the word *self*, which were placed with the pronoun (e.g., *yourself/you).* Exocentric compounds, because

of their idiosyncratic meanings, were not included in morphological families. For example, in the exocentric compound *roughhousing*, the head word does not describe a type of house but is a metaphorical referent to a type of activity that may or may not occur in a house.

Nagy and Anderson (1984) identified two types of semantic relatedness between a derived word and its root word: (a) semantically transparent, where readers should be able to infer a word's meaning immediately or with reasonable textual context, based on knowledge of the root word (e.g., *misrepresent/represent*), and (b) semantically opaque, where readers require either substantial textual context to connect a derived word with its root word or where the connection is not readily discernible (*condescend*/*descend).* An analysis to determine the degree to which the Becker et al. designations are semantically transparent was conducted, the results of which can be found in accompanying on-line materials (Appendix C). This analysis suggests that Becker et al.'s coding for derivative words follows a similar pattern as that of Nagy and Anderson (1984) with approximately one of every 10 family members having a semantically opaque relationship to the lead word.

Identifying additional members of the core vocabulary morphological families. In choosing the criterion of *U*10+ as the minimum requirement for identifying the core vocabulary, we were aware that many of the words in this group would have less frequent family members that, when combined, could add to the likelihood that students will encounter one or more family members in school texts. For example, the word *snow* has a predicted appearance of 140 appearances per million. None of its inflected or derivative forms appear within the *U*10+ group but four forms are evident in the portion of the corpus with predicted appearances of 1 to 9 times per million words: *snowed* (1), *snowing* (2), *snows* (4), and *snowy* (8). All but one of these words is an inflected ending, meaning that these family members are highly transparent.

Further, the likelihood that students will see these words in texts is high in that the four inflected and derived forms combine to account for 15 predicted appearances per million words.

To provide a comprehensive representation of the size of a morphological family, we extended our analysis to include family members among words predicted to occur one to nine times per million words of text—a group that we refer to as $1 \leq U \leq 9$. This group of words consists of 13,882 words within the *EWFG* (Zeno et al., 1995). The aim of this analysis was to identify family members for the original group of lead words, not to account for new lead words and their morphological families in the $1 \leq U \leq 9$ group. For example, the words *miracle* and *miracles* appear in the ($1 \leq U \leq 9$) group but this family was not added to the core vocabulary.

The assignment of words to morphological families occurred in three stages. First, the principal investigator coded all of the words within the *U*10+ group. Next, the second investigator coded 700 of the *U*10+ sample of the words. The inter-rater agreement rating between investigators was 95.7%. In the third stage of coding, the principal investigator identified the family members of the original lead words within the $1 \leq U \leq 9$ group. A research associate with substantial experience in conducting corpora analyses independently coded 10% of this group with an inter-rater level of 90.3%.

Most discrepancies in coding pertained to compound words, which were more prominent in the coding of the $1 \leq U \leq 9$ group since compound words increase as frequency rankings decrease (Nagy & Anderson, 1984).A second discrepancy in the coding of the $1 \leq U \leq 9$ group was the failure of a coder to exhaustively include all family members, especially words with prefixes and suffixes. For example, changes in spelling from a lead word to derivatives mean that searches using a lead word (e.g., *reverse*) could miss particular derivatives (e.g., *reversible,*

irreversible). All disagreements were discussed between coders. Once consensus was reached, agreed-upon adjustments were made to the database and guidelines for scoring were elaborated.

Identification of core vocabulary at different developmental levels of text. The *EWFG* (Zeno et al., 1995) provides data on appearances of words at individual grades from 1 through 13. These data were the basis for determining grade bands at which lead words become prominent. We clustered the appearances of lead words and family members according to the six grade bands that make up the CCSS staircase of text complexity (NGACBP & CCSSO, 2010: K-1[3], 2-3, 4-5, 6-8, 9-10, and 11-CCR. An illustration of the number of predicted appearances at different grade levels for members of a word family is given in Table 1.

The grade band at which a family of words becomes prominent was established by computing the percentage of a word family's appearances as a function of the total words for each grade band in the *EWFG* database. The criterion for prominence within a grade band corpus was the same as the one used for including words within the overall corpus: 10 appearances per million words. As illustrated in Table 1, the word family *improve* attains this level at the Grades 4-5 grade band.

Multilevel models to explore changes in rank. The sums of the total appearances of families at grade bands were transformed into ranks. We then used multilevel modeling to determine whether the ranks at each grade band changed as texts became more complex (i.e., grade bands got higher) and whether these variations depend on the word's overall frequency. For these analyses, each word's rank at each grade level band was considered as multiple observations for each word nested within time. First, a two-level unconditional growth model

[3] We describe the first grade band as K-1 to maintain consistency in describing the developmental levels of words in the *EWFG* and in analyzing the core vocabulary in the CCSS exemplars, even though the first level in the *EWFG* was Grade 1 only.

was used to answer the question of whether ranks varied across grade bands. Next, the predictor frequency was used to predict the intercept and slope to determine how variation in rank related to a word's overall frequency. A random effect for the intercept and slope was also included due to the variability present in the data. Grade bands were considered time points with the data centered at the fourth to fifth grade band, meaning that rank at the intercept would be interpreted as the rank of the word within fourth and fifth grade texts. In terms of the interaction with frequency, we used case 3 (due to our cross-level interaction γ_{11}) within Preacher, Curran, and Bauer's (2006) online tool to determine regions of significance.

Identifying Exemplar Texts

The CCSS exemplar texts. The proof of concept analysis that applied the word family database to a new set of texts was conducted with the 200 exemplars identified by writers of the CCSS in the standards and in Appendix B (NGACBP & CCSSO, 2010). CCSS developers described these texts as exemplifying the level of complexity and quality that students need to attain at particular points across their school careers to ensure attainment of reading levels associated with CCR at high school graduation.

The process for identifying the CCSS exemplar texts is described in Appendix B of the Standards. First, the CCSS work group asked teachers, educational leaders, and researchers to suggest texts with the level of complexity and quality of texts appropriate for different grade levels. From the nominations, the work group selected classic or historically significant texts and contemporary works of comparable literary merit, cultural significance, and rich content. Finally, the work group vetted the selections according to the quantitative and qualitative recommendations in Appendix A of the Standards and other criteria that included subject matter, publication date, and authorship.

Copies of all 200 CCSS exemplar texts were obtained. Numbers of texts and words at each grade band in the analysis appear in Table 2. When texts had 40 or fewer pages, the entire text was scanned. For texts longer than 40 pages, 10% of the text from the middle portion of the text was scanned. The principal investigator conducted an analysis of seven texts chosen randomly from the entire set of exemplar texts longer than 40 pages to determine whether a sample from the middle of a text produced similar outcomes as samples from the first third and last third of a text. All analyses produced similar outcomes, confirming the choice of sampling from the middle of texts that were longer than 40 pages. Two independent reviewers checked the accuracy of the Optical Character Recognition (OCR) process for all scanned texts.

Analytic scheme for vocabulary in the CCSS exemplar texts. A unique text analysis program was developed for this study to establish the degree to which the word family database (i.e., the 2,451 word families and their members) accounted for the total and unique words in the CCSS exemplar texts. The digital program provided two forms of output: (a) the percentage of tokens and types in the CCSS exemplar texts accounted for by the $U10+$ portion of the word family database and (b) a similar analysis for the $1 \leq U \leq 9$ portion of the word family database. Each CCSS exemplar text was analyzed separately. Data on CCSS exemplar texts were averaged for each text type (i.e., narrative, informational) and the two text types combined at each of the six grade bands.

Results

Establishing the Size of Morphological Families in a Core Vocabulary

1. How many morphological families are present among words predicted to have at least 10 appearances per million words of written English?

Our analysis identified 2,451 lead words within the group of words with frequencies of *U*10+ in the *EWFG* corpus (Zeno et al., 1995). Each family had an average of 2.15 members. Of the 2,451 families, 42.3% had one member (i.e., the lead word), 50% had 1 to 3 members in addition to the lead word, and 7.7% had 4 to 13 members and the lead word.

The final row in Table 3 summarizes the predicted appearances of the lead words and family members in a corpus of one million words of texts: 62.5% of a million-word corpus for the former and an additional 19% for the family members.

1.a. How many additional members of the 2,451 morphological families are present among words with predicted appearances of 1 to 9 per million words?

As summarized in the final row of Table 3, 6,023 members of the original 2,451 morphological families were identified within words with frequencies of $1 \leq U \leq 9$ within the EWFG (Zeno et al., 1995). With the addition of these words, the size of families increases by an average of 2.46 members. The portion of the 2,451 morphological families that continue to consist only of the lead word decreases from 42.3% o 7.5%. This group includes words such as *salmon*, where the singular and plural forms are the same, *much*, which refers to a generalized quantity, and function words such as *with* and *at*. The majority of families—54%—have 1 to 3 members in addition to the lead word. Typical of this group is the word *map* with the plural and inflected forms of *maps, mapping,* and *mapped*. The number of families with 4 or more family members in addition to the lead word rises to 39% with the addition of family members from the $1 \leq U \leq 9$ group.

The information in the bottom line of Table 3 indicates that, while the number of family members from the $1 \leq U \leq 9$ group (n = 6,023) is greater in number than the group from which the morphological families were originally identified (n = 5,275), the addition of these family

members increases the predicted number of appearances of a morphological family by only 1.9%. The volume of words that belong to the 2,451 families, however, is substantial. When viewed from the perspective of the 19,469 most-frequent words in written English (i.e., those with $U \geq 1$), 58% are members of the 2,451 morphological families.

1.b. How are morphological families distributed across grade bands?

Table 3 summarizes the grade band data from the *EWFG* (Zeno et al., 1995) for the 2,451 morphological families. These data indicate that over half of the morphological families (53%) appear in texts at the Grades K-1 level. Most morphological families have appeared by the end of the elementary-school period (85%). Morphological families continue to be added to the core vocabulary during the middle- to high-school period, although the total percentage of the core vocabulary (15%) does not match the volume of morphological families that are present in elementary-school texts.

We conducted an additional analysis to consider whether, as word families such as *principles, regulations,* and *doctrine* become more prominent in the higher grades, word families that appear with frequency in the lower grades (e.g., *cat, luck, neat* in Grades K-1 and *tales, rough, herd* in Grades 2-3) decrease in frequency. Results of these analyses are shown in Table 4. Ranks of words varied significantly across the bands of grade-level texts with an average rank of 649.77 for the lead words within Grade 4-5 texts (see Model 1). How these ranks varied across time depended on the word's overall frequency value (see Model 2, *timeXfreq* = 0.44, SE = 0.004, $p < .001$). Overall, for each incremental higher-grade band (i.e., moving from the 4-5 grade-level band to the 6-8 grade-level band), the average rank for the lead words increases significantly by 105.06 ($p < .001$). Keeping in mind that a larger number indicates a lower rank (e.g., a word with a rank of 200 is less prevalent relative to other words in

a text than a word with a rank of 100), this indicates that, on average, particular words in the core vocabulary become less prevalent as texts became more complex. Thus, core words are most prevalent in texts at lowest end of the grade-level spectrum and less frequent words begin to displace them as texts become more complex in the higher grades. With that said, there is significant variability in the intercept and in the slope.

In terms of the interaction with frequency, results suggested that the interaction between grade-level band and frequency is significant for all words with a *U* function greater than 50.77, which was about 41% of the lead words in the sample (n = 1002). The pattern is graphically illustrated in Figure 1. For words with higher values (i.e., *U* functions of 150 or 300), rank increases (meaning less prevalent compared to other words) across the grade-level bands with words of higher frequency. In contrast, words with a *U* function of 50 (or less) tend to have similar ranks across grade-level bands. These results may suggest that the most frequent words become less prevalent in texts whereas words that are somewhat less frequent retain their overall prevalence as texts become increasingly complex. A significant variability in the intercept and slope continues, however, suggesting that these trends should be interpreted as average trends and that individual word trajectories can vary significantly.

Table 5 presents examples of this variation with six words (*little, much, a, abandoned, baked,* and *occasionally*). The ranks of each word are shown at each grade band as well as the difference between ranks in K–1 relative to Grades 11-CCR texts. What is clear is that, even for words of similar frequency, different trends are occurring.

Core Vocabulary in CCSS Exemplar Texts

2. What percentages of the total and unique words are accounted for by the core vocabulary within CCSS exemplar texts at different grade bands?

Our second research question addresses the degree to which the morphological families in the $U10+$ and $1 \leq U \leq 9$ groups account for the total and unique words in a new and unique set of texts—the exemplar texts of the CCSS. In discussing these results, we will use the terms *tokens* to refer to the total number of words in texts and *types* to describe the number of unique words in texts. Summary data on types and tokens are provided in Table 6.

Tokens. With the exception of the texts at the Grade 2-3 band, percentages of tokens accounted for by the 2,451 morphological families with predicted appearances of $U \geq 1$ per million of text are an average of 1.04% higher for narrative texts than informational ones. At the Grade 2-3 band, the percentage of tokens accounted for by the 2,451 families in the two text genres is quite similar.

Across the six grade levels of CCSS exemplars (see Table 6), the average percentage of tokens accounted for by the 2,451 morphological families is 91.5%. The percentage is particularly high for Grades K-1: 97.1. The percentage drops to around 92% at Grades 2-3 and, across the remaining grades, hovers close to 90% with approximately .7% fewer of the tokens in texts accounted for at each successive grade level. One grade band shows a somewhat different pattern: tokens for the Grades 6-8 band show a slightly larger drop from the previous grade band and are quite similar to the texts of the next grade band (9-10).

Types. The number of different words in a text is also a critical feature that can influence word learning (Endress & Hauser, 2011). In the texts at the Grades K-1 band, approximately 96% of the unique words come from the 2,451 word families. The percentage drops by 6.7% in texts at Grades 2-3 and then continues to decline by an average of 2.7% until Grades 11-CCR where the total percentage of types from the core vocabulary is 74.9%.

Discussion

In the decades since the shift away from texts with controlled vocabulary to authentic texts (with unknown vocabulary characteristics) in school reading programs (Fitzgerald, Elmore, Relyea-Kim, Hiebert, & Stenner, 2016), the relative frequencies of words have been used to index text complexity (e.g., MetaMetrics, 2000), but rarely have they been used to understand the task of learning words and their meanings within school texts. An exception is the work of Nagy and Anderson (1984), which estimated the number of words in written English based on morphological families. Implicit in their inquiry was the goal of finding a way to reduce the enormity of the task of "discovering" from context the meanings of the huge number of unknown words students will encounter in their school texts over their K-12 school careers. Like Nagy and Anderson, we focused on viewing the vocabulary demands of school texts from the perspective of morphological relatedness. Our work is also distinguished from that of Nagy and Anderson in a number of ways: (a) a focus on the presence of families within the words predicted to occur most frequently in school text, rather than to predict the number of morphological families within the entire lexicon of written English, (b) use of a database representing texts from grades 1 through college, rather than only grades 3-9, (c) establishing when across students' school careers these word families are predicted to become prominent, and (d) conducting a proof of concept study (Levac et al., 2010; Yang, 2005) in which the presence of the word families is examined in a unique set of texts—the CCSS exemplar texts.

The Size of the Core Vocabulary When Viewed as Morphological Families

Nagy and Anderson (1984) concluded that the ability to make morphological inferences coupled with proficiency in contextual analysis could go a long way in aiding students in dealing with the many words in the English lexicon. In their meta-analysis of 20 studies of

incidental word learning, Swanborn and de Glopper (1999) found that approximately 15% of the unknown words in texts were learned in context or incidentally. Among the variables that Swanborn and de Glopper identified as influencing whether a word is learned in context is the proportion of unknown to known words in a text. That is, a lower density of unknown words in a text produced a higher likelihood of learning the unknown words from context.

If morphological problem solving coupled with contextual analysis is to work for students as a way of dealing with the many unknown words in text, we assume that at least part of the lexicon needs to be known well. Several scholars have suggested ratios of known to unknown vocabulary required for students to comprehend texts (e.g., Betts, 1946; Clay, 1991; Stahl & Heubach, 2007), although considerable ambiguity and debate surrounds these estimates (e.g., Halladay, 2012). The current study does confirm the presence of a core group of words that is predicted to occur frequently in texts: 2,451 lead words with 4.61 members from the $U \geq 1$ group. At the present time, we simply do not know how facility with the meanings of the 2,451 morphological families affects students' ability to negotiate text. But we believe that the core vocabulary merits the attention of researchers to determine its role in proficient reading at different levels and the potential to support incidental acquisition of word knowledge during reading through morphological inferencing.

Having identified a core vocabulary in which the majority of words have relatively frequent relatives that students are likely to encounter in school makes it possible to conceive of vocabulary programs that capitalize upon and nurture students' problem solving abilities. Studies involving a range of different learners have demonstrated that providing students with instruction and practice using morphological components supports the ability to solve unknown words (Bowers, Kirby, & Deacon, 2010; Carlisle, 2010; Goodwin & Ahn, 2010; Reed, 2008).

Although studies of morphological instruction alone have not demonstrated effects on generalized comprehension (Wright & Cervetti, in press), a hypothesis to be tested is whether instruction of word families that occur frequently at particular developmental levels might be successful in improving students' generalized comprehension. Prior vocabulary studies have focused on a broad array of words, not on morphological families that students are likely to encounter in texts at particular grade levels. Choosing families of words that students are likely to see often at a particular grade level provides naturalistic opportunities for repeated exposure and for practice in applying their knowledge of these words to solving the meanings of new words. The identification of morphological families that are prominent at particular developmental levels, one of the contributions of the present study, could be useful in designing interventions that aim to increase students' comprehension.

We do want to caution, however, that the availability of a database of morphological families does not mean that a list should be widely distributed to students so they can memorize the lead words. Gaining mastery over the corpus will likely require a mix of activities—some explicit instruction, some creative word play, lots of wide reading, and many opportunities to use these words in writing activities. In addition, if we want students to leverage this knowledge for word-solving, instruction in multiple contextual and morphological problem-solving strategies may be needed. But these suggestions are nothing more than informed hunches—hunches that require solid, extensive pedagogical research prior to establishing best practice policies.

Core Vocabulary within the Different Grade Bands

The present analyses identified when words become prominent at different grade bands on the basis of predictions in an existing word database (Zeno et al., 1995). The database of

word families was then used for a proof of concept to determine whether these predictions held up with the CCSS exemplar texts. Neither of these analyses provides data on students' ability to recognize the words and their meanings. Biemiller (2005) showed that, at least for English-speaking students, the meanings of particular words are acquired in a predictable developmental sequence. The points at which students know words in different grade band vocabularies need to be established before the word family database is useful for instructional or intervention purposes. Variations in the vocabularies of a cohort of primary-level students' vocabularies are considerable, reflecting linguistic, economic, and social factors (Biemiller & Slonim, 2001). Determining what portions of a grade-level vocabulary require instruction for which students and how this instruction is best provided for students with varying needs should be a priority in future research on the core vocabulary. Furthermore, variations in the complexity of words for students within a grade band cohort can also be considerable. For example, factors such as concreteness and length influence the complexity of words (Balota, Yap, & Cortese, 2006).

Most morphological families of the 2,451 word families are predicted to have appeared in students' texts by the end of the elementary-school period (85%). But the core vocabulary is not stagnant across the grades. Over time, two types of changes can be seen in the core vocabulary. First, there continue to be additions to the core vocabulary at all grade levels, including Grades 11-CCR. The number of words added to the core vocabulary gets smaller with movement up the CCSS's staircase of text complexity but the words that are added are likely important additions to grade-level corpora. Morphological families added in the Grades 11-CCR include *discrimination* and *transmitted* where both the lead words and family members illustrate the layers of affixation that Nagy and Townsend (2012) attributed to complex academic vocabulary (e.g., *indiscriminately, transmission*).

We also hypothesize, from our experience with the 1≤*U*≤9 database, that there are additional morphological families within this group that, while not having a single member with predicted value of *U*10+, have substantial predictive value in high school texts. For example, words within the 1≤*U*≤9 group such as *calculate* with 6 additional family members and *modify* with 7 additional members have predicted appearances, as a group, of 143 and 125, respectively, in texts at Grades 11-CCR. These examples suggest that a productive line of future work would be to identify additional morphological families with lead words within the 1≤*U*≤9 corpus and to determine their prominence in texts over the middle to high-school period.

Another way in which the core vocabulary does not remain stagnant is the changing prominence of some words in the corpus at different grade bands. The analysis of the ranks of words at particular frequency levels provides insight into these changes. Results suggested that, on average, words within the core vocabulary decreased in prevalence (i.e., had a higher number rank where lower ranks meant more prevalence). This pattern lends support to the hypothesis that the core vocabulary is particularly dominant in the earlier years and that words with lower frequencies come into play at higher grade levels. The decrease in prevalence was most prominent for the most frequent words, suggesting that across time, words that were less frequent within the core vocabulary stayed similarly prevalent whereas the most frequent words decreased in prevalence. This suggests that less frequent core vocabulary words, many of which are academic in nature, take on a more important role with texts in higher grade bands.

Core Vocabulary in the CCSS Exemplar Texts

The grade-level analyses of the 2,451 word families were based on the predicted appearances in written English according to the *EWFG* (Zeno et al., 1995). According to Zeno et al., their database of 17,274,580 total words came from 60,527 text samples drawn from

6,333 texts. Our second interest in the present study lay in the degree to which this database of word families accounts for the total and unique words in an independent but important set of texts at different grade bands—those identified within the CCSS as exemplars of the appropriate complexity from Grades K through college. The analyses showed that the 2,451 lead words and family members with predicted frequencies of $U \geq 1$ accounted for an average of 91.5% of the words in the CCSS exemplar texts from Grades K to CCR. The drop of 5% in the percentage of core vocabulary from texts at Grades K-1 to Grades 2-3 was substantial but, even after that point, the core vocabulary continued to account for the vast majority of the words in texts.

The rhetoric surrounding complex texts and the challenges faced by many American students on the NAEP (National Center for Education Statistics, 2015) might lead to the expectation that differences in core vocabulary distributions between early primary texts and CCR texts would be even greater than was the case in the present analysis of the CCSS exemplars. We did see differences in how ranks of words within the core vocabulary changed across texts of different grade band levels with moderately frequent words beginning to take on higher ranks as students move through the grades. However, the core vocabulary remains prominent throughout the texts of schooling, suggesting that it may continue to provide leverage for meaning making in texts of higher grades.

The percentage of total words accounted for by the 2,451 morphological families remains high across the grade bands but the amount of lexical diversity increases substantially at higher grade bands. At Grade K-1, less than 5% of the unique words are not within the core vocabulary. By the middle-school grades, 20% of the unique words in texts are not part of the core vocabulary and, at high school, approximately 25% of unique words are not in this group.

The demands of vocabulary in texts can be seen to escalate substantially in middle school and beyond as the diversity of vocabulary increases.

Subsequent analyses are also needed to address qualitative differences in the rare vocabulary in texts across the grades. Table 7 has been provided to illustrate the nature of tokens and types at each of the six grade bands. Even a cursory examination of the rare words in Table 7 shows noticeable differences across the grades. The non-core words in the primary-grade examples pertain to familiar concepts (e.g., *ham*, *pumpkin).* In Grades 4-5 and 6-8, the non-core words may be easily explainable (e.g., *dimpled, grumbles*) but may not be immediately familiar to students, especially when reading independently. Across the high school to CCR period, rare words seem to become increasingly more abstract (e.g., *insular*).

Texts are also changing in ways other than vocabulary as students move through the grades. For one, texts get longer at higher grade levels. Especially for struggling readers, the presence of an additional rare word per 100 likely means something quite different in a 1,000-word text than in a 100-word text. Sentences in texts are also getting longer and more complex as students move across the grades. Longer sentences reflect the presence of clauses and phrases, which make demands on memory and comprehension processes (Just & Carpenter, 1980), in addition to those made by unfamiliar and often abstract concepts.

Limitations

Any project, but particularly one with the focus of the present study on a large database and also a large text corpus, involves numerous choices, all of which influence the outcomes. One such choice was to use the *U*10+ criterion as the cutoff point for inclusion in the database of lead words. How quickly a word becomes part of students' repertoires of known words is likely a complex interaction of word features (e.g., polysemy, orthographic transparency,

concreteness, interest) and text factors (e.g., ratio of known to unknown words, repetition of known words). The criterion of *U*10+ was chosen because words with this frequency and higher accounted for at least 90% of the words in elementary texts (Hiebert, 2005, 2013).

Another critical choice was to regard compound words as morphological variants of the lead words, which was usually the second root word in the compound word. We think the case is strong for our decision. At an intuitive level, a fishhook is more about hooks than fish, and a basketball more naturally falls into the ball than the basket bucket. But we have taken only a tiny step into a domain of scholarship that should become a particular focus of future studies that consider proficiency with the core vocabulary.

We also chose to exclude proper names from the 2,451 morphological families. This choice is typical in the identification of academic or school-based vocabulary (see, e.g., Anglin, 1993; Coxhead, 2000; Gardner & Davies, 2014; Nagy & Anderson, 1984). Proper names are not part of morphological families because most proper names do not have connections to semantic networks. For example, the name *Detroit* refers to "strait" in French but, in English, is not associated with that meaning. In designing instruction, of course, attention needs to be paid to proper names. Some texts can be quite dense with proper names, especially the magazine articles that are often the text types that dominate state and national assessments (Hiebert, 2005).

The choice to limit morphological members to those with $U \geq 1$ means that this database does not provide an exhaustive listing of all members of the 2,451 families, as illustrated by the identification of the word *unspeakable* in the Grades 9-10 excerpt in Table 7 as rare, even thought he word *speak* is part of the 2,451 word families The *EWFG* (Zeno et al., 1995) database has 124,403 words with *U* functions less than 1. A database of this size requires a

sampling procedure, which is typically done in studies of morphological family size such as that of Nagy and Anderson (1984). A sampling procedure, however, is not feasible in a digital analysis that seeks to identify all of the types and tokens in texts represented by morphological families, as was the case in the present study.

Our interest in the developmental appearances of the words in the 2,451 morphological families was another reason for not including words with *U* function less than 1 since the *EWFG* database does not provide information on the grade-level appearances of these words. We did conduct a preliminary analysis of 5% of the 2,451 lead words, randomly selected, to determine the level of prediction that was lost by the choice to limit the analysis to the *U>1* portion of the corpus. This preliminary analysis showed that 123 lead words had an average of 13.7 family members within the *U*<1 group. The number of additional family members was great but, across the 123 lead words, these family members added only an average of 1.3 additional appearances per million words. The insubstantial addition to the predicted appearances of this large group of words confirmed our decision to focus on the portion of the corpus that accounts for the majority of words in texts.

Conclusion

This descriptive study does not prescribe what, how, or when to teach particular words within the English lexicon. At the same time, the analyses did show that a majority of the words in texts across the school years can be parsed into a relatively small number of morphological families. These findings suggest a shared vocabulary in texts across the grade bands—a core vocabulary that merits additional investigation.

We offer the core vocabulary as a resource to researchers for theoretical and empirical investigations of the relationship of vocabulary to comprehension at different developmental

levels and with different types of texts. The amenability of this vocabulary to intervention and the effects of such interventions on comprehension and knowledge acquisition also merit attention. We believe that such research can contribute to models of vocabulary learning and instruction and comprehension that, over time, could increase the robustness and efficacy of curriculum and instruction. Ultimately, digital databases and tools hold promise for addressing the vocabulary gap and providing instructional solutions that could enable many more students to enter workplaces and communities with the literacy levels required for the 21st century.

References

Anglin, J.M. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development*, *58* (10, Serial 238).

Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (a travel update). In M. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 285–375). Amsterdam, NL: Academic Press.

Beck, I. L., McKeown, M.G., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction*. New York, NY: Guilford Press.

Becker, W. C., Dixon, R., & Anderson-Inman, L. (1980). *Morphographic and root word analysis of 26,000 high frequency words*. Eugene, OR: University of Oregon Follow Through Project, College of Education.

Betts, E. A. (1946). *Foundations of reading instruction*. New York, NY: American Book Co.

Biemiller, A. (2005). Size and sequence in vocabulary development: Implications for choosing words for primary grade vocabulary instruction. In E.H. Hiebert & M.L. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (pp. 223-242). Mahwah, NJ: Erlbaum.

Biemiller, A. (2010). *Words worth teaching: Closing the vocabulary gap*. Columbus, OH: McGraw-Hill SRA.

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, *93*(3), 498-520.

Bowers, P. N., Kirby, J. R., & Deacon, S. H. (2010). The effects of morphological instruction on literacy skills a systematic review of the literature. *Review of Educational Research*, *80*(2), 144-179.

Breland, H M., Jones, R.J., Jenkins, L., Paynter, M., Pollack, J., & Fong, Y.F. (1994). The college board vocabulary study. *ETS Research Report Series* (No. 1), i-51.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904-911.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977-990.

Carlisle, J. F. (2010). Effects of instruction in morphological awareness on literacy achievement: An integrative review. *Reading Research Quarterly*, *45*(4), 464-487.

Carlisle, J. F., & Katz, L. A. (2006). Effects of word and morpheme familiarity on reading of derived words. *Reading and Writing*, *19*(7), 669-693.

Carlisle, J. F., & Stone, C. (2005). Exploring the role of morphemes in word reading. *Reading Research Quarterly*, *40*(4), 428-449.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American heritage word frequency book*. Boston, MA: Houghton Mifflin.

Clay, M.M. (1991). *Becoming literate: The construction of inner control*. Portsmouth, NH: Heinemann.

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

Dale, E., & Chall, J.S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, **27**: 11–20; 37-54.

Dale, E., & O'Rourke, J. (1981). *The living word vocabulary*. Chicago, IL: World Book-Childcraft International.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159-190.

Davies, M. (2012). *The Corpus of American Soap Operas*: *100 million words, 2001-2012* Retrieved from http://corpus2.byu.edu/soap/.

Dolch, E. W. (1936). A basic sight vocabulary. *The Elementary School Journal*, *36*(6), 456-460.

Endress, A. D., & Hauser, M. D. (2011). The influence of type and token frequency on the acquisition of affixation patterns: Implications for language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 77-95

Fitzgerald, J., Elmore, J., Relyea-Kim, E.J., Hiebert, E.H., & Stenner, A.J. (2016). Has first-grade core-reading-program text complexity changed across six decades? *Reading Research Quarterly*, *51*(1), 7-28.

Fry, E. (1980). The new instant word list. *The Reading Teacher*, *34*(3), 284-289.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305-327.

Gates, A. I. (1962). The word recognition ability and the reading vocabulary of second-and third-grade children. *The Reading Teacher*, *15*(6), 443-448.

Goodwin, A. P., Gilbert, J. K., Cho, S. J., & Kearns, D. M. (2014). Probing lexical representations: Simultaneous modeling of word and reader contributions to multidimensional lexical representations. *Journal of Educational Psychology*, *106*(2), 448-468.

Goodwin, A. P., Gilbert, J. K., & Cho, S. J. (2013). Morphological contributions to adolescent word reading: An item response approach. *Reading Research Quarterly*, *48*(1), 39-60.

Goodwin, A. P., & Ahn, S. (2010). A meta-analysis of morphological interventions: Effects on literacy achievement of children with literacy difficulties. *Annals of Dyslexia*, *60*(2), 183-208.

Graves, M. F., Elmore, J., Bowen, K., Sanford-Moore, E.E., Copeland, M., Fitzgerald, J., Koons, H., & Stenner, A.J. (December 3, 2014). *The vocabulary of core reading programs*. Paper presented at the annual meeting of the Literacy Research Association, Marco Island, FL.

Graves, M.F., Juel, C.F., Graves, B.B., & Dewitz, P.F. (2010). *Teaching reading in the 21st century: Motivating all learners* (5th ed.). Hoeboken, NJ: Pearson.

Halladay, J. L. (2012). Revisiting key assumptions of the reading level framework. *The Reading Teacher*, *66*(1), 53-62.

Hazenberg, S., & Hulstun, J.H. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics, 17*(2), 145-163.

Hiebert, E.H. (2013). Core vocabulary and the challenge of complex text. In S. Neuman & L. Gambrell (Eds.), *Quality Reading Instruction in the Age of the Common Core State Standards* (pp. 149-161). Newark, DE: IRA.

Hiebert, E.H. (2005). In pursuit of an effective, efficient vocabulary curriculum for the elementary grades. In E.H. Hiebert & M. Kamil (Eds.), *The teaching and learning of vocabulary: Bringing scientific research to practice (*pp. 243-263). Mahwah, NJ: LEA.

Hiebert, E.H. (2002). Standards, assessment, and text difficulty. In A.E. Farstrup & S.J. Samuels (Eds.). *What research has to say about reading instruction* (3rd Ed., pp. 337-369) Newark, DE: International Reading Association.

Hiebert, E.H., & Raphael, T.E. (1996). Psychological perspectives on literacy and extensions to educational practice. In D.C. Berliner & R.C. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 550-602). New York: Macmillan.

Hill, D. R. (2008). Graded readers in English. *ELT journal*, *62*(2), 184-204.

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

Johnson, D. D., Smith, R. J., & Jensen, K. L. (1972). Primary children's recognition of high-frequency words. *The Elementary School Journal*, *73*(3), 162-167.

Jones, N. (1991). Development of morphophonemic segments in children's mental representations of words. *Applied Psycholinguistics*, *12*(2), 217-239.

Just, M.A., & Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329-354.

Kucera, H., & Francis, W.N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978-990.

Leech, G., & Rayson, P. (2014). *Word frequencies in written and spoken English: Based on the British National Corpus*. New York, NY: Routledge.

Levac, D., Colquhoun, H., & O'Brien, K.K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, *5*(1), 69. doi: 10.1186/1748-5908-5-69

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203-208.

McCutchen, D., Logan, B., & Biangardi-Orpe, U. (2009). Making meaning: Children's sensitivity to morphological information during word reading. *Reading Research Quarterly, 44*(4), 360-376.

MetaMetrics (2000). *The Lexile framework for reading*. Durham, NC: Author. Retrieved from lexile.com\about_meta\press\21098b.htm

Milone, M. (2009). *The development of ATOS: The Renaissance readability formula*. Wisconsin Rapids, WI: Renaissance Learning.

Mugglestone, L. (2013). *The Oxford history of English*. New York, NY: Oxford University Press.

Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, *19*(3), 304-330.

Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, *47*(1), 91-108.

Nation, I. S. P. (2013). *Learning vocabulary in another language*. New York, NY: Cambridge University Press.

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

National Center for Education Statistics (2012). *The Nation's report card: Vocabulary results from the 2009 and 2011 NAEP reading assessments* (NCES 2013–452). Washington, DC: Institute of Education Sciences, U.S. Department of Education.

National Center for Education Statistics (2015). *The Nation's report card: 2015 Mathematics and reading assessments* (NCES 2015-136). Washington, DC: Institute of Education Sciences, U.S. Department of Education.

National Governors Association Center for Best Practices & Council of Chief State School Officers (NGACBP & CCSSO). (2010). *Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects with Appendices A-B*. Washington, DC: Author. Retrieved from http://www.corestandards.org/the-standards

Pearson, P.D., Hiebert, E.H., & Kamil, M.L. (2007). Vocabulary assessment: What we know and what we need to know. *Reading Research Quarterly*, *42*(2), 282-296.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*(4), 437-448.

Rabin, J., & Deacon, H. (2008). The representation of morphologically complex words in the developing lexicon. *Journal of Child Language*, *35*(2), 453–465.

Reed, D. K. (2008). A synthesis of morphology interventions and effects on reading outcomes for students in grades K–12. *Learning Disabilities Research & Practice*, *23*(1), 36-49.

Ricketts, J., Nation, K., & Bishop, D. V. (2007). Vocabulary is important for some, but not all reading skills. *Scientific Studies of Reading*, *11*(3), 235-257.

Roe, B., Smith, S., & Burns, P.C. (2011). *Teaching reading in today's elementary schools* (11th ed.). Boston, MA: Cengage Learning.

Scott, J.A., Lubliner, S., & Hiebert, E.H. (2006). Constructs underlying word selection and assessment tasks in the archival research on vocabulary instruction. In J.V. Hoffman,

D.L. Schallert, C.M. Fairbanks, J. Worthy, & B. Maloch (Eds.), *55th Yearbook of the National Reading Conference* (pp. 264-275). Oak Creek, WI: NRC.

Sénéchal, M., Ouellette, G., & Rodney, D. (2006). The misunderstood giant: On the predictive role of early vocabulary to future reading. In D. K. Dickinson & S. B. Neuman, *Handbook of early literacy research* (Vol. 2, pp.173-182). New York, NY: Guildford Press.

Spache, G. (1953). A new readability formula for primary-grade reading materials. *The Elementary School Journal*, *53*(7), 410-413.

Stahl, S. A., & Heubach, K. M. (2005). Fluency-oriented reading instruction. *Journal of Literacy Research*, *37*(1), 25-60.

Stallman, A. C., Commeyras, M., Kerr, B., Reimer, K., Jimenez, R., Hartman, D. K., & Pearson, P. D. (1989). Are "new" words really new? *Literacy Research and Instruction*, *29*(2), 12-29.

Swanborn, M. S., & de Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, *69*(3), 261-285.

Thorndike, E. L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychology, 8*(6), 323-332.

Thorndike, E. L. (1921). *The teacher's word book*. New York, NY: Teachers College, Columbia University Press.

Thorndike, E. L. (1932). *A teacher's word book of 20,000 words*. New York, NY. Teacher's College, Columbia, University Press.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's handbook of 30,000 words*. New York, NY. Teacher's College, Columbia, University Press.

Venezky, R. L. (1999). *The American way of spelling: The structure and origins of American English orthography*. New York, NY: Guilford Press.

West, M. P. (Ed.). (1953). *A general service list of English words: With semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London, UK: Addison-Wesley Longman Limited.

Wright, T. S., & Cervetti, G. N. (in press). A systematic review of the research on vocabulary instruction that impacts text comprehension. *Reading Research Quarterly*.

Yang, M. C. (2005). A study of prototypes, design activity, and design outcome. *Design Studies*, *26*(6), 649-669.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, MA: Touchstone Applied Science Associates, Inc.

Table 1

Illustration of Predicted Appearances of a Word and its Family Members by Grade Levels

Word	K-1	2-3	4-5	6-8	9-10	11-CCR
improve	0	8	45	132	111	239
improved	1	3	21	65	64	154
improvement	0	0	5	20	24	79
improvements	0	0	2	13	19	59
improves	0	0	0	7	5	10
improving	0	0	5	14	13	56
Appearances in grade-band corpus	1	11	78	251	236	597
Predicted appearances per million words	1	7	47	99	140	233

Table 2

Number of Texts and Words in Exemplar Texts

Grade Band	# of Texts	Total Words	Average Length of Texts
K-1	14[1]	7,647	546.21
2-3	25	33,396	1,335.84
4-5	29	61,730	2,128.62
6-8	33	157,905	4,785.0
9-10	46	268,960	5,846.96
11-CCR	53	386,288	7,288.45
Total	200	915,926	4,579.63

[1]The three wordless books recommended by the CCSS for Grades K-1 are not included in the count of texts.

Table 3

Analysis of Family Members Within U10+ and *1≤U≤9 Frequency Groups*

	Lead Word		*U*10+		1≤*U*≤9		All Words in Families	
	# of Words	Predicted Appearance in One Million Words	# of Words	Predicted Appearance in One Million Words	# of Words	Predicted Appearance in One Million Words	# of Words	Predicted Appearance in One Million Words
K-1	1,307	582,219	1,932	166,684	3,419	10,518	6,658	759,421
2-3	480	24,993	518	14,987	1,078	3,351	2,076	43,331
4-5	290	10,758	258	5,475	680	2,328	1228	18,561
6-8	221	4,776	97	1,701	485	1,640	803	8,117
9-10	108	1,814	18	235	261	848	387	2,897
11-CCR	45	556	1	10	100	297	146	863
Total	2,451	625,116	2,824	189,092	6,023	18,982	11,298	833,190

Table 4

Models for Examination of Variation in Ranks of Words Across Grade Bands

	Model 1: Unconditional Growth Model L1: $Rank_{ij} = \beta_{0j} + \beta_{1j}time_{ij} + r_{ij}$ L2: $\beta_{0j} = \gamma_{00} + u_{0j}; \beta_{1j} = \gamma_{10} + u_{1j}$			Model 2: Effect of Frequency L1: $Rank_{ij} = \beta_{0j} + \beta_{1j}time_{ij} + r_{ij}$ L2: $\beta_{0j} = \gamma_{00} + \gamma_{01}freq_{1j} + u_{0j};$ $\beta_{1j} = \gamma_{10} + \gamma_{11}freq_{1j} + u_{1j}$		
Fixed Effects	Est.	SE	*p*	Est.	SE	*p*
Intercept	649.77	81.396	<.001	137.55	27.786	<.001
Time (grade band)	105.06	18.505	<.001	-6.13	8.193	.454
Frequency	----------			2.01	0.015	<.001
TimeXfreq	----------			0.44	0.004	<.001
Random Effects						
Var (Int)	15979364.20	463882.060	<.001	1599134.85	53216.435	<.001
Cov (Int,Slope)	3139222.45	97351.454	<.001	17411.37	11093.596	.117
Var (Slope)	757389.24	24002.891	<.001	79673.74	4760.920	<.001
Model Fit						
AIC	263266.445			256901.271		
BIC	263312.021			256962.039		

Table 5

Illustrations of Trends for Differences in Ranks of Words at Different Grade Bands

	Most frequent (*U* function >50)			Less frequent (*U* function <50)		
Exemplar	*little*	*much*	*a*	*abandoned*	*baked*	*occasionally*
U function	1067	1026	24070	18	10	31
Rank K,1	3112	698	21811	0	35	1
Rank 2,3	3214	1795	43052	5	53	9
Rank 4,5	2676	2116	47790	17	38	50
Rank 6-8	3155	3458	74090	43	57	214
Rank 9,10	1820	2202	49620	36	35	202
Rank 11-CCR	1871	2938	71694	73	32	390
Δ (time6-time1)	-1,241	2,249	49,883	73	-3	389

Table 6

Means (& Standard Deviations) Types & Tokens in CCSS Exemplar Texts of 2,451 Morphological Families (%)

Word Frequency Group	K-1		2-3		4-5		6-8		9-10		11-CCR	
	N[1]	I[2]	N	I	N	I	N	I	N	I	N	I
TOKENS												
U10+	97.2	92.8	89.7	88.6	89.6	87.6	88.1	86.3	87.6	85.5	87.0	85.0
With $1 \leq U \leq 9$	97.7 (.5)	96.6 (.08)	91.9 (3.7)	92.3 (1.6)	91.7 (2.2)	90.7 (1.2)	90.4 (2.4)	89.5 (3.4)	90.0 (6.2)	89.0 (13.3)	89.8 (6.0)	88.6 (6.6)
N & I combined	97.1 (.65)		92.1 (1.7)		91.1 (1.0)		89.8 (2.6)		90.0 (13.3)		89.1 (6.3)	
TYPES												
U10+	95.1	90.2	87.3	81.4	77.8	80.7	71.5	72.2	68.1	71.7	64.6	66.7
With $1 \leq U \leq 9$	96.4 (.01)	94.9 (.4)	90.6 (2.4)	87.0 (5.9)	84.3 (1.1)	86.2 (2.5)	78.8 (8.5)	79.9 (2.8)	76.1 (13.9)	79.4 (25.8)	73.6 (7.8)	75.7 (9.1)
N & I combined	95.6 (.52)		88.9 (6.7)		85.5 (1.0)		79.6 (11.2)		76.1 (21.4)		74.9 (5.7)	

[1]N stands for Narrative
[2]I stands for Informational

Table 7

Excerpts of Texts[1] in Six Grade Bands Illustrating Tokens and Types From 2,451 Morphological Families

	K-1	2-3	4-5	6-8	9-10	11-CCR
Title	*Green Eggs & Ham*	*From Seed to Pumpkin*	*Tuck Everlasting*	*Roll of Thunder*	*I Have A Dream*	*Society & Solitude*
Token (Type) %	96.4 (96.0)	90.9 (87.2)	92.2 (85.4)	89.8 (75.1)	90.5 (77.7)	88.4 (76.2)
	I[2] will not eat them here or there. I will not eat them anywhere. I do not eat green eggs and **ham[3].** I do not like them,…	The farmer tends the **pumpkin** patch to keep weeds out. Weeds take water from the soil. **Pumpkin** plants need that water to grow.	…tall water grasses whispering away from its sides, releasing it. Here and there the still surface of the water **dimpled,** and bright rings spread noiselessly and **vanished**.	He **sulked** for a while with a few **audible grumbles** which no one paid any attention to, but finally he fell asleep and did not awaken…	We can never be satisfied as long as the **Negro** is the victim of the **unspeakable** horrors of police **brutality.** We can never be satisfied,…	But how **insular** and **pathetically solitary** are the people we know! Nor dare they tell what they think of each other when they meet in the street.

[1]Samples of approximately 25 words

[2]Words in gray are members of the 2,451 morphological families (predicted to appear once or more per million words of text).

[3]Words bolded in black are not within the 2,451 morphological families.

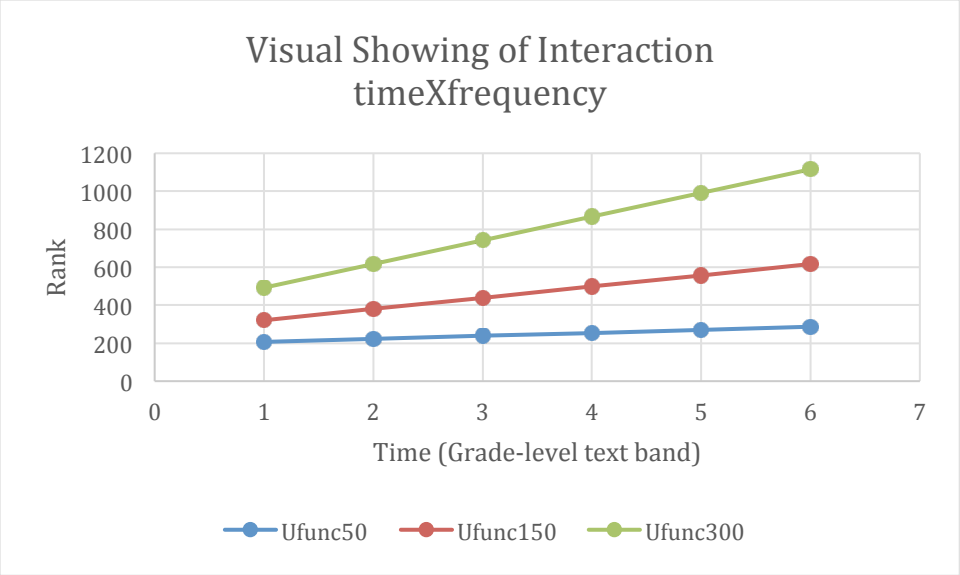

Figure 1. Graphical representation of interaction between grade band and frequency showing that the slope of the increasing frequency rank (suggesting lower prevalence) across grade bands differs depending on a lead word's frequency. For words with a frequency of 50, ranks of words do not increase significantly across grade bands. For words with a frequency of 150 or 300, ranks of words increase significantly (suggesting lower prevalence) across grade bands. Note that time=1 for K,1 texts; time=2 for gr2-3 texts; time=3 for gr4-5 texts; time=4 for gr6-8 texts; time=5 for gr9-10 texts; and time=6 for gr11&college texts.

Appendix A

Major Word Frequency Projects Over The Past 95 Years

Word List	Date of Publication	Source of Sample	# Tokens	# Types
Teacher's Word Book	Thorndike, (1921, 1932) Thorndike & Lorge (1944)	Children's literature, Bible & English classics, elementary-school textbooks, books on trades & domestic arts, daily newspapers, correspondence	4,565,000	10,000; 20,000; 40,000
The general service list of English words	West (1953)	Visual inspections by semanticists of 5 million words from various sources (e.g., encyclopedias, magazines, textbooks, novels)	5 million words	2,000 highest frequency words
Computational Analysis of Present-Day English	Kucera & Francis (1967)	500 samples from 15 categories of adult printed material in 1961	1,014,232	50,406
American Heritage Word Frequency Book	Carroll, Davies, & Richman (1971)	1,045 titles used in schools (Grades 3-9) based on a survey of educators in 1969; 1,000 different publications	5,088,721	86,741

College Board Corpus	Breland, Jones, Jenkins, Paynter, Pollack, & Fong (1994)	High school and first-year college courses, including articles from newspapers & magazines	14,360,884	Not specified
Educator's Word Frequency Guide	Zeno, Ivens, Millard, & Duvvuri (1995)	60,527 text samples from 6,333 textbooks, literature, popular fiction, & nonfiction in use in American schools & colleges	17,274,580	154,941
Hyperspace Analogue to Language (HAL)	Lund & Burgess (1996)	3,000 Usenet newsgroups during February 1995	131 million words	Not specified
SUBTLEX$_{US}$	Brysbaert & New (2009)	Subtitles of U.S. films and television series gathered from the website www.opensubtitles.org	51 million	Not specified
Corpus of Contemporary American	Davies (2009)	160,000 texts, including 20 million words each year from 1990 to 2011, with each yearly	450 million	100,815

English (COCA)		corpus evenly divided across 5 genres: spoken, fiction, popular magazines, newspapers, & academic journals		
British National Corpus (BNC)	Leech & Rayson (2014)	90% from written samples from newspapers, fiction, non-fiction books, and speeches & 10% from oral samples	100 million	794,771
Corpus of American Soap Operas (SOAP)	Davies (2012)	Transcripts from 10 American soap operas from 2001 to 2012	100 million words	Not specified

Appendix B

A Comparison of the Rankings of the 5,500 Most-Frequent Words Across Databases

To consider whether rankings of high-frequency words in the *EWFG* are similar to those of other databases, the rankings of the first 5,500 words on the *EWFG* were compared with those of the same words in five databases: (a) three databases (Corpus of Contemporary American English (COCA), Davies, 2009; British National Corpus (BNC), Leech & Rayson, 2014; Kucera-Francis, 1967) from conventional texts (i.e., books, newspapers) and (b) two (Corpus of American Soap Operas (SOAP), Davies, 2012; *SUBTLEX*$_{\textbf{\textit{US}}}$, Brysbaert & New, 2009) from written language corpora that approximate oral language (e.g., television scripts).

The correlations in Table B-1 indicate that rankings of words on the *EWFG* are similar to those in the three databases of conventional written language (i.e., COCA, BNC, Kucera-Francis) but have less overlap with the rankings of databases that are proxies of oral language (i.e., SOAP, *SUBTLEX*$_{\textbf{\textit{US}}}$). In that our interest is in the lexicon of written text, this analysis validated our choice of the *EWFG*.

Table B-1

Correlations Between 2,451 Lead Words: Six Data Bases (U Functions)

	EWFG	COCA	BNC	Kucera-Francis	SOAP
EWFG					
COCA	.948				
BNC	.950	.996			
Kucera-Francis	.949	.986	.994		
SOAP	.651	.715	.673	.619	
SUBTLEX$_{\textbf{US}}$	.689	.653	.673	.653	.841

Appendix C

Analysis of Semantically Transparent and Semantically Opaque Assignments in Becker et al. (1980) Database

Becker et al.'s (1980) assignments of derived words to a root word were compared to those of Nagy and Anderson (1984). Nagy and Anderson coded words according to six levels of semantic relatedness between a word and its root word or immediate ancestor. Their coding system was reliable only when the six categories were collapsed into two groups. Consequently, they used a two-category dichotomy: (a) semantically transparent: a word's meaning can be inferred immediately or with reasonable textual context (*misrepresent/represent*) or (b) semantically opaque: a word's meaning requires either substantial textual context to connect derived word with root word or is not discernible (*condescend*/*descend).*

In the Nagy and Anderson sample, 13.2% of all words had affixes. Of this group, 84.4% had semantically transparent relations to the root word and 14.6% had semantically opaque relations. Nagy and Anderson (1984) provided 32 examples of target words and immediate ancestors for the semantically transparent group. Becker et al.'s coding of semantically transparent words agreed 100% with Nagy and Anderson's coding.

To exemplify semantically opaque words, Nagy and Anderson provided 35 examples. Becker et al. coded 67% of this group in the same manner as Nagy and Anderson (e.g., *visualize* with *visual; ominous* with *omen*). For the other 33% of the words in this group, Becker et al. left the target word intact and did not place it with an ancestor. For example, the word *prefix* was classified as having the root word *prefix* and not the semantically opaque choice of *fix* identified by Nagy and Anderson.

If our database follows the distribution of semantically transparent and opaque words described within the lexicon by Nagy and Anderson (1984), we would anticipate that approximately 9.3% of the affixed words in the Becker et al. sample have semantically opaque connections to the lead word. That is, approximately one in 10 every morphological family members would have an opaque connection with the root word.

Semantic opaqueness, however, may be a function of frequency. In the Nagy and Anderson sample of words with semantically opaque relationships to root words, 48% had predicted frequencies of $U<1$, 43% were in the $1 \leq U \leq 9$, and only 9% were in the $U10+$ group. Additional analyses are required to understand the nature of semantic opaqueness as a function of developmental level of text and also in the challenge that these words pose for readers at different developmental levels.