METAMETRICS RESEARCH BRIEF

The Vocabulary of Core Reading Programs

Jeff Elmore, Senior Research Engineer, Kim Bowen, Ph.D., Research Associate, Michael Graves, Ph.D., Professor Emeritus, University of Minnesota, and Jill Fitzgerald, Ph.D., Distinguished Research Scientist, MetaMetrics; Research Professor and Professor Emerita, The University of North Carolina at Chapel Hill

August 2015

OBJECTIVE

As the first part of a comprehensive research endeavor to identify the words that U.S. students are likely to encounter in grades K-12, this study involved analyzing four current, best-selling, elementary core reading programs. From this analysis, frequency lists for words and word families appearing in these programs were developed. These frequency lists are available for researchers, publishers, and educators.

METHODS

Data Sources:
Based on the Education Market Research database (2012) and discussions with publishers about recent consolidations, four current, best-selling, core reading programs were identified. The words tallied for the study included all words that were intended for students to read from the student editions of each of the programs.

Procedures:
First, the texts were digitized and edited according to a standardized protocol. Through computer analysis, the number of tokens (running words) and types (unique words) were tabulated by program and grade.

To reduce the presence of potential non-words arising from Optimal Character Recognition (OCR) errors and other irregularities, character strings not appearing in the 450-million-word Corpus of Contemporary American English (Davies & Gardner, 2010) were removed from further analyses.

Next, words were grouped into families at four different levels. In Level 1, every different word form was counted as unique (types). In Level 2, a family consisted of the base word and its inflected forms. Level 3 added forms to the family that include the 10 most frequent prefixes and derivational suffixes (White, Sowell, & Yanagihara, 1989). Level 4 added forms that include over 250 prefixes and suffixes listed in the English Lexicon Project (ELP) database (Balota et al., 2007).

To group words into families, bases and families were identified using the ELP database. For words not found in the ELP database, a modified version of Morfessor 2.0 (Virpioja, Smit, Grönroos, & Kurimo, 2013) was used to segment words into morphemes, and each word form was assigned to one of the four levels of word families.

Finally, the reliability of the computer determinations was checked by comparing them to the decisions of two judges with extensive experience in language and education. The computer determinations were concordant with the judges' decisions 85% of the time.

ANALYSES

Descriptive statistics of tokens and types at each of the four levels of word families were calculated.

RESULTS & DISCUSSION

As shown in Table 1, the four programs contain a total of 2,111,344 tokens (running words). The number of words per grade increases as grades increase across all four programs. Still, comparison of the longest and shortest programs shows some substantial differences. For example, overall, Program 1 includes almost twice as many words as Program 2; and in Grade 1, Program 4 includes markedly fewer words than the other programs.

Table 1: Number of tokens included, by grade, program, and overall.

	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	All Grades*
Program 1	13,939	81,757	121,942	142,853	138,395	165,349	664,235
Program 2	12,884	39,100	56,115	71,390	88,805	112,710	381,004
Program 3	18,416	64,000	90,740	103,581	121,343	123,151	521,231
Program 4	8,196	78,684	86,033	112,755	126,734	132,472	544,874
All Programs	53,435	263,541	354,830	430,579	475,277	533,682	2,111,344

*The sum of the number of tokens in each grade

As shown in Table 2, each program includes a wide range of unique word families per grade (from 1,066 in the shortest program in Grade 1 to 8,060 in the longest program in Grade 6) with more unique word families appearing at each grade level across all programs. Thus, children face an increasingly sizeable task each year. The largest increase occurs between Grade 1 and 2, with each program including more than twice as many types in the Grade 2 books than in the Grade 1 books. Comparison of the number of different word families contained in the programs across all six grade levels reveals only modest differences, with the largest difference being between Program 2 with 12,926 unique Level 2 word families and Program 1 with 14,158 unique Level 2 word families.

Table 2: Number of unique Level 2 word families included, by grade, program, and overall.

	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	All Grades*
Program 1	1,778	4,258	6,037	6,882	6,912	8,060	14,158
Program 2	1,510	3,540	4,454	5,916	6,104	7,691	12,926
Program 3	1,633	3,741	5,224	6,159	6,782	7,738	13,301
Program 4	1,066	3,653	4,627	6,127	6,683	7,676	12,952

*Total number of unique tokens when all grades are combined

As shown in Table 3, the first 100 word families account for 44-45% of the tokens at each grade level and the first 4,000 word families account for nearly 90% of the tokens, and these percentages are relatively constant across the grade levels and programs. Clearly, it is essential that students know words from these 4,000 word families if they are to avoid frequently stumbling over words as they progress through these core reading programs. At the same time, students need to know a number of additional words, many of which occur infrequently, if they are to fully comprehend and learn from programs like these.

Table 3. Percentages of tokens among the first 100, 500, 1000, and ~ 4000 most frequent families by grade.

	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
First 100	44%	44%	44%	44%	45%	45%
First 500	68%	68%	67%	65%	66%	66%
First 1000	76%	76%	75%	74%	74%	74%
First ~4000	88%	88%	88%	87%	87%	86%

As Hiebert and Fisher (2006) have noted, there is very limited knowledge about the number of repetitions children need in order to learn words. Gates (1930), for example, studied beginning readers and very frequent words and estimated that students with low reading ability need about 55 repetitions. Nation (2014), studied adults learning English as a foreign language and found that they need about 12 repetitions. As shown in Table 4, in the programs studied, the majority of words (about 80%) are repeated 9 or fewer times (sum of “2 to 9” and “1 Occurrence”), with more than 30% occurring only once. By themselves, core reading programs may not be able to provide enough repetition to enable students to learn the majority of words they contain.

Table 4. Number of occurrences of unique Level 2 word families in Grade 1 books in each of the four programs.

	55 or More	21 to 54	10 to 20	2 to 9	1 Occurrence
Program 1	2%	4%	12%	45%	37%
Program 2	2%	6%	10%	45%	37%
Program 3	3%	7%	11%	46%	33%
Program 4	2%	6%	10%	48%	34%

If a student is able to generalize meanings from base words to words with inflected forms, the learning task is substantially reduced. For example, in Program 1, instead of having to learn 21,575 words, a student must learn 14,158 words, a reduction of 34% in the learning task. This is a level of morphological generalization that most native English-speaking children have achieved, even in Grade 1. If a student can generalize meaning from base words to many derived forms of words (Level 4), the learning task is further reduced, although less significantly – on average an additional 13% over the reduction from inflected forms alone. This is a level of morphological generalization beyond that of most children, even in Grade 6. For readers of the programs in this study, being able to recognize the inflected forms of a base word substantially reduces the learning task; while being able to recognize derived forms, even a lot of derived forms, reduces the learning task, but to a lesser degree.

Table 5. Number of words and word families contained in each program for individual word forms, and Level 2, Level 3, and Level 4 word families.

	Level 1: Individual Forms	**Level 2: Root Plus Inflections**	**Level 3: Root Plus Inflections and Common Derived Forms**	**Level 4: Root Plus Affixes in ELP**
Program 1	21,575	14,158	12,656	11,767
Program 2	19,296	12,926	11,606	10,911
Program 3	20,045	13,301	11,989	11,309
Program 4	19,645	12,952	11,592	10,828

Importance of Understanding the Vocabulary Learning Task Students Face

Research has shown that some children from lower socioeconomic backgrounds and some English as a foreign language learners are likely to have smaller vocabularies than their middle class, native-English speaking peers (Fernald, Marchman, & Weisleder, 2012; Goldenberg, 2013; Hart & Risley, 1995; White, Graves, & Slater, 1990). What research has not yet shown is which words children of various ages, socioeconomic backgrounds, and linguistic backgrounds do and do not know (Pearson, Hiebert, & Kamil, 2007). Once we know which words appear in the material students read, and which of those words students know, we will be in a position to provide differentiated instruction to those students with vocabularies markedly smaller than those of their peers.

Additional Research Needed

As noted in the introduction to this research brief, the study described here is part of a comprehensive research endeavor. The value of the work will increase markedly with additional research. We have identified three lines of research that are particularly important: (1) descriptions of the words in other corpora (content area texts, high school material, etc.), as well as parts of these corpora such as elementary level general- and content-specific academic vocabulary and secondary level general- and content-specific academic vocabulary; (2) information on the knowledge students of various ages, socioeconomic backgrounds, and linguistic backgrounds have of the words in various corpora; and (3) understanding of the factors that affect word learning including better estimates of the number of repetitions young readers need, on average, to learn words.

The findings of this study were used to create word frequency lists. These lists are available for researchers, publishers, and educators for instructional use, including:

- determining which words to teach to individual students
- planning district or state vocabulary curriculum
- designing instructional programs for reading and/or vocabulary, and
- developing assessments

REFERENCES

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). *The English Lexicon Project. Behavior Research Methods*, *39*, 445-459.

Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists*. New York: Routledge.

Education Market Research. (2012). *Reading market 2012*. New York: Author.

Fernald, A., Marchman, V. A., & Weisleder, A. (2012). SES differences in language processing skill and vocabulary are evident at

18 months. *Developmental Science*, *16*, 234–248.

Gates, A. I. (1930). *Interest and ability in reading.* New York: Macmillan.

Goldenberg, C. (2013). Unlocking the research on English learners. *American Educator*, *37*(2), 4-12, 38.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experiences of young American children.* Baltimore: P. H. Brookes.

Hiebert, E.H. & Fisher, C.W. (2006). Fluency from the first: What works with first graders. In T. Rasinksi, C.L.Z. Blachowicz, & K. Lems (Eds.), *Fluency instruction: Research-based best practices.* New York: Guilford Press.

Nation, P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language, 26*(2), 1–16.

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

Virpioja, S., Smit, P., Grönroos, S-A., & Kurimo, M. (2013). Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline. Aalto University publication series *SCIENCE + TECHNOLOGY,* 25/2013. Aalto University, Helsinki.

White, T. G., Graves, M. F. & Slater, W. H. (1990). Growth of reading vocabulary in diverse elementary schools: Decoding and word meaning. *Journal of Educational Psychology*, *82*, 281–290.

White, T. G., Sowell, J., & Yanagihara, A. (1989). Teaching elementary students to use word-part clues. *The Reading Teacher, 42*, 302-308.

MetaMetrics® is focused on improving education for students of all ages. The organization develops scientific measures of academic achievement and complementary technologies that link assessment results with instruction. For more than twenty years, MetaMetrics' work has been increasingly recognized worldwide for its distinct value in differentiating instruction and personalizing learning. Its products and services for reading, mathematics and writing provide valuable insights about academic ability and the potential for growth, enabling students to achieve their goals at every stage of development.

METAMETRICS®, the METAMETRICS® logo and tagline, LEXILE®, LEXILE® FRAMEWORK, and the LEXILE® logo are trademarks of MetaMetrics, Inc., and are registered in the United States and abroad. The trademarks and names of other companies and products mentioned herein are the property of their respective owners. Copyright © 2015 MetaMetrics, Inc. All rights reserved.

MetaMetrics
LINKING ASSESSMENT WITH INSTRUCTION