METAMETRICS RESEARCH BRIEF

Lexile Word Frequency Profiles

Jeff Elmore

AUGUST 2016

OBJECTIVE

This report describes the development of a new type of frequency measure for words that better reflects the developmental nature of word exposure. The relationship between word frequency and word knowledge has been well documented (Brysbaert, Buchmeier, Conrad, Jacobs, Bölte, & Böhl, 2011; Rudell, 1993). Indeed, word frequency is the operational measure of semantic difficulty in the equation powering the Lexile Analyzer® (Stenner, Horabin, Smith, & Smith, 1988; Stenner, Burdick, Sanford, & Burdick, 2007). However, the underlying theoretical explanation of why word frequency predicts word knowledge is exposure. Readers are exposed more often to more frequent words, and thus have greater knowledge of them (Klare, 1963). The connection between word frequency and word knowledge therefore is more meaningful if the word frequencies more accurately reflect the degree of exposure to a word for the average developing reader.

Leveraging the power of the 1.4-billion-word MetaMetrics® corpus of texts intended for readers in U.S. K-12 schools, we developed the Lexile Word Frequency Profile, a set of frequency measures describing the developmental trajectory of a word's occurrences along the Lexile scale. We demonstrate the meaningfulness and utility of Lexile Word Frequency Profiles by examining a few specific word profiles and using the profiles to predict other measures of word familiarity.

METHODS

Lexile measures (Stenner et al., 1988): a developmental scale that measures reader ability and text complexity on a common scale using semantic and syntactic features. Independent psychometric studies of the Lexile scale (Mesmer, 2007; White & Clement, 2001) indicate that it is a valid and reliable measure of reader ability and text complexity.

Data Sources

MetaMetrics Corpus: The corpus used in this study comprised 91,935 texts including textbooks, trade books, leveled readers, and other texts such as supplemental textbook material intended for K-12 students. Lexile text measures typically range from above 200L to below 1400L. The corpus contains 255,744 unique words appearing at least 10 times, totaling 1,390,320,260 running words.

Age-of-Acquisition Ratings: A database of word familiarity ratings based on the estimated age at which a person will know the meaning of a word (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). The database contains ratings from ages 1.5 to 25 for 51,625 words and is used to assess the additional predictive power of Lexile Word Frequency Profiles over a single frequency measure.

PROCEDURES AND ANALYSES

Books were digitized and edited according to the guidelines for Lexile analysis. Occurrences of each word were tallied for each Lexile Zone. A Lexile Zone contains all texts within a 100L range, for example the 200L Zone contains texts measuring from 200L to 299L. Since each text in the corpus has a Lexile measure, frequency counts can be generated for occurrences in texts within each Lexile Zone. For example, we can count the number of times the word *rabbit* appeared in 0L to 99L texts, 100L to 199L texts, etc. A Lexile Word Frequency Profile consists of a set of frequency measures from below 0L to 2200L. In addition to a raw count of each word in each zone, several other kinds of counts were calculated and are described below.

Word Family Frequency Counts

To account for the derivational nature of the English language, frequency counts were generated for several levels of morphological word family relationships. For example, the word *uncommonly*, which may be relatively infrequent, is likely a more familiar word because it is composed of the relatively frequent word *common* and two frequent affixes *un-*, and *-ly*. Because not all derivations are equally transparent, four levels of morphological relationships were considered and word frequency counts were calculated at each level (Elmore, Fitzgerald, Graves, & Bowen, 2015):

- Level 1—every word form is counted uniquely.
- Level 2—base words and their inflected forms and derived forms with the suffixes such as *-ed* (past), *-en* (past participle), and *-ing* (present participle) are counted together.
- Level 3—all the forms in Levels 1 and 2 plus the 10 most frequent prefixes and suffixes such as *-ly* and *un-* are counted together.
- Level 4—all the forms in Levels 1 through 3 plus 107 prefixes and 108 suffixes listed in the English Lexicon Project (ELP) database (Balota et al., 2007) are all counted together. For example, the word *pseudoscientific* would be considered a part of the *science* word family.

Raw, Relative, Cumulative, and Reverse Cumulative Counts

For each word, four types of counts were calculated at each Lexile Zone and for each word family level: raw, relative, cumulative, and reverse cumulative. Relative counts are calculated as the frequency of a particular word in a particular zone divided by the total number running words in that zone. Cumulative counts tally the number of occurrences in a particular Lexile Zone and all previous zones. Reverse cumulative counts

tally the number of occurrences in a Lexile Zone and all subsequent zones (e.g., the number of occurrences in all texts above 1200L). Finally, confidence intervals were calculated for all of the counts (Brown, Cai, DasGupta, 2001).

All possible combinations of Lexile Zones, word family levels, and types of counts were calculated. For example, for the word *jump*, one could access a count of the number of times either *jump, jumps, jumper,* or *jumping* occurred in all texts 600L and below, or the relative frequency of just the word 'jumper' in only texts from the 200L Lexile Zone, with an estimated 95% confidence interval.

After Lexile Word Frequency Profiles were created for all words, several pairs of words with approximately the same overall frequency were compared to illustrate the additional information contained in the profile above and beyond a single frequency measure.

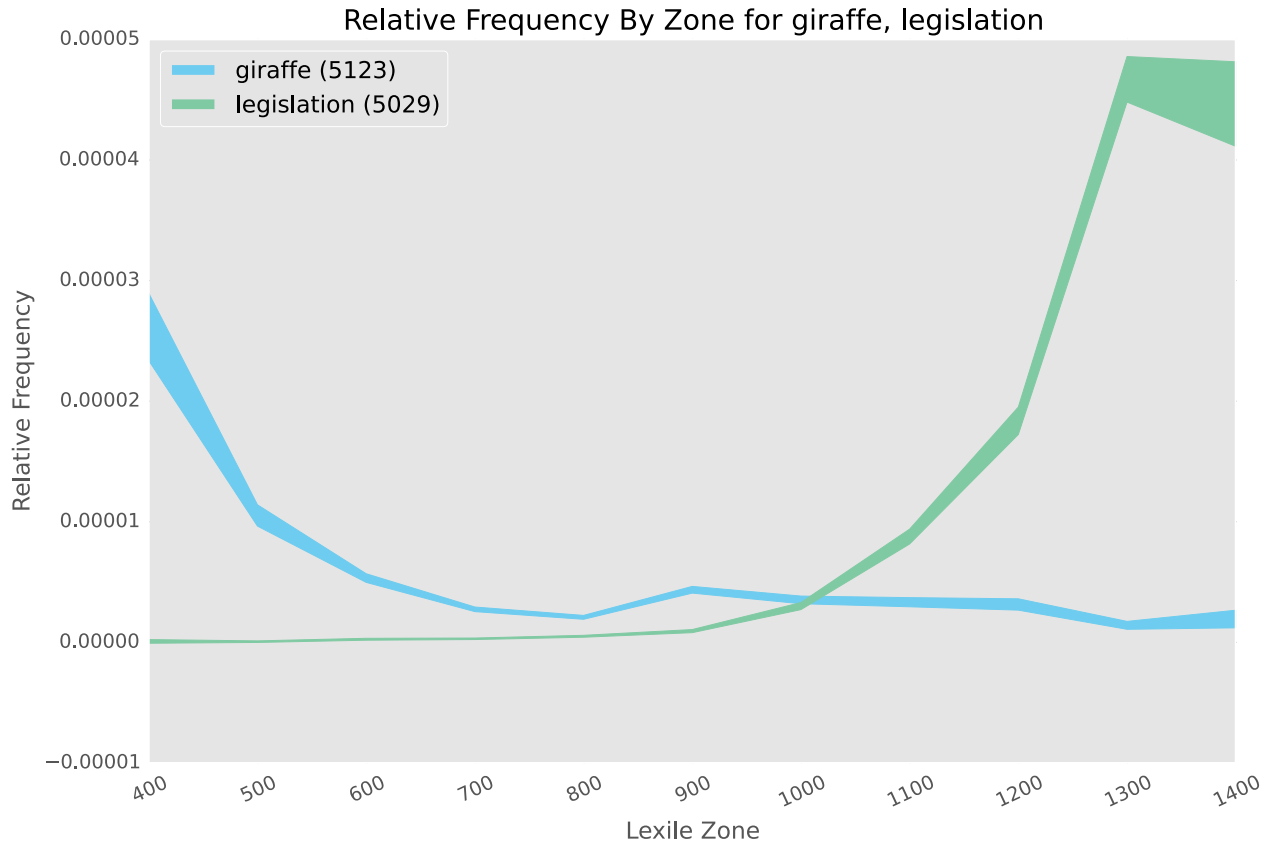
Finally, the power of Lexile Word Frequency Profiles to predict another measure of word familiarity was compared with the baseline of a prediction using a single summative word frequency measure. To evaluate the additional value of Lexile Word Frequency Profiles, two machine learning models called "Random Forest Regression" models (Breiman, 2001) were developed to predict age-of-acquisition ratings (Kuperman, et al., 2012): (1) a model using only the overall frequency, and (2) a model using all 624 Lexile Word Frequency Profile measures. Accuracy of the models was evaluated using the out-of-bag R^2 measures.

RESULTS

In total, 624 individual frequency measures were calculated for 255,744 words. Next, we selected several pairs of words with similar overall frequencies to demonstrate the additional information that Lexile Word Frequency Profiles provide. Although two words may occur about the same number of times overall, they can often have significant differences in the patterns of their use as reflected in their profiles.

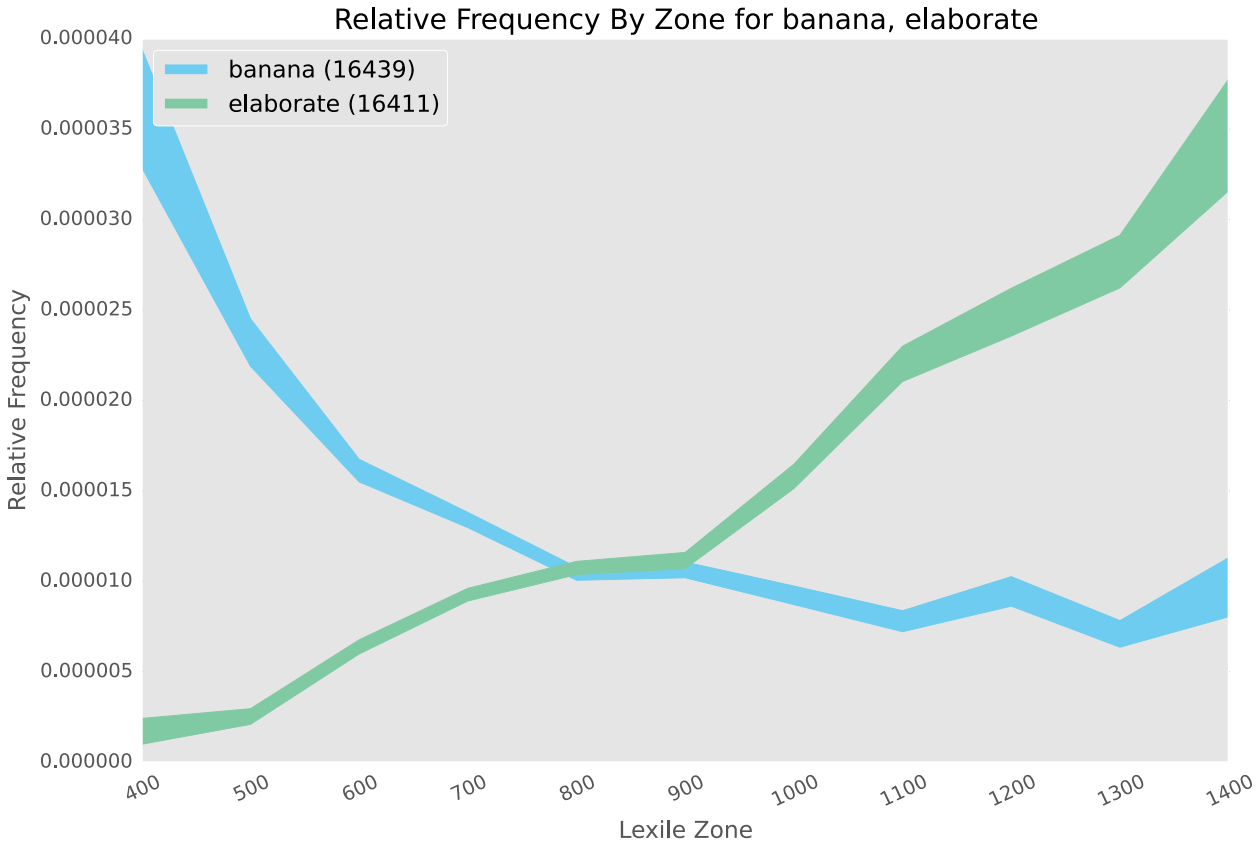
It was hypothesized that so-called academic words that tend to appear in school reading materials but not in stories or in conversations, like *analyze* or *consider*, would likely be more concentrated in higher Lexile texts; whereas, words representing familiar concrete objects such as animals and foods would likely be more concentrated in lower Lexile texts. For example, Figure 1 shows the Lexile Word Frequency Profiles for the words *giraffe* and *legislation* which both appear approximately 5,000 times overall, but have radically different profiles with *giraffe* occurring most frequently in lower Lexile texts and *legislation* occurring most frequently in higher Lexile texts above 1100L.

Figure 1.



Similarly, in Figure 2, *banana* and *elaborate* both occur approximately 16,000 times overall, but have a similar pattern of difference in their Lexile Word Frequency Profiles to the words in Figure 1.

Figure 2.



These examples suggest that Lexile Word Frequency Profiles do indeed capture meaningful differences in word usage for important categories of words such as academic words and concrete nouns like animals and foods.

However, anecdotal evidence from a few examples is less compelling than examining differences over tens of thousands of words. To assess more broadly the additional value of Lexile Word Frequency Profiles, we assessed the power of the profiles to predict another word familiarity measure. Two different random forest models were fit to predict age-of-acquisition ratings for approximately 50,000 words. The first model was a baseline model using a single summative measure of word frequency. The second model was a Lexile Word Frequency Profile model using all 624 measures. The baseline model accounted for 25% of the variance in age-of-acquisition measures while the Lexile Word Frequency Profile model accounted for 75% of the variance in age-of-acquisition measures.

CONCLUSION

Considering evidence from both a visual inspection of pairs of similarly frequent words and quantitative analyses related to predicting other word familiarity metrics, Lexile Word Frequency Profiles appear to provide a rich source of information about the familiarity of words. The profiles may offer insight into how different kinds of words exhibit different patterns of usage in texts of varying complexity level corresponding with different levels of likely exposure for the average reader. More accurate information about word exposure is potentially of value to publishers, curriculum developers, educators, and researchers.

REFERENCES

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Brown, L.D., Cai, T.T., & DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science*, *16*(2), 101–133.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A.M., Bölte, J., & Böhl, A. (2011). The word frequency effect: a review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*(5), 412–424.

Elmore, J., Fitzgerald, J., Graves, M., & Bowen, K. (2015). The Vocabulary of Elementary Grades Disciplinary Textbooks (MetaMetrics Research Brief). Durham, NC: MetaMetrics, Inc..

Klare, G.R. (1963). *The measurement of readability*. Ames, IA: Iowa State University Press.

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

Rudell, A.P. (1993). Frequency of word usage and perceived word difficulty: Ratings of Kucera and Francis words. *Behavior Research Methods, Instruments, & Computers*, *25*(4), 455-463.

Stenner, A.J., Horabin, I., Smith, D.R., & Smith, M. (1988). The Lexile Framework. Durham, NC: MetaMetrics, Inc.

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2007). The Lexile Framework for Reading Technical Report. Durham, NC: MetaMetrics, Inc.