Word Maturity: A New Metric for Word Knowledge

Thomas K. Landauer, Kirill Kireyev & Charles Panaccione

To cite this article: Thomas K. Landauer, Kirill Kireyev & Charles Panaccione (2011) Word Maturity: A New Metric for Word Knowledge, Scientific Studies of Reading, 15:1, 92-108, DOI: 10.1080/10888438.2011.536130

To link to this article: http://dx.doi.org/10.1080/10888438.2011.536130

Published online: 18 Jan 2011.

Article views: 419

View related articles

Citing articles: 11 View citing articles
A new metric, Word Maturity, estimates the development by individual students of knowledge of every word in a large corpus. The metric is constructed by Latent Semantic Analysis modeling of word knowledge as a function of the reading that a simulated learner has done and is calibrated by its developing closeness in information content to that of a simulated literate adult. Individual human learner knowledge is aligned with the simulation by adaptive testing. Evidence of accuracy, example applications to vocabulary assessment, teaching and reading research, properties of the metric, and a conjecture about its possible wider importance are described.

This article introduces a new metric for the development of reading-comprehension-oriented word knowledge. In brief, a computational simulation based on Latent Semantic Analysis (LSA) first creates learning trajectories for each unique orthographic word-form in a text corpus that is representative in size and content of the lifetime reading of a literate adult. A metric-based adaptive test is used to align individual reader knowledge with the simulation. The technology can then make estimates of degree of knowledge of every individual word for individual readers relative to that of an average literate adult.

We begin with a brief discussion of the need and value of such a metric, comparing it with the current state of the art in vocabulary assessment. Next come sections on LSA and how it is used to create the new metric. This includes a review of evidence that LSA does the things that the metric requires. This is
followed by further motivation for such a metric. In the Discussion section, we describe formal properties of the metric more thoroughly, then exemplify some of its anticipated reading-relevant assessment and tutorial applications and describe some needed and desired next steps and possible future work. Finally, we offer a conjecture on how the metric might help to ameliorate current wide disparities in reading achievement.

THE NEED FOR A NEW METRIC

Because early vocabulary acquisition is obviously a positive self-reinforcing growth function (see Landauer & Dumais, 1997, for a model thereof), lack of normal mastery in early school years puts many students at lifelong academic disadvantage (Chall & Jacobs, 2003; Hart & Risley, 1995.) Our hypothesized solution is to move vocabulary deficient learners from a shallow growth curve to a steeper one while the job is not too big. We return to this idea in the final section of the article.

The problem directly addressed is that current vocabulary assessment and instruction is insufficient for such goals. Although estimating the average frequency or school level of the vocabulary known by a reader is a well-developed technology (e.g., Laufer, Elder, Hill, & Congdon, 2004, Meara & Buxton, 1987; Nation, 1983; Read 1988), with some recent exceptions (e.g., Frishkoff, Collins-Thompson, Perfetti, & Callan, 2008), what is measured is essentially only the number, kinds, usage frequency, or grade levels in which they are used. Biemiller’s (2008); Beck, McKeown, and Kucan’s (2002); and similar prescriptions for appropriate words and kinds of words to teach are steps in this direction. But to suggest how much is missed, simulations described next estimate that the number times a word has previously been encountered account for only about half of the variance in how well it is known. Other simulations to be described suggest that when and how a word is learned account for most of the difference. However, it is really the result of all factors, known and unknown, that we want to measure. A measure that accurately captures the developmental course of every individual word for every individual reader, rather than just average levels was the primary goal.

Other capabilities that a better measure was also hoped to enable included (a) predicting what words are not well enough known by a given reader to comprehend a particular text, (b) what words are most important for that purpose, and (c) what words a given reader is ready to learn.

THE WORD MATURITY METRIC

Word Maturity (WM) measures the evolution of the information carried by an orthographic word-form (hereafter “word” unless modified) from its first textual
encounter to the information that it caries for a simulated adult reader. The information in a word is represented by a mathematical vector, a string of real numbers. The vectors in turn are derived by LSA (see Landauer, 2002, 2007, Landauer & Dumais, 1997, or Martin & Berry, 2007) from a language corpus representative of the total cumulative reading of an average adult reader of the language. LSA uses the statistical method Singular Value Decomposition (SVD) to represent every word and passage (usually a paragraph) in such a way that the vector standing for a passage is always the sum of the vectors standing for each of the words it contains and the vector for each word is always the average of the vectors for all the passages in which it occurs.

Because an LSA representation doesn’t change as a function of where in a sentence or passage a word occurs or what other words it occurs with, the nature of a word in the metric is only the information carried by its LSA vector, not its interactions with other words. We view this as a major advantage because of its rigorous specificity, but readers should avoid confusion with other meanings and uses of “word.”

Properties of WM

The WM metric provides a quantitatively expressed trajectory of word knowledge development for every separate word in a corpus of potentially any size, from the word’s first encounter to its adult status. This contrasts with current vocabulary measures based on sampled word frequencies or provenances from which they come or for which they are considered suitable. Because it derives characteristic functions (Item Response Theory [IRT] curves1) for every word type, the metric enables unusually broad-range tests, especially adaptive ones. The vectors provide additional dynamic information about a word’s maturing semantics by their changing similarity to other words and to passages.

However, although it is important that a metric measure phenomena of interest, its most essential property is to provide an accurate and consistent yardstick against which a variety of otherwise differently measured but related phenomena can be compared. WM is or closely approaches being a linear distance measure. At any given point in a word’s development, WM measures how close a word’s current information content is to the information content that the word contains for a literate adult. It also simultaneously undergoes developmental changes in its similarity to other words.

The graduated scale based on WM has a zero point and is divided into equal subdivisions—the difference between maturities \(x\) and \(y\) is the same as between

---

1IRT is a psychometric method in which the difficulty of test items and the abilities of test takers are statistically separated (see “Item Response Theory,” 2010). Characteristic functions are logistic curves relating test item difficulty to test-taker performance.
y and z. Equal intervals can be attained by standard psychometric methods such as Rasch IRT scaling (see, e.g., Laufer et al., 2004) but would be prohibitively difficult to create for thousands of words. The calibration is in number of paragraphs cumulatively added to the text corpus on which it is based, simulating equal amounts of reading. Of importance, its values are independent of their causes (see “Kelvin,” 2008). For vocabulary, independence of cause requires that the measure give the same result no matter how a word was learned—for example, by contextual reading, definition memorization, direct instruction, or oral exposure. This makes it possible to compare related phenomena that would otherwise be incommensurate. For example, it could make possible determining whether 3 hr of phonemic awareness training causes words to become more like that of an adult than 4 hr of oral fluency practice.

**LSA Vectors Have the Properties Needed by the Metric**

We first need to estimate how much of the information in text is carried by words alone and how much by the order of the words, which the metric ignores. Landauer (2002) estimated lower limits of approximately 80% being in word-forms alone based on both the ratio of possible combinations of vocabulary words to the number of possible orders of words in sentences, and on the mutual information between human and LSA measures of document similarities. The remaining 20% are presumably attributable to information in syntactic and grammatical interactions between words. Similar proportions of information are evidenced by humanlike simulations of word, passage, and document similarity in automatic essay and summary assessment (Foltz, Gilliam, & Kendall, 2000; Foltz, Laham, & Landauer, 1999; Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005; Landauer & Laham, 2000; Landauer, Laham, & Foltz, 2003), information retrieval within and across languages (Dumais, Letsche, Littman, & Landauer, 1997), tutorial systems (Franzke et al., 2005; Graesser, McNamara, Louwerse, & Cai, 2004) and many more. The fact that LSA agrees so widely with human interpretations of language meaning also implies by statistical induction that it can be expected to hold for passages not encountered in the training corpus to a large degree, an expectation confirmed by its success in the applications just cited.

**APPLICATIONS OF WM**

WM has been used so far only in research laboratories at the University of Colorado and Pearson Knowledge Technologies where it is being developed. For a short preview—before more in the Discussion section—we have used it to construct an adaptive test of individual reader knowledge of each of more than 27,000
individual words, automatically identify words that a given reader probably does not know, and choose sentences and distracters for cloze items. We note that test sentences automatically chosen from a corpus rather than being humanly authored still requires human vetting.

Vocabulary Assessment

It is both obvious and well demonstrated that learning to read depends on learning words and vice versa (Beck, McKeown & Omanson, 1987; Beck, Perfetti, & McKeown, 1982; Perfetti & Hart, 2001; Saragi, Nation, & Meister, 1978; Verhoven & van Leeuwe, 2008; Wagner, 2005). Thus, trying to improve methods for improving vocabulary should also add to our understanding of reading comprehension and its promotion. However, an interlocking set of factors make the effort very challenging. Roughly stated, average high school students need to know 60,000 to 90,000 unique words by the time they graduate, but they encounter only about 20,000 or less outside of school, and only about 200 per year are explicitly taught in school and permanently retained; almost all the rest must come from reading (Anderson & Freebody, 1981; Landauer & Dumais, 1997; Nagy, Herman, & Anderson, 1985; Sternberg, 1987). But to learn vocabulary by reading, readers need to understand what they read (Anderson & Freebody, 1981). And to understand an average sentence they need to know 90 to 98% of its words (Nagy & Scott, 2000).

So, there is a vicious catch-22. In 1987, Nagy and Herman remarked that “the size of the task is such that just teaching more words cannot be seen as the answer” (p. 19). And although the 2000 National Reading Panel (National Institute of Child Health and Human Development, 2000) report concluded that explicitly teaching words has significantly positive effects (Stahl & Fairbanks, 1986; Sternberg, 1987), it remains clear that far too many students still fail to learn enough.

Our proffered solution lies in the fact that the more words readers know, the faster they learn other words. Simulated students who comprehend better increase their vocabularies more by reading. Correspondingly, in LSA, text and word knowledge is mathematically reciprocal. When Landauer and Dumais (1997) replaced some words with nonsense syllables and recalculated the LSA space, the vectors for the nonsense syllables became significantly more like those of the real words when the corpus was larger. In a recent experiment in our lab using the WM metric, we found that the same words, if first encountered after simulated reading of 64,000 paragraphs not containing them, were learned more than twice as well as when they were encountered after only 16,000 paragraphs.

Finally, a meta-analysis by Stahl and Fairbanks (1986) reported that effective explicit direct teaching in school (e.g., explaining word meanings and tutoring
strategy use) resulted in increased knowledge of words with an effect size of approximately .97, thus doubling the number of words learned in a school environment. However, that would seemingly take the annual total vocabulary gain only from about 200 to about 400 words.

WM Versus Word Frequency

Current state of the art in vocabulary assessment is based on testing small samples of selected words and using IRT to obtain word difficulties and reader skill levels. Sometimes much simpler methods are used, a popular example being the YES/NO test (Meara & Buxton, 1987) in which, for example, the 1,000, 3,000, and 10,000 most frequent in common usage are tested by asking test-takers to check whether they know each word, and correcting bias by interleaving nonwords.

However, as Table 1 illustrates, a WM-based simulation found that words that occur with the same total frequency can have maturities that range from very low to very high. The seven example words shown in the table had each appeared 500 ± 38 times in the complete corpus and 50 ± 3 times at an earlier point—frequency ranges of only ± 15%. In contrast, WM for the same words ranged from 21% to 76% of adult level, a range of 55%. The difference is fairly typical. Word frequency measures miss the fact that words of the same frequency vary greatly in how much they have approached their adult status because many other factors affect word learning. In contrast, WM measures the closeness to adult meaning due to any factor that causes changes in LSA similarity to that of a simulated adult. Put differently, WM measures the effect of encountering words (and paragraphs, etc.), not the causes of encountering them. Note, however, because encounter is necessary, WM is necessarily highly correlated with frequency: by an estimate described later, the two share 63% of their variance.

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequencies in a Partial Corpus</th>
<th>WM in the Same Partial Corpus</th>
<th>Frequency in the Full Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>marble</td>
<td>54</td>
<td>.21</td>
<td>485</td>
</tr>
<tr>
<td>sunshine</td>
<td>49</td>
<td>.31</td>
<td>508</td>
</tr>
<tr>
<td>drugs</td>
<td>53</td>
<td>.42</td>
<td>532</td>
</tr>
<tr>
<td>carpet</td>
<td>48</td>
<td>.59</td>
<td>539</td>
</tr>
<tr>
<td>twin</td>
<td>48</td>
<td>.61</td>
<td>458</td>
</tr>
<tr>
<td>earn</td>
<td>53</td>
<td>.70</td>
<td>489</td>
</tr>
<tr>
<td>beam</td>
<td>47</td>
<td>.76</td>
<td>452</td>
</tr>
</tbody>
</table>
THE LSA BASIS FOR WM

Many readers will be familiar with LSA and some of its applications. However, its use here depends on a critical property that has not always been clearly stated. What has been misleading is saying that LSA reflects the frequency or probability with which words appear in the same passages: It doesn’t, at least not directly. LSA starts with data about how words appear in passages but infers something importantly different—how words combine to form meaningful passages, as follows.

LSA uses a matrix algebra technique called SVD, a method for decomposing an object into a combination of its parts. SVD makes the simplifying assumption that the combination is by addition. To put this idea in somewhat more mathematical terms, every component is represented by a different mathematical vector, a string of statistically independent real numbers. For words and passages of text, the string is typically 200 to 1,000 numbers long, 300 usually being close to optimal (Landauer, 2002; Landauer & Dumais, 1997). Another way to state this is that words and paragraphs are represented as points in a 300-dimensional space, a space with 300 orthogonal coordinates. The dimensionality of the space is chosen empirically to maximize the degree to which LSA similarities between paragraphs match human judgments and support other simulations of language phenomena. For intuition here, the reason for the optimum is that with too few dimensions, words don’t differ from each other as much as they do for humans, and if there are too many, they don’t resemble each other as much as they do for humans. The generalization between one word and another is on average most human-like at the optimum (see Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2004; Landauer, 2007).

The rule for assigning the vectors for words and passages can be verbally stated as

\[
\text{the meaning of a paragraph} \approx \text{the meaning of its first word} + \text{its second word} + \ldots + \text{its last word}
\]

To make this true, SVD finds a constant numerical (vector) value for each word such that the equation is true for every word and passage in the corpus to a least squares criterion. The dimensionality is kept the same everywhere LSA is used in WM. The similarity of a word and document, a document and another document, or between two words is typically measured by the cosine between them. Full expositions of the mathematics, algorithms, modes of application, and a wide

\[2\text{The measure of similarity is the cosine, defined as } \frac{(A \ast B)}{(\sqrt{A} \ast \sqrt{B})}, \text{ where } A \text{ and } B \text{ are vectors resulting from element by element multiplication}\]
A NEW METRIC FOR WORD KNOWLEDGE

variety of uses in education and language modeling can be found in Landauer, McNamara, Dennis, and Kintsch (2007) and some tools for using LSA can, as of this writing, be found at http://lsa.colorado.edu.

The most important result is that LSA represents the meaning of passages in a way that is largely independent of what literal words are used to express similar meanings. For an extreme example, consider the following two sentences:

“Cardiac surgeries are quite safe these days.”
“Nowadays, it is not at all risky to operate on the heart.”

These sentences have no words in common and their syntax is markedly different, but their LSA similarity (the cosine between the two combinations of words) is .7, some 6 SDs above chance. (Note that it is hard to construct such good examples, in part because highly synonymous sentences are hard to create, and partly because LSA vectors do not contain the order-dependent information conveyed by syntax and grammar.)

HOW THE WM METRIC ITSELF WORKS

WM is based on LSA vector representations of words that are learned by analysis of a corpus that mirrors the increasing experience with printed language of a typical reader. It simulates how close to adult status a word or a feature thereof is. Here is an overview of the process of producing a calibrated WM scale.

1. Obtain a large and representative text corpus divided into approximate paragraphs.
2. Use SVD to create an LSA “semantic space” for the entire corpus that contains all the paragraphs in the corpus.
3. Create a series of subcorpora by adding portions of the total corpus in an order that approximates the order of text encounters by typical learners of the language, for example, by using text at successive Lexile levels. Lexile measures use combinations of word frequencies and sentence lengths within paragraphs (note, causes of word learning, not results) to align paragraphs consistently with samples of human performance using cloze tests and IRT.
4. At each step, compute a new LSA semantic space for the cumulatively enlarged sample. This is the current space after simulated reading of all n paragraphs so far encountered.
5. Align the vectors of all the paragraphs now in the new space as well as possible with those of the same paragraphs in the larger full “adult” space. This employs the linear algebra method called Procrustes rotation, needed because SVD is good only up to rotation and adding new paragraphs of text requires simultaneous least squares fitting of the new and old data.
6. For each word, compute the current average vector of all the paragraphs now containing it. This is the new vector for the word.
7. Measure the new distance (cos) of each paragraph vector to the vector it will have in the full “adult” space.
8. Divide the distance between the first inserted part of the corpus and the last into equal numbers of paragraphs. The WM scale displays the evolving status of a word by how close its vector is to that of a simulated adult reader in units of equal amount of simulated reading. Individual word development can then be traced as a function of the amount of simulated reading. Linear interpolation between scale points is legal.

Figure 1 traces developmental trajectories of five example words for one simulated reader based on a corpus of 85,000 paragraphs. The trajectories were derived by the cumulative addition of paragraphs from an initial 10,000 paragraphs through 16 additions of 5,000 each. The 10,000 starting paragraph subcorpus is based on an estimate of exposure to written language before school entry, 85,000 a calibration point at about the amount of text that has been read by an average college student. The order of the additions up to 60,000 was based on Lexiles, the technique used by MetaMetrics Inc. to estimate reading difficulties of educational books and periodicals (Smith, Stenner, Horabin, & Smith, 1989). Lexile values are in turn based on the average number of words and their log word frequencies in roughly paragraph-length samples of text as scaled by Rasch Item Response Theory (Rasch, 1961/1980; Smith et al., 1989; Wright & Stone, 1979).
The remainder; from 61,000 to 85,000 paragraphs, are randomly ordered adult readings. Other methods of ordering the text could easily be substituted.

The emulation of order of reading is thus not highly precise, so the simulation will not track the trajectory of individual words and paragraphs perfectly. However, multiple random corpora sampled from a very much larger corpus find only moderate differences. Average trajectories from such multiple samplings give greater stability and vocabulary coverage.

In addition to tracing the overall trajectory of a word, related characteristics can be revealed by examining the similarity of a word to selected other words or to the same words at different stages. A good example is the evolution of the word *turkey* from predominantly about food to mostly about the name of the country straddling western Asia and Eastern Europe. To illustrate, here are cosine similarities between the trajectory value of the word *dinner* at four selected points in Turkey’s early to late maturation: .29, .40, .44, .20. And here for the word *Istanbul* at the same points: .09, .17, .40, .60. Notice that in Figure 1, the trajectory for the word *turkey* rises smoothly except for a plateau at the second to third step. The averaging of all the meanings of a word (which happens more globally for frequency based measures) hides a great deal. However, the combination of WM and LSA offers a tool for examining what is happening inside the word—in its vector representation. One could use the same method to study evolution of the meanings of paragraphs, or to investigate issues such as depth and breadth of word knowledge (Beck, McKeown, & Omanson, 1987).

Also note again that word meanings grow not only by their own contextual occurrences in text but also by even larger effects of the holistic accommodation of new paragraphs by SVD—about three times as important by some LSA simulations (Landauer & Dumais, 1997). Thus, tracing the details of these complex effects with WM might offer a fruitful way to analyze the organic nature of growth in reading comprehension.

There are some other old and new statistical models that also measure similarities between words and documents. In addition to keyword methods, the most prominent include HAL (Burgess, Livesay, & Lund, 1998), which measures word similarities by how they follow other words; Beagle (Jones & Mewhort, 2007), which we have found works about as well as LSA for sentence similarities but is not applicable for larger language units; and the probability-based TOPICS model (Blei et al., 2003; Griffiths & Steyvers, 2004), which measures similarity of words and documents to each other comparably to LSA but is inapplicable because its representations of words would change randomly with each text addition.

**HOW THE METRIC IS ALIGNED WITH HUMAN VOCABULARY DEVELOPMENT**

To align human readers with the simulator’s word knowledge, the simulator’s trajectory for each word in the corpus is represented by a best-fitting logistic growth
curve (technically a two-parameter IRT characteristic function.) Then, an adaptive test estimates the human’s overall WM maturity level by iteratively testing words with higher and lower growth functions. The individual learner’s overall vocabulary skill level is the computed midpoint. Others have used adaptive testing for vocabulary assessment (e.g., Laufer et al., 2004; Vispoel, 1998). New here is ability to separately trace developmental of every word in a large corpus. Finally an individual learner’s probability of knowing an individual word is estimated by how close to adult the learner’s trajectory for the word is at the midpoint of the trajectory for the word itself. The criterion for “knowing” a given word requires a separate parameter: a criterion for how close to its adult status it needs to be to be considered “learned.”

The relation between human vocabulary knowledge and WM has been evaluated by agreement with well-established vocabulary tests. The tests so far used have been designed for young readers; they include the following:

- Peabody Picture Vocabulary Test–III (Maddux, 1999). Rank order correlation, $\rho = 0.76$.
- Kaufman Brief Intelligence Test–II (Kaufman & Kaufman, 1990), $\rho = 0.73$.
- Kaufman Assessment Battery for Children–Expressive Vocabulary (Kaufman & Kaufman, 1985), $\rho = 0.83$.
- Kaufman Assessment Battery for Children–Verbal Knowledge (Kaufman & Kaufman, 1985), $\rho = 0.81$.

Correlations between the psychometric tests and corpus frequency were nearly the same as with WMs: .76, .84, .75 and .83: $M = 0.794$, vs. 0.768. Thus WM here shows the same sensitivity to word frequency as do conventional psychometric tests while measuring the wide differential developmental status and variabilities of words of the same frequency.

WM predicted which words were recommended to be taught in successive school grades by Biemiller, (2008), with a correlation of .51 ($n = 12,912$, $p < .0000$). Note that there is a fairly large amount of noise in Biemiller’s estimates due to the very coarse scaling by grade level, rater unreliability, school and class differences and overlap in endorsed words between grades.

**DISCUSSION**

We have described a new way of measuring vocabulary that provides more information about the evolution of word knowledge than previously feasible. However, we believe that the basis of the instrument and its resulting new capabilities are of more general significance. We begin with a deeper review of the metric’s formal properties, then describe some illustrative novel applications to reading research.
Measurement

Measurement has long been at the core of psychological research on reading. However, the things measured have always been classifiable as either causes or manifestations of knowledge about or the ability to use words, not indices of fundamental properties of a word in the way that a thermometer measures a fundamental property of matter. Good measures of how many words a reader knows abound, as well as ones of how “difficult” they are based on small samples, usually selected by common usage frequency, sometimes applying IRT. What has been missing is a measure based on the property of a word itself that is independent of what caused it to be used or the results of using it. The WM metric is based on a rigorously defined abstract mathematical property, the information represented in a word vector. This definition makes it a candidate fundamental scientific unit, a quantifiable aspect of nature, closely akin to length or weight (see, e.g., “Metric Space,” 2010).

The unit of measurement is a 300-element LSA vector derived by the application of SVD to a standard text corpus. The information content of each unit is determined by its vector length, here 300, and the number of bits in each element, thus converging to a continuous value as the elements and the number of bits per element increases. A calibrated scale based on these units is created by dividing the text corpus into equal cumulative amounts of reading experience, and recomputing every vector and the average distance (cos) between all the vectors of the current set of paragraphs and that of the same paragraphs in the “adult”-level corpus. Thus, the measure of how well a given word is known by a given reader for any reason or by any criterion can be ascertained by determining the point on the scale to which its vector is mathematically closest. Having such a unit and scale opens possibilities for new ways to assess, analyze, and teach vocabulary and reading, as illustrated next.

Further Applications of WM

One example is the adaptive test that yields an unusually broad estimate of the test-taker’s personal average word learning rate and developmental course because it can select test items using the characteristic functions of thousands of words, thus increasing range and precision. The test items are currently YES/NO types (Meara & Buxton, 1987), sometimes replaced or supplemented with cloze tests.

Another application finds words that if not known will make it difficult for a given reader to understand a particular sentence. For example, to ensure comprehension, the editors of a chapter on ecosystems for a middle school reading textbook manually chose six keywords out of the text to test and teach students prior to their reading the chapter. The six keywords chosen by the editors
were organism, photosynthesis, species, nonliving, nutrients, and reproduce. For comparison, WM was used to automatically choose the six best keywords. Its choices were organism, photosynthesis, species, bacteria, ocean, and river. (The probability of randomly picking the three words in common with six tries would be $\sim .003$.)

A third application automatically chooses sentences and distracters to use for cloze items taken from an approximately 500-million-word educational text corpus. WM is used to assure that (a) all the words other than the target are sufficiently known by the reader, and (b) the sentence with the target word omitted has a similar meaning to that of the target word so as to implicitly teach it by context. The distracters were chosen using WM along with other automated constraints, such as POS, n-gram probability, and moderate LSA semantic similarity to both word and sentence. Here are examples of two cloze items thus created. The first one was intended for a low vocabulary student as measured by WM:

\textit{All the living and nonliving things around a _____ and its environment.}

\begin{itemize}
  \item A. organism
  \item B. oxygen
  \item C. algae
\end{itemize}

The second is intended for a high vocabulary student.

\textit{Freshwater habitats can be classified according to the characteristic species of fish found in them, indicating the strong ecological relationship between a _____ and its environment.}

\begin{itemize}
  \item A. organism
  \item B. energy
  \item C. adaptation
\end{itemize}

These literacy-learning tools are all automated to be presented, responded to, and rapidly corrected with minimum need for teacher, textbook, or parental intervention so as to support otherwise too frequent, varied, and highly personalized use.

An especially promising advance would be the ability to identify and insert optimum reading experiences into developmental paths. The possibility comes from the simulation’s property that every passage added to the corpus, and presumably to reader experience, changes the meaning of every word and passage to some extent. The strategy would be to use simulations to predict what individual or kinds of text and their placement in the curriculum will produce desired effects in general and for given readers.

Another useful extension would be creating individual WM trajectories for individual learners. Different people have different interests, experience, and goals, and it would be good to be able to specialize the metric, for example, to science or history vocabulary. These last two ideas are the most technically challenging.
Some Issues Needing More Research

It needs stressing that all of the components of the WM metric, its creation and its applications, are to some extent unfinished or unsettled. We do not present here a finished educational technology, but one ready to be critiqued, evaluated, and iteratively improved. More research and evaluation is needed on several fronts. Most important, the content of the corpus needs to be optimized and standardized, including especially that of the adult calibration point. Because word meanings are always in flux even for adults, simulated word-knowledge levels need to be established across multiple and representative language samples. It is possible that a variety of different versions of the metric will be needed.

We earlier foresaw directly predicting scores on additional psychometrically established tests to better equate them. To that, we would add predicting scores on reading comprehension assessments, correlation with age, school grade, social background, and standardized assessments such as SAT, ACT, and GRE, plus intelligence tests, as well as measuring aspects of text such as complexity. All of these are known to be highly correlated with vocabulary measures based on corpus frequency. Of course empirical field trials for effectiveness and efficacy of the tutorial applications, especially as applied to low-performing Grade 3 to 6 students. Applying the word maturity measure to wider areas of reading comprehension assessment and tutorial goals is an obvious step.

A Conjecture About Wider Importance

We close with the previously mentioned conjecture on the possible long-term effects that WM might help make possible. Hart and Risley (2003) reported a 30 million token word exposure gap at age 3 and increasing gaps between the vocabulary disadvantaged and privileged into high school. Despite valiant national experiments these disparities still exist, as the “fourth-grade slump” gives one evidence. This is surely cautionary. However, we believe that the improved focus, range, quantity, personalization, assessment, and automation that computational modeling can support will be an important contributor to the solution. Beck et al. (2002) estimated that the range of vocabulary at school entry is 2,500 to 5,000 words, which suggests that learning about 650 more words a year for 2 years might bring an average low-starting learner into the normal range. That is less than two more words a day.

Current gains in vocabulary for low starters are hobbled by too little reading, lack of personalization, suboptimal practice distribution, too little and too slow feedback, and too little time on task. We believe that these barriers can be largely overcome by automation and Internet delivery to schools and homes coupled with new assessment and teaching technologies like those described here. We also believe that better measurement of the rate of gain for individual words and
individual learners that WM offers can help vocabulary and reading researchers learn how to better exploit these opportunities.

ACKNOWLEDGMENTS

We thank Jackson Stenner and MetaMetrics Inc. for the generous use of their corpus of Lexiled text. For funding of the work, we thank Marjorie Scardino, CEO, Pearson plc.

REFERENCES


