# A New Yardstick and Tool for Personalized Vocabulary Building

**Thomas K Landauer** **Kirill Kireyev** **Charles Panaccione**

Pearson Education,
Knowledge Technologies

{tom.landauer,kirill.kireyev,charles.panaccione}@pearson.com

## Abstract

The goal of this research is to increase the value of each individual student's vocabulary by finding words that the student doesn't know, needs to, and is ready to learn. To help identify such words, a better model of how well any given word is expected to be known was created. This is accomplished by using a semantic language model, LSA, to track how every word changes with the addition of more and more text from an appropriate corpus. We define the "maturity" of a word as the degree to which it has become similar to that after training on the entire corpus.

An individual student's average vocabulary level can then be placed on the word-maturity scale by an adaptive test. Finally, the words that the student did or did not know on the test can be used to predict what other words the same student knows by using multiple maturity models trained on random samples of typical educational readings. This detailed information can be used to generate highly customized vocabulary teaching and testing exercises, such as Cloze tests.

## 1 Introduction

### 1.1 Why "Vocabulary First"

There are many arguments for the importance of more effective teaching of vocabulary. Here are some examples:

(1) Baker, Simmons, & Kame'enui (1997) found that children who enter school with limited vocabulary knowledge grow much more discrepant over time from their peers who have rich vocabulary knowledge.

(2.) Anderson & Freebody (1981) found that the number of words in student's meaning vocabularies was the best predictor of how well they comprehend text.

(3) An unpublished 1966 study of the correlation between entering scores of Stanford Students on the SAT found the vocabulary component to be the best predictor of grades in every subject, including science.

(4) The number of words students learn varies greatly, from 0.2 to 8 words per day and from 50 to over 3,000 per year. (Anderson & Freebody,1981)

(5) Printed materials in grades 3 to 9 on average contain almost 90,000, distinct word families and nearly 500,000 word forms (including proper names.) (Nagy & Anderson, 1984).

(6) Nagy and Anderson (1984) found that on average not knowing more than one word in a sentence prevented its tested understanding, and that the probability of learning the meaning of a new word by one encounter on average was less than one in ten.

(7) John B. Carroll's (1993) meta-analysis of factor analyses of measured cognitive ability found the best predictor to be tests of vocabulary.

(8) Hart and Risley's large randomized observational study of the language used in households with young children found that the number of words spoken within hearing of a child was associated with a three-fold difference in vocabulary by school entry.

### 1.2 The Challenge

Several published sources and inspection of the number of words taught in recent literacy textbooks and online tools suggest that less than 400 words per year are directly tutored in American schools. Thus, the vast majority of vocabulary must be acquired from language exposure, especially from print because the oral vocabulary of daily living is usually estimated to be about 20,000

words, of which most are known by early school years. But it would obviously be of great value to find a way to make the explicit teaching of vocabulary more effective, and to make it multiply the effects of reading. These are the goals of the new methodologies reported here.

It is also clear that words are not learned in isolation: learning the meaning of a new word requires prior knowledge of many other words, and by most estimates it takes a (widely variable) average of ten encounters in different and separated contexts. (This, by the way, is what is required to match human adult competence in the computational language model used here. Given a text corpus highly similar to that experienced by a language learner, the model learns at very close to the same rate as an average child, and it learns new words as much as four times faster the more old words it knows (Landauer & Dumais, 1997).)

An important aside here concerns a widely circulated inference from the Nagy and Anderson (1984) result that teaching words by presenting them in context doesn't produce enough vocabulary growth to be the answer. The problem is that the experiments actually show only that the inserted target word itself is usually not learned well enough to pass a test. But in the simulations, words are learned a little at a time; exposure to a sentence increases the knowledge of many other words, both ones in the sentence and not. Every encounter with any word in context percolates meaning through the whole current and future vocabulary. Indeed, in the simulator, indirect learning is three to five times as much as direct, and is what accounts for its ability to match human vocabulary growth and passage similarity. Put differently, the helpful thing that happens on encountering an unknown word is not guessing its meaning but its contribution to underlying understanding of language.

However, a vicious negative feedback loop lurks in this process. Learning from reading requires vocabulary knowledge. So the vocabulary-rich get richer and the vocabulary-poor get relatively poorer. Fortunately, however, in absolute terms there is a positive feedback loop: the more words you know, the faster you can learn new ones, generating exponential positive growth. Thus the problem and solution may boil down to increasing the growth parameter for a given student enough to make natural reading do its magic better.

Nonetheless, importantly, it is patently obvious that it matters greatly what words are taught how, when and to which students.

The hypothesis, then, is that a set of tools that could determine what particular words an individual student knows and doesn't, and which ones learned (and sentences understood) would most help other words to be learned by that student might have a large multiplying effect. It is such a toolbox that we are endeavoring to create by using a computational language model with demonstrated ability to simulate human vocabulary growth to a reasonably close approximation. The principal foci are better selection and "personalization" of what is taught and teaching more quickly and with more permanence by application of optimal spacing of tests and practice—into which we will not go here.

### 1.3 Measuring vocabulary knowledge

Currently there are three main methods for measuring learner vocabulary, all of which are inadequate for the goal. They are:

1. Corpus Frequency. Collect a large sample of words used in the domain of interest, for example a collection of textbooks and readers used in classrooms, text from popular newspapers, a large dictionary or the Internet. Rank the words by frequency of occurrence. Test students on a random subset of, say, the 1,000, 2,000 and 5,000 most frequent words, compute the proportion known at each "level" and interpolate and extrapolate. This is a reasonable method, because frequently encountered words are the ones most frequently needed to be understood.
2. Educational Materials. Sample vocabulary lessons and readings over classrooms at different school grades.
3. Expert Judgments. Obtain informed expert opinions about what words are important to know by what age for what purposes.

Some estimates combine two or more of these approaches, and they vary in psychometric sophistication. For example, one of the most sophisticated, the Lexile Framework, uses Rasch scaling (Rasch, 1980) of a large sample of student vocabulary test scores (probability right on a test, holding student ability constant) to create a difficulty measure for sentences and then infers the difficulty of words, in essence, from the average difficulty of the sentences in which they appear.

The problem addressed in the present project goal is that all of these methods measure only the proportion of tested words known at one or more frequency ranges, in chosen school grades or for particular subsets of vocabulary (e.g. "academic" words), and for a very small subset—those tested - some of the words that the majority of a class knows. What they don't measure is exactly which words in the whole corpus a given student knows and to what extent, or which words would be most important for that student to learn.

A lovely analog of the problem comes from Ernst Rothkopf's (1970) metaphor that everyone passes through highly different "word swarms" each day on their way to their (still highly differentiated) adult literacy.

## 2 A new metric: Word Maturity

The new metric first applies Latent Semantic Analysis (LSA) to model how representation of individual words changes and grows toward their adult meaning as more and more language is encountered. Once the simulation has been created, an adaptive testing method can be applied to place individual words on separate growth curves - characteristic functions in psychometric terminology. Finally, correlations between growth curves at given levels can be used to estimate the achieved growth of other words.

### 2.1 How it works in more detail: LSA.

A short review of how LSA works will be useful here because it is often misunderstood and a correct interpretation is important in what follows. LSA models how words combine into meaningful passages, the aspect of verbal meaning we take to be most critical to the role of words in literacy. It does this by assuming that the "meaning" (please bear with the nickname) of a meaningful passage is the sum of the meanings of its words:

```
Meaning of passage =
 {meaning of first wd} +
 {meaning of second word} + …  +
 {meaning of last word}
```

A very large and representative corpus of the language to be modeled is first collected and represented as a term-by-document matrix. A powerful matrix algebra method called Singular Value Decomposition is then used to make every paragraph in the corpus conform to the above objective function—word representations sum to passage representations - up to a best least-squares approximation. A dimensionality-reduction step is performed, resulting in each word and passage meanings represented as a (typically) 300 element real number vector. Note that the property of a vector standing for a word form in this representation is the effect that it has on the vector standing for the passage. (In particular, it is only indirectly a reflection of how similar two words are to each other or how frequently they have occurred in the same passages.) In the result, the vector for a word is the average of the vectors for all the passages in which it occurs, and the vector for a passage is, of course, the average all of its words.

In many previous applications to education, including automatic scoring of essays, the model's similarity to human judgments (e.g. by mutual information measures) has been found to be 80 to 90% as high as that between two expert humans, and, as mentioned earlier, the rate at which it learns the meaning of words as assessed by various standardized and textbook-based tests has been found to closely match that of students. For more details, evaluations and previous educational applications, see (Landauer et al., 2007).

### 2.2 How it works in more detail: Word Maturity.

Taking LSA to be a sufficiently good approximation of human learning of the meanings conveyed by printed word forms, we can use it to track their gradual acquisition as a function of increasing exposure to text representative in size and content of that which students at successive grade levels read.

Thus, to model the growth of meaning of individual words, a series of sequentially accumulated LSA "semantic spaces" (the collection of vectors for all of the words and passages) are created. Cumulative portions of the corpus thus emulate the growing total amount of text that has been read by a student. At each step, a new LSA semantic space is created from a cumulatively larger subset of the full adult corpus.

Several different ways of choosing the successive sets of passages to be added to the training set have been tried, ranging from ones based on readability metrics (such as Lexiles or DRPs) to en-

tirely randomly selected subsets. Here, the steps are based on Lexiles to emulate their order of encounter in typical school reading.

This process results in a separate LSA model of word meanings corresponding to each stage of language learning. To determine how well a word or passage is known at a given stage of learning—a given number or proportion of passages from the corpus—its vector in the LSA model corresponding to a particular stage is compared with the vector of the full adult model (one that has been trained on a corpus corresponding to a typical adult's amount of language exposure). This is done using a linear transformation technique known as Procrustes Alignment to align the two spaces—those after a given step to those based on the full corpus, which we call its "adult" meaning.

Word *maturity* is defined as the similarity of a word's vector at a given stage of training and that at its adult stage as measured by cosine. It is scaled as values ranging between 0 (least mature) and 1 (most mature).

Figure 1 shows growth curves for an illustrative set of words. In this example, 17 successive cumulative steps were created, each containing ~5000 additional passages.

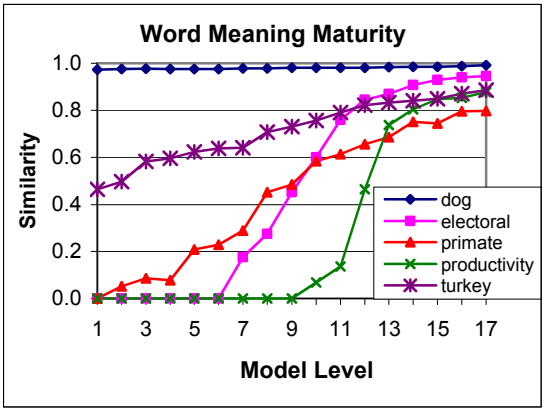


Figure 1. An illustration of meaning maturity growth of several words as a function of language exposure.

Some words (e.g. "dog") are almost at their adult meaning very early. Others hardly get started until later. Some grow quickly, some slowly. Some grow smoothly, some in spurts. Some, like "turkey," grow rapidly, plateau, then resume growing again, presumably due to multiple senses ("Thanksgiving bird" vs. "country") learned at different periods (in LSA, multiple "senses" are combined in a word representation approximately in proportion to their frequency.)

The maturity metric has several conceptual advantages over existing measures of the status of a word's meaning, and in particular should be kept conceptually distinct from the ambiguous and often poorly defined term "difficulty" and from whether or not students in general or at some developmental stage can properly use, define or understand its meaning. It is a mathematical property of a word that may or may not be related to what particular people can do with it.

What it does is provide a detailed view of the course of development of a word's changing representation—its "meaning", reciprocally defined as its effect on the "meaning" of passages in which it occurs,—as a function of the amount and nature of the attestedly meaningful passages in which it has been encountered. Its relation to "difficulty" as commonly used would depend, among other things, on whether a human could use it for some purpose at some stage of development of the word. Thus, its relation to a student's use of a word requires a second step of aligning the student's word knowledge with the metric scaling. This is analogous to describing a runner's "performance" by aligning it with well-defined metrics for time and distance.

It is nevertheless worth noting that the word maturity metric is not based directly on corpus frequency as some other measures of word status are (although its average level over all maturities is moderately highly correlated with total corpus frequency as it should be) or on other heuristics, such as grade of first use or expert opinions of suitability.

What is especially apparent in the graph above is that after a given amount of language exposure, analogous to age or school grade, there are large differences in the maturity of different words. In fact the correlation between frequency of occurrence in a particular one of the 17 intermediate corpora and word maturity is only 0.1, measured over 20,000 random words. According to the model--and surely common sense--words of the same frequency of encounter (or occurrence in a corpus) are far from equally well known. Thus, all methods for "leveling" text and vocabulary instruction based on word frequency must hide a great range of differences.

To illustrate this in more detail, Table 1, shows computed word maturities for a set of words that have nearly the same frequency in the full corpus

(column *four*) when they have been added only 50±5 times *(column two)*. The differences are so large as to suggest the choice of words to teach students in a given school grade would profit much from being based on something more discriminative than either average word frequency or word frequency as found in the texts being read or in the small sample that can be humanly judged. Even better, it would appear, should be to base what is taught to a given student on what that student does and doesn't know but needs to locally and would most profit from generally.

| Word | Occurrences in intermediate corpus (level 5) | Occurrences in adult corpus | Word maturity (at level 5) |
|---|---|---|---|
| marble | 54 | 485 | 0.21 |
| sunshine | 49 | 508 | 0.31 |
| drugs | 53 | 532 | 0.42 |
| carpet | 48 | 539 | 0.59 |
| twin | 48 | 458 | 0.61 |
| earn | 53 | 489 | 0.70 |
| beam | 47 | 452 | 0.76 |

**Table 1** A sample of words with roughly the same number of occurrences in both intermediate (~50) and adult (~500) corpus

The word maturity metric appears to perform well when validated by some external methods. For example, it reliably discriminates between words that were assigned to be taught in different school grades by (Biemiller, 2008), based on a combination of expert judgments and comprehension tests ($p < 0.03$), as shown in Table 2.

| grade 2, known by > 80% | grade 2, known by 40-80% | grade 6, known by 40-80% | grade 6, known by < 40% |
|---|---|---|---|
| n=1034 | n=606 | n=1125 | n=1411 |
| **4.4** | **6.5** | **8.8** | **9.5** |

**Table 2** Average level for each word to reach a 0.5 maturity threshold, for words that are known at different levels by students of different grades (Biemiller, 2008).

Median word maturity also tracks the differences ($p < 0.01$) between essays written by students in different grades as shown in Figure 2.

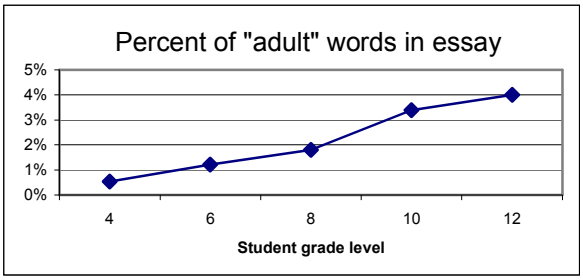


**Figure 2** Percentage of "adult" words used in essays written by students of different grade levels. "Adult" words are defined as words that reach a 0.5 word maturity threshold at or later than the point where half of the words in the language have reached 0.5 threshold.

## 2.3 Finding words to teach individual students

Using the computed word maturity values, a sigmoid characteristic curve is generated to approximate the growth curve of every word in the corpus. A model similar to one used in item response theory (Rasch, 1980) can be constructed from the growth curve due to its similarity in shape and function to an IRT characteristic curve; both curves represent the ability of a student. The characteristic curve for the IRT is needed to properly administer adaptive testing, which greatly increases the precision and generalizeability of the exam. Words to be tested are chosen from the corpus beginning at the average maturity of words at the approximate grade level of the student. Thirty to fifty word tests are used to home in on the student's average word maturity level. In initial trials, a combination of yes/no and Cloze tests are being used. Because our model does not treat all words of a given frequency as equivalent, this alone supports a more precise and personalized measure of a student's vocabulary. In plan, the student level will be updated by the results of additional tests administered in school or by Internet delivery.

The final step is to generalize from the assessed knowledge of words a particular student (let's call her Alice) is tested on to other words in the corpus. This is accomplished by first generating a large number of simulated students (and their word maturity curves) using the method described above. Each simulated student is trained on one of many ~ 12 million word corpora, size and content approximating the lifelong reading of a typical college student, that have been randomly sampled from a representative corpus of more than half a

billion words. Some of these simulated students' knowledge of the words being tested will be more similar to Alice than others. We can then estimate Alice's knowledge of any other word *w* in the corpus by averaging the levels of knowledge of *w* by simulated students whose patterns of tested word knowledge are most similar hers. The method rests on the assumption that there are sufficiently strong correlations between the words that a given student has learned at a given stage (e.g. resulting from Rothkopf's personal "swarms'.) While simulations are promising, empirical evidence as to the power of the approach with non-simulated students is yet to be determined.

## 3 Applying the method

On the assumption that learning words by their effects on passage meanings as LSA does is good, initial applications use Cloze items to simultaneously test and teach word meanings by presenting them in a natural linguistic context. Using the simulator, the context words in an item are predicted to be ones that the individual student already knows at a chosen level. The target words, where the wider pedagogy permits, are ones that are related and important to the meaning of the sentence or passage, as measured by LSA cosine similarity metric, and, ipso facto, the context tends to contextually teach their meaning. They can also be chosen to be those that are computationally estimated to be the most important for a student to know in order to comprehend assigned or student-chosen readings—because their lack has the most effect on passage meanings—and/or in the language in general. Using a set of natural language processing algorithms (such as n-gram models, POS-tagging, WordNet relations and LSA) the distracter items for each Cloze are chosen in such a way that they are appropriate grammatically, but not semantically, as illustrated in the example below.

In summary, Cloze-test generation involves the following steps:

1. Determine the student's overall knowledge level and individual word knowledge predictions based on previous interactions.
2. Find important words in a reading that are appropriate for a particular student (using metrics that include word maturity).
3. For each word, find a sentence in a large collection of natural text, such that the rest of the sentence semantically implies (is related to) the target word and is appropriate for student's knowledge level.
4. Find distracter words that are (a) level-appropriate, (b) are sufficiently related and (c) fit grammatically, but (d) not semantically, into the sentence.

| |
|---|
| All the living and nonliving things around an ___ is its environment.<br>A. organism B. oxygen C. algae |
| Freshwater habitats can be classified according to the characteristic species of fish found in them, indicating the strong ecological relationship between an ___ and its environment.<br>A. adaptation B. energy C. organism |

**Table 3** Examples of auto-generated Cloze tests for the same word (*organism*) and two students of lower and higher ability, respectively.

## 4 Summary and present status

A method based on computational modeling of language, in particular one that makes the representation of the meaning of a word its effect on the meaning of a passage its objective, LSA, has been developed and used to simulate the growth of meaning of individual word representations towards those of literate adults. Based thereon, a new metric for word meaning growth called "Word Maturity" is proposed. The measure is then applied to adaptively measuring the average level of an individual student's vocabulary, presumably with greater breadth and precision than offered by other methods, especially those based on knowledge of words at different corpus frequency. There are many other things the metric may support, for example better personalized measurement of text comprehensibility.

However, it must be emphasized that the method is very new and essentially untried except in simulation. And it is worth noting that while the proposed method is based on LSA, many or all of its functionalities could be obtained with some other computational language models, for example the Topics model. Comparisons with other methods will be of interest, and more and more rigorous evaluations are needed, as are trials with more various applications to assure robustness.

## 5 References


Richard C. Anderson, Peter Freebody. 1981. *Vocabulary Knowledge.* In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77-117). International Reading Association, Newark DE.

Scott K. Baker, Deborah C. Simmons, Edward J. Kameenui. 1997. *Vocabulary acquisition: Research bases.* In Simmons, D. C. & Kameenui, E. J. (Eds.), *What reading research tells us about children with diverse learning needs: Bases and basics.* Erlbaum, Mahwah, NJ.

Andrew Biemiller (2008). Words Worth Teaching. Co-lumbus, OH: SRA/McGraw-Hill.

John B Carroll. 1993. *Cognitive Abilities: A survey of factor-analytic studies.* Cambridge: Cambridge University Press*, 1993.*

Betty Hart, Todd R. Risley. 1995. *Meaningful differences in the everyday experience of young American children.* Brookes Publishing, 1995.

Melanie R. Kuhn, Steven A. Stahl. 1998. *Teaching children to learn word meanings from context: A synthesis and some questions.* Journal of Literacy Research, 30(1) 119-138.

Thomas K Landauer, Susan Dumais. 1997. *A solution to Plato's problem: The Latent Semantic Analysis theory of the Acquisition, Induction, and Representation of Knowledge*. Psychological Review, 104, pp 211-240.

Thomas K Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch. 2007. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum.

Cleborne D. Maddux (1999). Peabody Picture Vocabulary Test III (PPVT-III). Diagnostique, v24 n1-4, p221-28, 1998-1999

William E. Nagy, Richard C. Anderson. 1984. *How many words are there in printed school English?* Reading Research Quarterly*,* 19, 304-330.

Ernst Z. Rothkopf, Ronald D. Thurner. 1970. *Effects of written instructional material on the statistical structure of test essays.* Journal of Educacational Psychology*,* 61, 83-89.

George Rasch. (1980). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.