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.

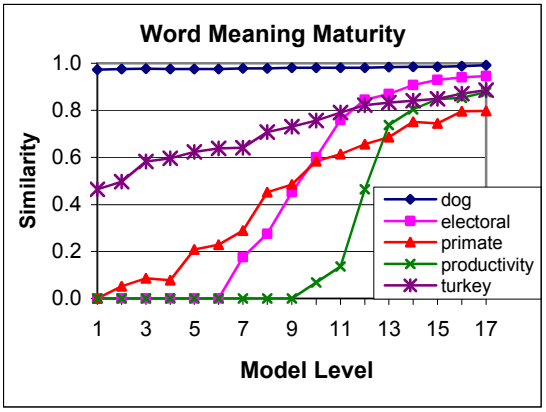

Figure 1. An illustration of meaning maturity growth of several words as a function of language exposure.

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.

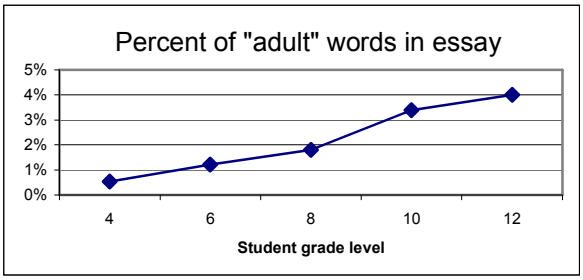

Figure 2 Percentage of "adult" words used in essays written by students of different grade levels. "Adult" words are defined as words that reach a 0.5 word maturity threshold at or later than the point where half of the words in the language have reached 0.5 threshold.

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. 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. 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.