Pearson’s Text Complexity Measure

White Paper

Thomas K. Landauer, Ph.D.

May 2011
About Pearson
Pearson, the global leader in education and education technology, provides innovative print and digital education materials for preK through college, student information systems and learning management systems, teacher licensure testing, teacher professional development, career certification programs, and testing and assessment products that set the standard for the industry. Pearson’s other primary businesses include the Financial Times Group and the Penguin Group. For more information about the Assessment & Information group of Pearson, visit http://www.pearsonassessments.com/.

About Pearson’s White Papers
Pearson’s white paper series shares the perspectives of our assessment experts on selected topics of interest to educators, researchers, policy makers and other stakeholders involved in assessment and instruction. Pearson’s publications in .pdf format may be obtained at: http://www.pearsonassessments.com/research.
Pearson’s Text Complexity Measure

Introduction

Pearson’s Knowledge Technologies group has developed a new measure of text complexity that is fundamentally different from current readability measures such as Flesch–Kincaid (Kincaid et al., 1975), Coleman-Liau (Coleman et al., 1975), Dale-Chall (Chall et al., 1995) and Lexiles (Stenner, 1996). These all rely on a small number of surface features such as the average number of words or characters per sentence, the log of the common usage frequency of the words in the sentences, and a few selected syntactic and grammatical features.

That is not to say that such combinations of features cannot predict readability usefully, especially when averaged over many sentences and passages and combined with good psychometrics. They do. But what we offer here is a new approach that stands on a broader and deeper foundation that enables greater accuracy of measurement and a wider range of analytic capabilities.

As an illustration of the weakness of conventional readability measures that the new approach improves upon, our new metric made it possible to estimate that word frequency accounts on average for only about 34% of the variance in how well a single word is known, varying from 60% for 4th grade readers to only 10% for 1st year college readers. This finding is an example of a language phenomenon of considerable importance for text complexity measurement, that meanings themselves are learned. Word meanings (and thus passage meanings) depend critically on what and how many other words are already known (Anderson & Freebody, 1981). Thus, word (and paragraph) learning is a synergistic and recursive process.

In contrast, the conventional readability metrics ignore these changes, instead treating words as stationary constants that need only be counted. Thus, measuring readability by number of words and their common usage frequency (e.g. Thorndike et al., 1944), and doing it the same way for 4th to 12th graders, misses much of the causality of varying comprehensibility (Landauer, Kireyev & Panaccione, 2011). The text complexity metric overcomes this lack by applying a well established and accurate computational language simulation model, Latent Semantic Analysis (LSA) (Deerwester et al., 1990, Landauer, T. K & Dumais, S. T., 1997), to mimic the way that word (and paragraph) meanings are learned through reading. In doing so, it applies many of the same underlying technologies as Pearson’s automatic reading, writing, speaking and listening assessment technologies (see Foltz et al., 1999), all of which match human judgments 90-100% as well as human judgments match each other.

The new application of the Word Maturity metric to text complexity uses LSA to trace the evolving similarity of words and passages to each other and to those of proficient adult readers as a function of how much and what has been read. Just as for human learners, the evolution is driven by cumulative encounters with new words in new contexts that convey and teach new meanings (e.g. Anderson & Freebody, 1981). By adding new words and contexts in a natural order and quantity, the simulation closely mimics human learning. Just as in human learning, the way that meaning grows for every word (and for every word for every reader) depends on the unique contexts and orders in which each word and passage is encountered.
To follow the same learning paths that human learners do, the simulation uses a very similar number and distribution of words and passages as those encountered by average American readers. Of course every reader reads different things to some extent. The system handles this by creating a large group of independent randomly chosen sets of passages of the same size from a very large text corpus (a collection of texts representing a selected content area, in this case hundreds of thousands of passages and words read by American students and early adults.) This assures that all the words in almost any given reading can be included by substitution.

As explained in the next section, it is the representation of words and passages as mathematical vectors (best thought of here simply as long strings of independent numbers) that enables word and passage meanings to change so as to consistently accommodate all the similarities between each word and each other one.

A particularly important result of this process is that the same words are learned as much as four times as fast if they occur when the amount of text already read is large than if they occur very early (Landauer & Dumais, 1997, Kireyev & Landauer, 2011).

In summary, the most important advance of the new measure lies in the fact it measures how the meanings of words themselves, and thus of passages, change for learners with increasing exposure to language, even though the spelled "word forms" don't change. These developmental changes are not captured at all by the standard methods; in them, words have the same meanings for readers of all ages and reading experience. Thus, although the new measure is as sensitive to frequency of encounter and length of passages as current measures, its estimates of text complexity include the changes of meaning and interrelationships between words with reading experience.

**Mathematical foundations of the measure**

The Word Maturity metric exploits a powerful statistical algorithm, Singular Value Decomposition (see Landauer & Dumais, 1997), to closely simulate the dynamic evolution of word and passage meanings encountered through a student’s life. The algorithm first finds the similarity of the vector standing for every word (and every paragraph) to the vector for every other word (and paragraph) in a corpus of words representative of those encountered by early readers. Then more paragraphs and their words are added to the corpus in a natural order, and the algorithm is rerun to make all the similarities fit together again. To make the items fit, their vector representations must change, and thus the meanings of the words and passages also change.

To measure the development of word knowledge, the vector similarity of each word to that of an average college educated adult is computed at each step. Thus, as more paragraphs and their words are added, the system tracks how close the meaning of each has become to that of an adult. This creates an individual learning curve for each separate word as a function of how much and what typical adults have read.2

An important component of the process is the ordering of the paragraphs added. In the research stage of this effort, published Lexile levels were used, starting with a roughly estimated number of paragraphs that a beginning reader would have read, and adding a series of 15 progressive sets of 5,000 paragraph-like passages. A sample of individual word maturation trajectories created in this manner is shown in Figure 1.
Note, however, that sufficiently similar results have been obtained with other sources used to order the entries, including other current readability measures. Note carefully, however, that it is the resulting highly varied shapes of the individual word trajectories that are important, not just the rising number of total words encountered, whose average would closely resemble the single trajectory that a standard measure of word frequency would provide.

**Figure 1.** Five contrasting examples of Word Maturity trajectories.

The x-axis in Fig. 1 is the number of paragraphs added to the corpus at each step; the y-axis shows how close the individual word meanings are to their simulated meanings for first year college students on a scale of 0 to 1. The word "dog" starts out almost as mature as it will get, "productivity" is unknown until the reader has encountered a total of around 40,000 paragraphs, the others words fall between with markedly different developmental paths--of which current measures are unaware.

The method also makes it possible to investigate in some detail how the meanings of a word change as it matures. This analysis is carried out by comparing a word's current vector to that of related words, which could eventually include, for example, the changing relative contributions of different senses to a word's overall meaning.

It is important to note that the science of word learning has for decades recognized that word learning depends much more strongly on what the reader already knows than on occurrences of the word itself (Landauer & Dumais, 1997) although that is also obviously necessary, and that LSA, and thus Word Maturity, faithfully model that mechanism as demonstrated in many other applications.
Figure 2. Estimating the maturity of a given word or passage

The vertical blue line shows an example average maturity, how near on average the words in a given text have become to their adult meanings as a function of the total number of paragraphs added.

Although the maturity of a word or passage could be assessed in other ways, for example by fitting logistic curves to each word trajectory, the current and easiest to understand is by just measuring how far each word in a text is from its adult level. From this both the average level of the text, shown as the vertical blue line in Figure 2, and of single maturity levels by where they intersect that line. Thus, for example, one could average all the growth curve heights or pick the least mature word in the text, the one that intersects nearest the bottom of the graph. For a real example, a recent evaluation under the Gates Foundation sponsored Student Achievement Partners collaboration, used an optimally weighted combination of the average maturity and the values of the ten "most difficult" words, the ones farthest from maturity.

Currently, a number of other variables are added to the Text Complexity measure, some identical or similar to the standard ones and others reflecting various other linguistic features--at this point most as constants--to account for other factors than word and paragraph maturity. Adding these features increases overall prediction accuracy modestly. However, importantly, because many of their effects on complexity also change with increasing occurrences in ever-changing contexts, in the future maturities of other features may also be tracked and integrated into the overall complexity measure, presumably increasing accuracy still more.
How well the metric assesses text complexity

The new measure has been evaluated against a number of text data sets of varying kinds in terms of both text content and criterion measures. The criterion measures include the reported school grades of texts used in NCLB reading comprehension tests and a widely used Item Response Theory (IRT)-based measure (the Stanford9) based on text comprehension questions. Table 1 below lists the data sets, their accuracy criteria and the number of text passages involved.

For these five data sets, text complexity measures were compared against the listed criteria.

<table>
<thead>
<tr>
<th>Description of Text data</th>
<th>Performance criterion</th>
<th>Number of texts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Summary Street:</strong> Randomly chosen whole texts from a large set of readings used in Pearson K-12 tutorial reading passages, the vast majority of which are informative text, all used in Pearson’s WriteToLearn® reading comprehension and iterative summarization formative tutorial exercises</td>
<td>Assigned grade level of readings</td>
<td>800</td>
</tr>
<tr>
<td><strong>State Readings:</strong> A Collection of diverse texts used as items in NCLB reading comprehension tests in 27 state and two national assessments.</td>
<td>Assigned grade level of readings</td>
<td>1221</td>
</tr>
<tr>
<td><strong>SAT9:</strong> Texts used in the well established Stanford9 reading comprehension test</td>
<td>Grade levels of use and Rasch IRT-based measures of student comprehension</td>
<td>110</td>
</tr>
<tr>
<td><strong>Common Core:</strong> Example texts from the Common Core State Standards Appendix B. selected by reading experts for use in varying grade levels</td>
<td>Appropriate grade levels according to experts</td>
<td>243</td>
</tr>
<tr>
<td><strong>NAEP:</strong> NAEP reading comprehension passages from NAEP 4,8, and 12th grade assessments</td>
<td>Grade levels of the test takers</td>
<td>39</td>
</tr>
</tbody>
</table>

Table 1. Data, criteria, number of texts
Table 2 lists Pearson correlations with the criterion variables for Pearson’s Text Complexity Measure and for two standard text complexity measures, Coleman-Liau and Flesch grade-level, and an optimal combination the two most general and important factors in readability formulas.

<table>
<thead>
<tr>
<th>Readings</th>
<th>Pearson’s Text Complexity Measure</th>
<th>Flesch Grade</th>
<th>Coleman-Liau</th>
<th>Word freq and sentence length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary Street readings</td>
<td>0.881</td>
<td>0.637</td>
<td>0.636</td>
<td>0.609</td>
</tr>
<tr>
<td>n = 800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>State readings</td>
<td>0.862</td>
<td>0.591</td>
<td>0.598</td>
<td>0.604</td>
</tr>
<tr>
<td>n = 1221</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT9 readings (vs. Grade)</td>
<td>0.803</td>
<td>0.673</td>
<td>0.732</td>
<td>0.566</td>
</tr>
<tr>
<td>n = 111</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT9 readings (vs. IRT)</td>
<td>0.844</td>
<td>0.678</td>
<td>0.763</td>
<td>0.685</td>
</tr>
<tr>
<td>n = 110</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Common Core readings</td>
<td>0.736</td>
<td>0.485</td>
<td>0.508</td>
<td>0.545</td>
</tr>
<tr>
<td>n = 243</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAEP readings</td>
<td>0.783</td>
<td>0.286</td>
<td>0.217</td>
<td>0.086</td>
</tr>
<tr>
<td>n = 39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. The first column lists the readings used to evaluate Pearson’s measure of text complexity, the second the correlations between them and Pearson’s Text Complexity measure. The third and fourth columns show correlations with the texts of two widely used traditional readability measures. The last column is for a simple optimized combination of the average word frequency and number of words per sentence.

Thus, on average, the text complexity measure correlated with criterion measures about a third better than the traditional measures.

Measures of the method’s validity

Text complexity is not yet a well-defined concept. For example, almost all of the measures thereof that researchers have put forward have assessed difficulties at the word and sentence levels and evaluated results either by correlation with grade levels in which they are used or placed on an abstract IRT scale. However, the ability to understand complex text is surely not limited to the skills tapped by the meanings of words and passages alone, as important as these may be. Other cognitive abilities must surely play a part. For one example, a goal of continuing Pearson’s text complexity research is to include difficulties caused by the order in which information is delivered, which is most obviously critical in math and science texts, but can also make narratives hard to follow even if almost all the words are well known and most of the sentences short. Other examples include logic, reasoning, argument resolution, specialized and quotidian world knowledge, and much of what goes under the rubric of intelligence.
Thus we will need more research on the effects on comprehension caused by a larger range of variables before we can measure validity properly. In the interim, however, to shed a modicum of light on the issue, a concurrent validity study was conducted of the correlation between word maturity on the scale described above and four well validated measures of abilities, each with somewhat different face-value conceptual relations to text complexity.

Described below are the test names, brief descriptions of the tests, the number of items in the correlations, and correlations of the tests item difficulties. (Spearman rho rather than Pearson’s r is used to increase the comparability of the various scales.)

<table>
<thead>
<tr>
<th>Test</th>
<th>Task Description</th>
<th>Number of items in test</th>
<th>Spearman correlation with item difficulties</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Peabody Picture Vocabulary Test, Fourth Edition (PPVT™-4)</strong> (Maddux, 1999)</td>
<td>Receptive (choose correct picture) and expressive (name picture) vocabulary test for ages 2.6-90 years</td>
<td>192</td>
<td>0.74</td>
</tr>
<tr>
<td><strong>Kaufman Brief Intelligence Test, Second Edition (KBIT-2)</strong> (Kaufman et al., 1990)</td>
<td>Verbal and non-verbal multiple choice cognitive ability score for ages 4-90 years</td>
<td>41</td>
<td>0.82</td>
</tr>
<tr>
<td><strong>Kaufman Assessment Battery for Children – Expressive Vocabulary (KABC-EV)</strong> (Kaufman et al., 1985)</td>
<td>Expressive vocabulary scale for ages 3-18 years</td>
<td>35</td>
<td>0.72</td>
</tr>
<tr>
<td><strong>Kaufman Assessment Battery for Children – Verbal Knowledge (KABC-VK)</strong></td>
<td>Updated version of KABC-II</td>
<td>45</td>
<td>0.75</td>
</tr>
</tbody>
</table>

*Table 3. Correlations with some of Pearson’s well-established tests conceptually related to text complexity. Without being specifically designed to assess "text complexity", these tests share a common concept, the variability of comprehensibility of different words. Thus, since they measure a variety of somewhat different concepts, both to each other and to text complexity, their moderately high correlation with Pearson’s text complexity measure seems as great as one could expect (or want) and lends considerable evidence of its validity.*

**In progress and planned extensions and revisions**

Pearson plans to exploit the capabilities offered by the Word Maturity metric in a variety of ways. The following are in early developmental stages.

- A small set of well-differentiated components of text complexity that teachers, parents and other educators will find intuitive. For example, a set of four such components might be *word difficulty, sentence difficulty, passage difficulty* and *overall comprehension difficulty*. (We note that results of complex measurements too often take the form of equally complex or arbitrary names, which are avoided.)
- Separate Text Complexity measures based on corpora restricted to particular genres, e.g. informative or narrative, or to different school and college subjects, e.g. biology and psychology.

- A measure of text comprehension difficulty due to the flow and order in which information is conveyed, which, for example, will be equally important for science and literature.

- Maturity trajectories for separate senses of words.

**Where from here?**

We think that the next step should be using the new technology to teach in addition to assessing. Given the enlarged ability to pinpoint the comprehension problems that texts present, it will be easier to choose and target the instruction we deliver. Indeed, we have already developed a companion measure of word and passage text complexity growth that assesses individual reader knowledge of an unlimited number of individual words using the same basic technology, as well as an *age-independent* vocabulary measure that, like Text Complexity, separately estimates knowledge of every word in an unlimited corpus, but here for any individual student. This will make it possible to teach specifically those words that a reader currently does not know and needs to learn.

**Endnotes**

1 These are "meanings" in the sense that "meaning" is used in LSA, which in turn agree with human judgments of the similarity of meaning of passages 90% as well as two humans agree with each other (Landauer, T. K. (2002)

2 This requires the constraint that there are be always the same number of independent dimensions, that is the 300 real numbers for each word and paragraph.

**Acknowledgements**

Evaluations and planning of our methods were greatly aided by a Gates Foundation sponsored discussion and collaboration group directed by Meredith and David Liben from Student Achievement Partners and chaired by David Coleman.
References


