A Two-Stage Approach for Generating Unbiased Estimates of Text Complexity

Kathleen M. Sheehan **Michael Flor** **Diane Napolitano**

Educational Testing Service
Princeton, NJ, USA
{ksheehan, mflor, dnapolitano}@ets.org

Abstract

Many existing approaches for measuring text complexity tend to overestimate the complexity levels of informational texts while simultaneously underestimating the complexity levels of literary texts. We present a two-stage estimation technique that successfully addresses this problem. At Stage 1, each text is classified into one or another of three possible genres: informational, literary or mixed. Next, at Stage 2, a complexity score is generated for each text by applying one or another of three possible prediction models: one optimized for application to informational texts, one optimized for application to literary texts, and one optimized for application to mixed texts. Each model combines lexical, syntactic and discourse features, as appropriate, to best replicate human complexity judgments. We demonstrate that resulting text complexity predictions are both unbiased, and highly correlated with classifications provided by experienced educators.

1 Introduction

Automated text analysis systems, such as readability metrics, are frequently used to assess the probability that texts with varying combinations of linguistic features will be more or less accessible to readers with varying levels of reading comprehension skill (Stajner, Evans, Orasan and Mitkov, 2012). This paper introduces TextEvaluator, a fully-automated text analysis system designed to facilitate such work.[1] TextEvaluator successfully addresses an important limitation of many existing readability metrics: the tendency to over-predict the complexity levels of informational texts, while simultaneously under-predicting the complexity levels of literary texts (Sheehan, Kostin & Futagi, 2008; Sheehan, Kostin, Futagi & Flor, 2010). We illustrate this phenomenon, and argue that it results from two fundamental differences between informational and literary texts: (a) differences in the way that common every-day words are used and combined; and (b) differences in the rate at which rare words are repeated.

Our approach for addressing these differences can be summarized as follows. First, a large set of lexical, syntactic and discourse features is extracted from each text. Next, either human raters, or an automated genre classifier is used to classify each text into one or another of three possible genre categories: informational, literary, or mixed. Finally, a complexity score is generated for each text by applying one or another of three possible prediction models: one optimized for application to informational texts, one optimized for application to literary texts, and one optimized for application to mixed texts. We demonstrate that resulting complexity measures are both unbiased, and highly correlated with text grade level (GL) classifications provided by experienced educators.

[1] TextEvaluator was previously called SourceRater.

Proceedings of the 2th Workshop of Natural Language Processing for Improving Textual Accessibility (NLP4ITA), pages 49–58,
Atlanta, Georgia, 14 June 2013. ©2013 Association for Computational Linguistics

Our paper is organized as follows. Section 2 summarizes related work on readability assessment. Section 3 describes the two corpora assembled for use in this study, and outlines how genre and GL classifications were assigned. Section 4 illustrates the problem of genre bias by considering the specific biases detected in two widely-used readability metrics. Section 5 describes the TextEvaluator features, methods and results. Section 6 presents a summary and discussion.

2 Related Work

Despite the large numbers of text features that may potentially contribute to the ease or difficulty of comprehending complex text, many widely-used readability metrics are based on extremely limited feature sets. For example, the Flesch-Kincaid GL score (Kincaid, et al., 1975), the FOG Index (Gunning, 1952), and the Lexile Framework (Stenner, et al., 2006) each consider just two features: a single measure of syntactic complexity (average sentence length) and a single measure of lexical difficulty (either average word length in syllables, average frequency of multi-syllable words, or average word familiarity estimated via a word frequency, WF, index).

Recently, more computationally sophisticated modeling techniques such as Statistical Language Models (Si and Callan, 2001; Collins-Thompson and Callan, 2004, Heilman, et al., 2007, Pitler and Nenkova, 2008), Support Vector Machines (Schwarm and Ostendorf, 2005), Principal Components Analyses (Sheehan, et al., 2010) and Multi-Layer Perceptron classifiers (Vajjala and Meurers, 2012) have enabled researchers to investigate a broader range of potentially useful features. For example: Schwarm and Ostendorf (2005) demonstrated that vocabulary measures based on trigrams were effective at distinguishing articles targeted at younger and older readers; Pitler and Nenkova (2008) reported improved validity for measures based on the likelihood of vocabulary and the likelihood of discourse relations; and Vajjala and Meurers (2012) demonstrated that features inspired by Second Language Acquisition research also contributed to validity improvements. Importantly, however, while this research has contributed to our understanding of the types of text features that may cause texts to be more or less comprehensible, evaluations focused on the presence and degree of genre bias have not been reported.

3 Corpora

Two text collections are considered in this research. Our training corpus includes 934 passages selected from a set of previously administered standardized assessments constructed to provide valid and reliable feedback about the types of verbal reasoning skills described in U.S. state and national assessment frameworks. Human judgments of genre (informational, literary or mixed) and GL (grades 3-12) were available for all texts. Genre classifications were based on established guidelines which place texts structured to inform or persuade (e.g., newspaper text, excerpts from science or social studies textbooks) in the informational category, and texts structured to provide a rewarding literary experience (e.g., folk tales, short stories, excerpts from novels) in the literary category (see American Institutes for Research, 2008). We added a Mixed category to accommodate texts classified as incorporating both informational and literary elements. Nelson, Perfetti, Liben and Liben (2012) describe an earlier, somewhat smaller version of this dataset. We added additional passages downloaded from State Department of Education web sites, and from the National Assessment of Educational Progress (NAEP). In each case, GL classifications reflected the GLs at which passages were administered to students. Thus, all passages classified at Grade 3 appeared on high-stakes assessments constructed to provide evidence of student performance relative to Grade 3 reading standards.

Two important characteristics of this dataset should be noted. First, unlike many previous corpora, (e.g., Stenner, et al., 2006; Zeno, et al., 2005) accurate paragraph markings are included for all texts. Second, while many of the datasets considered in previous readability research were comprised entirely of informational text (e.g., Pitler and Nenkova, 2008; Schwarm and Ostendorf, 2005; Vajjala and Meurers, 2012) the current dataset covers the full range of text types considered by teachers and students in U.S. classrooms.

Table 1 shows the numbers of informational, literary and mixed training passages at each targeted GL. Passage lengths ranged from 112 words at Grade 3, to more than 2000 words at Grade 12.

Average passage lengths were 569 words and 695 words in the informational and literary subsets, respectively.

Grade Level	Genre			Total
	Inf.	Lit.	Mixed	
3	46	60	8	114
4	51	74	7	132
5	44	46	12	102
6	41	40	6	87
7	36	58	6	100
8	70	63	18	151
9	23	23	2	48
10	26	49	2	77
11	15	24	0	39
12	47	15	22	84
Total	399	452	83	934

Table 1. Numbers of passages in the model development/training dataset, by grade level and genre.

A validation dataset was also constructed. It includes the 168 texts that were published as Appendix B of the new Common Core State Standards (CCSSI, 2010), a new standards document that has now been adopted in 46 U.S. states. Individual texts were contributed by teachers, librarians, curriculum experts, and reading researchers. GL classifications are designed to illustrate the "staircase of increasing complexity" that teachers and test developers are being encouraged to replicate when selecting texts for use in K-12 instruction and assessment in the U.S. The staircase is specified in terms of five grade bands: Grades 2-3, Grades 4-5, Grades 6-8, Grades 9-10 or Grades 11+. Table 2 shows the numbers of informational, literary and "Other" texts (includes both Mixed and speeches) included at each grade band.

Grade Band	Genre			Total
	Inf.	Lit.	Other	
2-3	6	10	4	20
4-5	16	10	4	30
6-8	12	16	13	41
9-10	12	10	17	39
11+	8	10	20	38
Total	54	56	58	168

Table 2. Numbers of passages in the validation dataset, by grade band and genre.

4 Genre Bias

This section examines the root causes of genre bias. We focus on two fundamental differences between informational and literary texts: differences in the types of vocabularies employed, and differences in the rate at which rare words are repeated. These differences have been examined in several previous studies. For example, Lee (2001) documented differences in the use of "core" vocabulary within a corpus of informational and literary texts that included over one million words downloaded from the British National Corpus. Core vocabulary was defined in terms of a list of 2000 common words classified as appropriate for use in the dictionary definitions presented in the Longman Dictionary of Contemporary English. The analyses demonstrated that core vocabulary usage was higher in literary texts than in informational texts. For example, when literary texts such as fiction, poetry and drama were considered, the percent of total words classified as "core" vocabulary ranged from 81% to 84%. By contrast, when informational texts such as science and social studies texts were considered, the percent of total words classified as "core" vocabulary ranged from 66% to 71%. In interpreting these results Lee suggested that the creativity and imaginativeness typically associated with literary writing may be less closely tied to the type or level of vocabulary employed and more closely tied to the way that core words are used and combined. Note that this implies that an individual word detected in a literary text may not be indicative of the same level of processing challenge as that same word detected in an informational text.

Differences in the vocabularies employed within informational and literary texts, and subsequent impacts on readability metrics, are also discussed in Appendix A of the Common Core State Standards (CCSSI, 2010). The tendency of many existing readability metrics to underestimate the complexity levels of literary texts is described as follows: "The Lexile Framework, like traditional formulas, may underestimate the difficulty of texts that use simple, familiar language to convey sophisticated ideas, as is true of much high-quality fiction written for adults and appropriate for older students" (p. 7).

Genre bias may also result from genre-specific differences in word repetition rates. Hiebert and

Mesmer (2013, p.46) describe this phenomenon as follows: “Content area texts often receive inflated readability scores since key concept words that are rare (e.g., *photosynthesis*, *inflation*) are often repeated which increases vocabulary load, even though repetition of content words can support student learning (Cohen & Steinberg, 1983)”.

Table 3 provides empirical evidence of these trends. The table presents mean GL classifications estimated conditional on mean WF scores, for the informational (*n* = 399) and literary (*n* = 452) passages in our training dataset. WF scores were generated via an in-house WF index constructed from a corpus of more than 400 million word tokens. The corpus includes more than 17,000 complete books, including both fiction and nonfiction titles.

	Informational			Literary		
Avg. WF	N	GL	SD	N	GL	SD
51.0–52.5	2	12.0	0.0	0	--	--
52.5–54.0	16	10.8	1.9	0	--	--
54.0–55.5	68	9.6	2.0	1	10.0	--
55.5–57.0	89	7.8	2.7	18	9.9	1.9
57.0–58.5	96	6.6	2.3	46	9.2	2.0
58.5–60.0	78	5.3	1.8	92	7.6	2.4
60.0–61.5	44	4.6	1.8	142	6.2	2.4
61.5–63.0	6	3.7	0.8	119	5.5	2.1
63.0–64.5	0	--	--	31	4.5	1.9
64.5–66.0	0	--	--	3	4.0	1.7
Total	399	57.4	2.1	452	60.6	1.9

Table 3. Mean GL classifications, by Average WF score, for informational and literary passages targeted at readers in grades 3 through 12.

The results in Table 3 confirm that, consistent with expectations, texts with lower average WF scores are more likely to appear on assessments targeted at older readers, while texts with higher average WF scores are more likely to appear on assessments targeted at younger readers. But note that large genre differences are also present. Figure 1 provides a graphical representation of these trends. Results for informational texts are plotted with a solid line; those for literary texts are plotted with a dashed line. Note that the literary curve appears above the informational curve throughout the entire observed range of the data. This suggests that a given value of the Average WF measure is indicative of a *higher* GL classification if the text in question is a literary text, and a *lower* GL classification if the text in question is an informational text. Since a readability measure that includes this feature (or a feature similar to this feature) without also accounting for genre effects will tend to yield predictions that fall *between* the two curves, resulting GL predictions will tend to be too high for informational texts (positive bias) and too low for literary texts (negative bias).

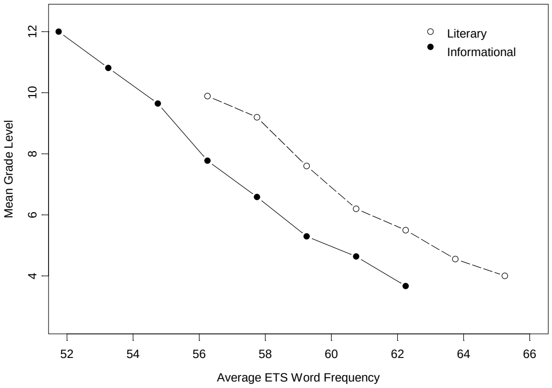

Figure 1. Mean text GL plotted conditional on average WF score. (One literary mean score based on evidence from a single text is not plotted.)

Figure 2 confirms that this evidence-based prediction holds true for two widely-used readability metrics: the Flesch-Kincaid GL score and the Lexile Framework[2]. Each individual plot compares Flesch-Kincaid GL scores (top row), or Lexile scores (bottom row) to the human GL classifications stored in our training dataset, i.e., classifications that were developed and reviewed by experienced educators, and were subsequently used to make high-stakes decisions about students and teachers, e.g., requiring students to repeat a grade rather than advancing to the next GL. The plots confirm that, in each case, the predicted pattern of over- and under-estimation is present. That is, on average, both Flesch-Kincaid scores and Lexile scores tend to be slightly too high for informational texts, and slightly too low for literary texts, thereby calling into doubt any cross-genre comparisons.

[2] All Lexile scores were obtained via the Lexile Analyzer available at www.lexile.com. Scores are only available for a subset of texts since our training corpus included just 548 passages at the time that these data were collected. Corresponding human GL classifications were approximately evenly distributed across grades 3 through 12.

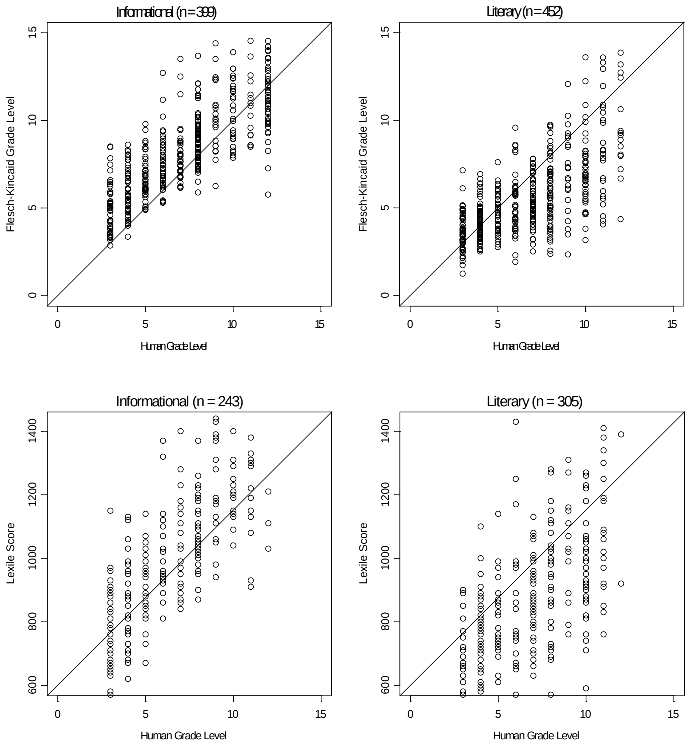

Figure 2. Passage complexity scores generated via the Flesch-Kincaid GL score (top) and the Lexile Framework (bottom) compared to GL classifications provided by experienced educators.

5 Features, Components and Results

5.1 Features

The TextEvaluator feature set is designed to measure the ease or difficulty of implementing four types of processes believed to be critically involved in comprehending complex text: (1) processes involved in word recognition and decoding, (2) processes associated with using relevant syntactic knowledge to assemble words into meaningful propositions, (3) processes associated with inferring connections across propositions or larger sections of text, and (4) processes associated with using relevant prior knowledge and experience to develop a more complete, more integrated mental representation of a text. (See Kintsch, 1998).

A total of 43 candidate features were developed. Since many of these were expected to be moderately inter-correlated, a Principal Components Analysis (PCA) was used to locate clusters of features that exhibited high within-cluster correlation and low between-cluster correlation. Linear combinations defined in terms of the resulting feature clusters provided the independent variables considered in subsequent investigations. Biber and his colleagues (2004) justify this approach by noting that, because many important aspects of text variation are not well captured by individual linguistic features, investigation of such characteristics requires a focus on "constellations of co-occurring linguistic features" as opposed to individual features (p. 45).

The PCA suggested that more than 60% of the variation captured by the full set of 43 features could be accounted for via a set of eight component scores, where each component is estimated as a linear combination of multiple correlated features, and only 3 of the 43 features had moderately high loadings on more than one component, and most loadings exceeded 0.70. The individual features comprising each component are described below.

Component #1: Academic Vocabulary. Ten features loaded heavily on this component. Two are based on the *Academic Word List* described in Coxhead (2000). These include: the frequency per thousand words of all words on the Academic Word List, and the ratio of listed words to total words. In a previous study, Vajjala and Meurers (2012) demonstrated that the ratio of listed words to total wards was very effective at distinguishing texts at lower and higher levels in the Weekly Reader corpus. Two additional features focus on the frequency of nominalizations, including one estimated from token counts and one estimated from type counts. Four additional features are based on word lists developed by Biber and his colleagues. These include the frequency per thousand words of academic verbs, abstract nouns, topical adjectives and cognitive process nouns (see Biber, 1986, 1988; and Biber, et al., 2004). Two measures of word length also loaded on this dimension: average word length measured in syllables, and the frequency per thousand words of words containing more than 8 characters.

Component #2: Syntactic Complexity. Seven features loaded heavily on this component. These include features determined from the output of the Stanford Parser (Klein and Manning, 2003), as well as more easily computed measures such as average sentence length, average frequency of long sentences (>= 25 words), and average number of words between punctuation marks (commas, semicolons, etc.). Parse-based features include average number of dependent clauses, and an automated version of the word "depth" measure introduced by Yngve (1960). This last feature, called Average Maximum Yngve Depth, is designed to capture variation in the memory load imposed by sentences with varying syntactic structures. It is estimated by first assigning a depth classification to each word in the text, then determining the maximum depth represented within each sentence, and then averaging over resulting sentence-level estimates to obtain a passage-level estimate. Several studies of this word depth measure have been reported. For example, Bormuth (1964) reported a correlation of -0.78 between mean word depth scores and cloze fill-in rates provided by Japanese EFL learners.

Component #3: Concreteness. Words that are more concrete are more likely to evoke meaningful mental images, a response that has been shown to facilitate comprehension (Coltheart, 1981). Alderson (2000) argued that the level of concreteness present in a text is a useful feature to consider when evaluating passages for use on reading assessments targeted at L2 readers. A total of five concreteness and imageability measures loaded heavily on this dimension. All five measures are based on concreteness and imageability ratings downloaded from the MRC psycholinguistic database (Coltheart, 1981). Ratings are expressed on a 7 point scale with 1 indicating least concrete, or least imageable, and 7 indicating most concrete or most imageable.

Component #4: Word Unfamiliarity. This component summarizes variation detected via six different features. Two features are measures of average word familiarity: one estimated via our inhouse WF Index, and one estimated via the TASA WF Index (see Zeno, et al., 1995). Both features have negative loadings, suggesting that the component is measuring vocabulary difficulty as opposed to vocabulary easiness. The other features with high loadings on this component are all measures of rare word frequency. These all have positive loadings since texts with large numbers of rare words are expected to be more difficult. Two types of rare word indices are included: indices based on token counts and indices based on type counts. Vocabulary measures based on token counts view each new word as an independent comprehension challenge, even when the same word occurs repeatedly throughout the text. By contrast, vocabulary measures based on type counts assume that a passage containing five different unfamiliar words may be more challenging than a passage containing the same unfamiliar word repeated five times. This difference is consistent with the notion that each repetition of an unknown word provides an additional opportunity to connect to prior knowledge (Cohen & Steinberg, 1983).

Component #5: Interactive/Conversational Style. This component includes the frequency per thousand words of: conversation verbs, fiction verbs, communication verbs, 1st person plural pronouns, contractions, and words enclosed in quotes. Verb types were determined from one or more of the following studies: Biber (1986), Biber (1988), and Biber, et al. (2004).

Component #6: Degree of Narrativity. Three features had high positive loadings on this dimension: Frequency of past perfect aspect verbs, frequency of past tense verbs and frequency of 3rd person singular pronouns. All three features have previously been classified as providing positive evidence of the degree of narrativity exhibited in a text (see Biber, 1986 and Biber, 1988).

Component #7: Cohesion. Cohesion is that property of a text that enables it to be interpreted as a "coherent message" rather than a collection of unrelated clauses and sentences. Halliday and Hasan (1976) argued that readers are more likely to interpret a text as a "coherent message" when certain observable features are present. These include repeated content words and explicit connectives. The seventh component extracted in the PCA includes three different types of cohesion features. The first two features measure the frequency of content word repetition across adjacent sentences within paragraphs. These measures differ from the cohesion measures discussed in Graesser et al. (2004) and in Pitler and Nenkova (2008) in that a psychometric linking procedure is used to ensure that results for different texts are reported on comparable scales (See Sheehan, in press). The frequency of causal conjuncts (*therefore, consequently,* etc.) also loads on this dimension.

Component #8: Argumentation. Two features have high loadings on this dimension: the frequency of concessive and adversative conjuncts (*although, though, alternatively, in contrast,* etc.), and the frequency of negations (*no, neither,* etc.), Just and Carpenter, (1987).

5.2 An Automated Genre Classifier

A preliminary automated genre classifier was developed by training a logistic regression model to predict the probability that a text is classified as *informational* as opposed to *literary*. A significant positive coefficient was obtained for the Academic Vocabulary component defined above, suggesting that a high score on this component may be interpreted as an indication that the text is more likely to be informational. Significant negative coefficients were obtained for Narrativity, Interactive/Conversational Style, and Syntactic Complexity, indicating that a high score on any of these components may be interpreted as an indication that the text is more likely to be literary. Two individual features that were not included in the PCA were also significant: the proportion of adjacent sentences containing at least one overlapping stemmed content word, and the frequency of 1st person singular pronouns. These features were not included in the PCA because they are not reliably indicative of differences in text complexity (See Sheehan, in press; Pitler and Nenkova, 2008.) Results confirmed, however, that these features are useful for predicting a text's genre classification.

Alternative decision rules based on this model were investigated. Table 4 summarizes the levels of precision (P), recall (R) and F1 = 2RP/(R+P) obtained for the selected decision rule which was defined as follows: Classify as informational if P(Inf) >= 0.52, classify as literary if P(inf) < 0.48, else classify as mixed. This decision rule is defined such that few texts are classified into the mixed category since, at present, the training dataset includes very few mixed texts. The table shows decreased precision in the Validation dataset since many more mixed texts are included, and the majority of these were classified as informational.

Dataset	Genre	N	R	P	F1
Training	Inf	399	.84	.79	.81
Training	Lit	452	.88	.79	.83
Training	Mixed	83	.01	.09	.01
Validation	Inf	67	.91	.56	.69
Validation	Lit	56	.80	.80	.80
Validation	Mixed	45	.07	1.0	.13

Table 4. Levels of Precision, Recall and F1 obtained for 1, 089 texts in the training and validation datasets. Speeches are not included in this summary.

5.3 Prediction Equations

We use separate genre-specific regression models to generate GL predictions for texts classified as informational, literary, or mixed. The coefficients estimated for informational and literary texts are shown in Table 5. Note that each component is significant in one or both models. The table also highlights key genre differences. For example, note that the Interactive/Conv. Style score is significant in the Inf. model but not in the Literary model. This reflects the fact that, while literary texts at all GLs tend to exhibit relatively high interactivity, similarly high interactivity among inf. texts tends to only be present at the lowest GLs. Thus, a high Interactivity is an indication of low complexity if the text in question is an informational text, but provides no statistically significant evidence about complexity if the text in question is a literary text.

Component	Informational	Literary
Academic Voc.	1.126*	.824*
Word Unfamiliarity	.802*	.793*
Word Concreteness	-.610*	-.483*
Syn. Complexity	.983*	1.404*
Lexical Cohesion	-.266*	-.440*
Interactive/Conv. Style	-.518*	*ns*
Degree of Narrativity	*ns*	-.361*
Argumentation	.431*	*ns*

Table 5. Regression coefficients estimated from training texts. **p* < .01, *ns* = not significant.

5.4 Validity Evidence

Two aspects of system validity are of interest: (a) whether genre bias is present, and (b) whether complexity scores correlate well with judgments provided by professional educators, i.e., the educators involved in selecting texts for use on high-stakes state reading assessments. The issue of genre bias is addressed in Figure 3. Each plot compares GL predictions generated via TextEvaluator to GL predictions provided by experienced educators. Note that no evidence of a systematic tendency to under-predict the complexity levels of literary texts is present. This suggests that our strategy of developing distinct prediction models for informational and literary texts has succeeded in overcoming the genre biases present among many key features.

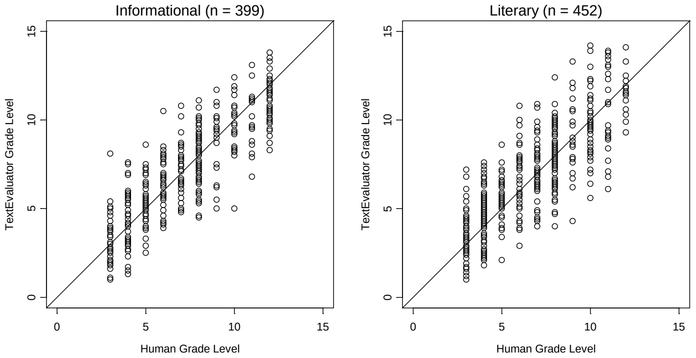

Figure 3. TextEvaluator GL predictions compared to human GL classifications for informational and literary texts.

TestEvaluator performance relative to the goal of predicting the human grade band classifications in the validation dataset was also examined. Results are summarized in Table 6 along with corresponding results for the Lexile Framework (Stenner, et al., 2006) and the REAP system (Heilman, et al., 2007). All results are reprinted, with permission, from Nelson, et al., (2012). In each case, performance is summarized in terms of the Spearman rank order correlation between the readability scores generated for each text, and corresponding human grade band classifications. 95% confidence limits estimated via the Fisher *r* to *z* transformation are also listed.

System	Lower 95% Bound	Correlation Coefficient	Upper 95% Bound
TextEvaluator	0.683	0.76	0.814
REAP	0.427	0.54	0.641
Lexile	0.380	0.50	0.607

Table 6. Correlation between readability scores and human grade band classifications for the 168 Common Core texts in the validation dataset.

The comparison suggests that, relative to the task of predicting the human grade band classifications assigned to the informational, literary and mixed texts in Appendix B of the new Common Core State Standards, TextEvaluator is significantly more effective than both the Lexile Framework and the REAP system.

6 Summary and Discussion

In many recent studies, proposed readability metrics have been trained and validated on text collections composed entirely of informational text, e.g., Wall Street Journal articles (Pitler and Nenkova, 2008), Encyclopedia Britannica articles (Schwarm and Ostendorf, 2005) and Weekly Reader articles (Vajjala and Meurers, 2012). This paper considers the more challenging task of predicting human-assigned GL classifications in a corpus of texts constructed to be representative of the broad range of reading materials considered by teachers and students in U.S. classrooms.

Two approaches for modeling the complexity characteristics of these passages were compared. In Approach #1, a single, non-genre specific prediction equation is estimated, and that equation is then applied to texts in all genres. Two measures developed via this approach were evaluated: the Lexile Framework and the REAP system.

Approach #2 differs from Approach #1 in that genre-specific prediction equations are used, thereby ensuring that important genre effects are accommodated. This approach is currently only available via the TextEvaluator system.

Measures developed via each approach were evaluated on a held-out sample. Results confirmed that complexity classifications obtained via TextEvaluator are significantly more highly correlated with the human grade band classifications in the held-out sample than are classifications obtained via the Lexile Framework or REAP system.

This study also demonstrated that, when genre effects are ignored, readability scores for informational texts tend to be overestimated, while those for literary texts tend to be underestimated. Note that this finding significantly complicates the process of using readability metrics to generate valid cross-genre comparisons. For example, Stajner, et al. (2012) conclude that SimpleWiki may not serve as a "gold standard" of high accessibility because comparisons based on readability metrics suggest that it is more complex than Fiction. We intend to further investigate this finding using TextEvaluator since conclusions that are not impacted by genre bias can then be reported. Additional planned work involves investigating additional measures of genre, and incorporating these into our genre classifier.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

American Institutes for Research (2008). *Reading framework for the 2009 National Assessment of Educational Progress.* Washington, DC: National Assessment Governing Board.

Biber, D. (1986). Spoken and written textual dimension in English: Resolving the contradictory findings. *Language, 62*: 394-414.

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., et al., (2004). *Representing language use in the university: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*. TOEFL Monograph Series, MS-25, January 2004. Princeton, NJ: Educational Testing Service.

Bormuth, J.R. (1964). Mean word depth as a predictor of comprehension difficulty. California *Journal of Educational Research, 15,* 226-231.

Cohen, S. A. & Steinberg, J. E. (1983). Effects of three types of vocabulary on readability of intermediate grade science textbooks: An application of Finn's transfer feature theory. *Reading Research Quarterly, 19*(1), 86-101.

Collins-Thompson, K. and Callan, J. (2004). A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL 2004*, Boston, USA.

Coltheart, M. (1981). The MRC psycholinguistic database, *Quarterly Journal of Experimental Psychology, 33A*, 497-505.

Common Core State Standards Initiative (2010). *Common core state standards for English language arts & literacy in history/social studies, science and technical subjects*. Washington, DC: CCSSO & National Governors Association.

Coxhead, A. (2000) A new academic word list. *TESOL Quarterly, 34(2)*, 213-238.

Gunning, R. (1952). *The technique of clear writing.* McGraw-Hill: New York.

Graesser, A.C., McNamara, D. S., Louwerse, M.W. and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments & Computers, 36*(2), 193-202.

Halliday, M. A.K. & Hasan, R. (1976) *Cohesion in English*. Longman, London.

Hiebert, E. H. & Mesmer, H. A. E. (2013). Upping the ante of text complexity in the Common Core State Standards: Examining its potential impact on young readers. *Educational Researcher, 42*(1), 44-51.

Heilman, M., Collins-Thompson, K., Callan, J. & Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'07*), 460-467.

Just, M. A. & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Boston: Allyn & Bacon.

Kincaid, J.P., Fishburne, R.P, Rogers, R.L. & Chissom, B.S. (1975). Derivation of new readability formulas (automated readability index, Fog count and Flesch reading ease formula) for navy enlisted personnel. Research Branch Report 8-75. Naval Air Station, Memphis, TN.

Kintsch, W. (1998). Comprehension: A paradigm for cognition. Cambridge, UK: Cambridge University Press.

Klein, D. & Manning, C. D. (2003). Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423-430.

Lee, D. Y. W. (2001) Defining core vocabulary and tracking its distribution across spoken and written genres. *Journal of English Linguistics. 29*, 250-278.

Nelson, J., Perfetti, C., Liben, D. and Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. Technical Report, The Council of Chief State School Officers.

Pitler, E. & Nenkova, A (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Association for Computational Linguistics, 186-195.

Schwarm, S. & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (ACL'05), 523-530.

Sheehan, K.M. (in press). Measuring cohesion: An approach that accounts for differences in the degree of integration challenge presented by different types of sentences. *Educational Measurement: Issues and Practice.*

Sheehan, K.M., Kostin, I & Futagi, Y. (2008). When do standard approaches for measuring vocabulary difficulty, syntactic complexity and referential cohesion yield biased estimates of text difficulty? In B.C. Love, K. McRae, & V.M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, Washington D.C.

Sheehan, K.M., Kostin, I., Futagi, Y. & Flor, M. (2010). *Generating automated text complexity classifications that are aligned with targeted text complexity standards*. (ETS RR-10-28). Princeton, NJ: ETS.

Si, L. & Callan, J. (2001). A statistical model for scientific readability. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM),* 574-576.

Štajner, S., Evans, R., Orasan, C., & Mitkov, R. (2012). What Can Readability Measures Really Tell Us About Text Complexity?. In *Natural Language Processing for Improving Textual Accessibility (NLP4ITA) Workshop Programme* (p. 14).

Stenner, A. J., Burdick, H., Sanford, E. & Burdick, D. (2006). How accurate are Lexile text measures? *Journal of Applied Measurement*, 7(3), 307-322.

Vajjala, S. & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, 163-173.

Yngve, V.H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society, 104,* 444-466.

Zeno, S. M., Ivens, S. H., Millard, R. T., Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.