

Online Readability and Text Complexity Analysis with *TextEvaluator*

Diane Napolitano, Kathleen M. Sheehan, and Robert Mundkowsky

Educational Testing Service

660 Rosedale Road, 12R

Princeton, NJ 08541, USA

{dnapolitano, ksheehan, rmundkowsky}@ets.org

Abstract

We have developed the *TextEvaluator* system for providing text complexity and Common Core-aligned readability information. Detailed text complexity information is provided by eight component scores, presented in such a way as to aid in the user's understanding of the overall readability metric, which is provided as a holistic score on a scale of 100 to 2000. The user may select a targeted US grade level and receive additional analysis relative to it. This and other capabilities are accessible via a feature-rich front-end, located at <http://texteval-pilot.ets.org/TextEvaluator/>.

1 Introduction

Written communication is only effective to the extent that it can be understood by its intended audience. A metric of readability, along with detailed information about aspects of the text which contribute its complexity, can be an indispensable aid to any content developer, teacher, or even reader. ETS's *TextEvaluator*¹ stands apart from similar systems (e.g.: Coh-Metrix (Graesser et al., 2011), Reading Maturity (Landauer, 2011), ATOS (Milone, 2008), Lexile (Stenner et al., 2006), and, perhaps most famously, Flesch-Kincaid (Kincaid et al., 1975)) in that it not only provides a single, holistic score of overall complexity, but also additional complexity information in the form of eight contributing components. The other systems known to us only provide one of these types of analysis. In addition to this,

¹*TextEvaluator* was previously known as *SourceRater*.

TextEvaluator will also provide the user with information on how its overall score and each of its component scores correspond to ideal values relative to a user-specified targeted grade level. All of this information is aligned with the current set of US grade school (K–12) text complexity standards outlined by the Common Core (CCSSI, 2010).

TextEvaluator's overall complexity scores are highly correlated with human grade level classifications, as shown in (Nelson et al., 2012). This study compared six systems as part of the Gates Foundation's Race to the Top initiative. Of these systems, the overall complexity scores computed by *TextEvaluator* were shown to have the highest Spearman rank correlation (0.756) between human grade level classifications on a set of 168 Common Core exemplar texts. *TextEvaluator* differs from these systems in that the computation of its overall complexity score relies on its set of eight component scores, each of which is a linear combination of four to ten fine-grained features. Most of these features are derived from information provided by part-of-speech tags and syntactic parses, unlike many competing systems which tend to only incorporate two or three basic features, such as average sentence length and average word frequency. Also unlike other systems, *TextEvaluator* differentiates between the two primary genres proposed by the Common Core: Informational texts, and their more challenging counter-parts, Literary texts. Internally, *TextEvaluator* makes use of either a model of Informational or Literary text complexity, in order to produce its final, overall score of complexity.

In this paper, we provide an overview

of how one can obtain and interpret *TextEvaluator* analyses received via the web. We provide a pilot version of our system at <http://texteval-pilot.ets.org/TextEvaluator/>.

Additional information on the overall complexity score and the component scores, as well as validity information, can be found in Sheehan et al. (2014) and Sheehan et al. (2013). Much of this information is also provided on the website's *About TextEvaluator* page.

1.1 Limitations

At this time, text submitted to *TextEvaluator* for analysis must be plain ASCII or UTF-8 text, free of images and tables. Many of *TextEvaluator*'s features make use of paragraphs, so it is recommended that at least one hard return is inserted between each paragraph. Manual word-wrapping will be corrected, and bulleted or numbered lists will be converted into one sentence per item.

TextEvaluator was designed for short reading passages, such as news articles or short stories, the sort of material one might expect to see during an exam or classroom assignment. We are currently investigating its use with longer (greater than 5-6 pages in length) texts. It is currently not suitable for poetry, plays, or texts that contain fewer than 2-3 sentences.

TextEvaluator was designed for use with materials that are publication-ready. Student assignments, such as essays, and transcripts of free-response monologues or dialogues, are not appropriate for *TextEvaluator*. *TextEvaluator* simply may not be able to analyze such transcripts or noisy text such as casual, online exchanges, due to its reliance on a syntactic parser (?). If the input contains at least one sentence that the parser cannot find a valid parse for, *TextEvaluator* cannot proceed with the analysis and will reject the text.

At this time, there is no programmatic API to *TextEvaluator* that is available to the public. However, batch-mode processing may be possible by contacting ETS via the information provided on the website.

2 Submitting a Text for Analysis

Upon visiting the *TextEvaluator* website, the user is asked to provide up to two pieces of information: a valid e-mail address and a client code. We first validate the provided e-mail address by sending an e-mail containing a link to the page described in Section 3.² Then, rather than have the user wait for their results to be computed, *TextEvaluator* will notify the user via e-mail when their results are ready.

An e-mail address is mandatory but a client code is not; specifying a valid client code gives the user access to some additional analyses. A client code can be obtained by purchasing a license for commercial use from ETS. However, this paper will focus primarily on the version of the website that is freely accessible for research and academic use.

Research and academic use is limited to texts that are 1600 words or less in length, and it is the responsibility of the user to truncate their texts. With a client code, the length of the text is not constrained. The user may either upload a plain text file or copy and paste such text into the larger input box. The user is then encouraged to provide a title for their submission, should they potentially have several *TextEvaluator* analyses on their screen at one time.

TextEvaluator will provide an additional set of analyses relative to a specified targeted grade level which ranges from US grades 2 to 12. At this time, the user is required to select a targeted grade. If a client code was entered, the user will be able to select additional targeted grades on the page containing their results.


3 The Results Page

The user will receive a link to their results via e-mail as soon as *TextEvaluator* has completed its analysis. Without the use of a client code, this web page will look similar to the one presented in Figure 1. Above the "Summary" tab, one can see the optional title they provided, or a title provided by *TextEvaluator*, along with two large boxes. The leftmost one will state whether or not the overall complexity of the submitted text is above, within, or below the expected range of complexity for your targeted grade

²This initial step is slated to be removed in a future version of the website.

Summary

Text Formatting Attributes	
Word Total:	1033
Sentence Total:	63
Average Words Per Sentence:	16
Paragraph Total:	25
Average Words Per Paragraph:	41
Quoted Words Total:	105
Flesch-Kincaid Grade Level Prediction:	5

 **Tip**
If these values are not consistent with your expectations, please review our [TextEvaluator formatting guidelines](#).

Level of Difficulty Relative to Grade		
Dimension of Variation/Component Score	Effect on Complexity	Value
Sentence Structure		
Syntactic Complexity (Increases Complexity)	↑	56
Vocabulary Difficulty		
Academic Vocabulary (Increases Complexity)	↑	22
Word Unfamiliarity (Increases Complexity)	↑	49
Concreteness (Decreases Complexity)	↓	61
Connections Across Ideas		
Lexical Cohesion (Decreases Complexity)	↓	39
Interactive/Conversational Style (Decreases Complexity)	↓	74
Level of Argumentation (Increases Complexity)	↑	65
Organization		
Degree of Narrativity (Decreases Complexity)	↓	57
Overall Text Complexity		
TextEvaluator Complexity Score		710
Classification Relative to Target Grade Level		Above

Figure 1: The results page one will see without the use of a client code. In this example, a targeted grade level of 4 was selected.

level. This information is also presented towards the bottom of the Summary tab, and will be explained later in this section. The rightmost box displays the text's overall complexity score on a scale of 100 (appropriate for first-grade students) to 2000 (appropriate for high-proficiency college graduates). As with the above/within/below analysis, this information is also presented towards the bottom of the Summary tab.

The box in the lefthand column of the Summary tab provides information regarding the contents of your text as discovered by *TextEvaluator*. If any of this information appears incorrect to the user, they are encouraged to reformat their text and submit it again. We also provide the Flesch-Kincaid grade level (Kincaid et al., 1975) of the text, should the user be interested in comparing their Common Core-aligned complexity score to one aligned to a previous standard.

TextEvaluator's analysis of the submitted text can be found in a table in the righthand column of the page. The scores of the eight components are presented, each on a scale of 0 to 100, with information regarding whether or not a higher value for that com-

ponent leads to a *more* complex or *less* complex text. This information is communicated via the arrows in the second column of this table. Each component score is the scaled result of a Principal Components Analysis which combines at least four but no more than ten distinct features per score. Provided is a brief description of each component; however, for more information, the reader is again referred to Sheehan et al. (2014), Sheehan et al. (2013), and the website's *About TextEvaluator* page.

3.1 Sentence Structure

Currently, the only component in this category, *Syntactic Complexity*, encapsulates all information regarding how complex the sentences are within the submitted text. It relies on information from syntactic parse trees³ (Manning et al, 2014) and part-of-speech tags (Toutanova et al., 2003), as well as basic measures such as the number of extremely long sentences and the size of the longest paragraph.⁴ As de-

³As provided by Stanford's shift-reduce parser, version 3.5.1: <http://nlp.stanford.edu/software/srparser.shtml>

⁴We make use of both a syntactic parser and a tagger in order to differentiate between possessives and the contractive form of "is". "s" forms tagged as POS by the tagger are re-attached to

scribed in section 1.1, should the parser fail to find a valid parse for any sentence in the text, *TextEvaluator* will be unable to calculate the features necessary to compute the text’s Syntactic Complexity score, and thus, unable to compute its overall complexity score.

3.2 Vocabulary Difficulty

This category’s components measure the amount of:

- *Academic Vocabulary*, words that are more characteristic of academic writing than that of fiction or conversation;
- Rare words, as determined by consulting two different word frequency indices and encapsulated in the *Word Unfamiliarity* component; and
- *Concreteness*, which describes the abstractness or difficulty one might have imagining the words within the text.

The two word frequency indices were created from one containing more than 17 million word tokens focused on primary school (K–12) reading materials, and one containing more than 400 million word tokens spanning primary and post-graduate school. Features in the Concreteness component are based on values of the perceived concreteness and imageability of each content word in the text as provided by the MRC Psycholinguistic Database (Coltheart, 1981).

3.3 Connections Across Ideas

The components within this category indicate the amount of difficulty the reader may have following the concepts presented throughout the text. *Lexical Cohesion* combines several features which are computed based on the number of overlapping lemmas between pairs of sentences in each paragraph. *Interactive/Conversational Style* is concerned with the use of verbs, contractions, and casual, spoken-style discourse, common to Literary texts. By comparison, the *Level of Argumentation* component provides a measurement of more formal, argumentative discourse, much more common in Informational texts. This component encapsulates words

the preceding noun and this modified tag structure is provided as input to the parser.

and phrases that are commonly found in argumentative discourse, such as subordinating concessive phrases (“although”, “however”, “on the contrary”), synthetic negations (“nor”, “neither”), “negative” adverbs (“seldom”, “hardly”, “barely”), and causal conjunctive phrases (“as a result”, “for this reason”, “under the circumstances”).

3.4 Organization

This category also only has one component, the *Degree of Narrativity*. This component differs from Interactive/Conversational Style in that it makes use of the number of words found within quotation marks, referential pronouns, and past-tense verbs, all of which are primary features of any written narrative.

3.5 The Overall Complexity Score

The determination of *TextEvaluator*’s overall complexity score is genre-dependent, relying on the idea that some features of text complexity will function differently for *Informational* and *Literary* texts. Thus, *TextEvaluator* will actually compute a different overall complexity score for each genre, each trained as a linear regression model of the component scores. Should the text actually be a combination of the two, a weighted average of the two scores is presented as the overall complexity score. The decision of which score to present to the user as the final, overall complexity score is determined by calculating the probability that the text is Informational. If that value is within a certain range, the text is said to be Informational, Literary, or Mixed. Regardless of the text’s genre, the complexity score’s relativity to the targeted grade level is determined the same way.

The notion of a text being above, within, or below the expected range of complexity relative to the targeted grade level is described further by the presentation of these ranges in Table 1. A text is “above” the allowable range of complexity if, for the chosen targeted grade, its complexity score is greater than the *Max* value, “below” if it is less than the *Min* value, or “within” if it is equal to, or between, the *Min* and *Max* values. The method used to establish this alignment between *TextEvaluator* complexity scores and the Common Core text complexity scale is described in (Sheehan, 2015). There, three evaluations of the proposed ranges are presented:

Target GL	Min	Max
2	100	525
3	310	590
4	405	655
5	480	720
6	550	790
7	615	860
8	685	940
9	750	1025
10	820	1125
11	890	1245
12	970	1360

Table 1: The *TextEvaluator*/Common Core alignment table, showing the expected ranges of complexity relative to a targeted grade level. Although complexity scores can be as high as 2000, only the ones presented in the ranges here have been externally validated by the Common Core (Sheehan, 2015).

one based on the 168 exemplar texts listed in Appendix B of (CCSSI, 2010); one based on a set of ten texts intended for readers who are career-ready; and one based on a set of 59 texts selected from textbooks assigned in typical college courses. In each case, results confirmed that *TextEvaluator*'s classifications of texts being above, within, or below each range of complexity are aligned with classifications provided by reading experts.

At this stage, users who provided a client code at the start of their analysis will be able to select and see analysis for a different targeted grade level. They will also receive an additional piece of information in the form of color-coding on each component score, relative to the selected targeted grade. Each score will be highlighted in red, yellow, or green, should the value for that component be either too high, a bit too high, or within or below the ideal range for a text in the same genre as the input text and at that targeted grade.

4 Conclusion

In this paper we have presented *TextEvaluator*, a tool capable of analyzing almost any written text, for which it provides in-depth information into the text's readability and complexity. This information is further summarized with a holistic score with both a high correlation to human judgement (Nelson et

al., 2012) and external validity. (Sheehan, 2015) It is these characteristics that lead us to believe that *TextEvaluator* is a useful tool for educators, content-developers, researchers, and readers alike.

References

- Coltheart, M. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4): 497 – 505.
- Common Core State Standards Initiative. (2010, June). *Common Core State Standards for English language arts and literacy in history/social studies, science and technical subjects*. Washington, DC: CCSSO and National Governors Association.
- Graesser, A.C., McNamara, D.S, and Kulikowich, J.M. 2011. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5): 223 – 234.
- Kincaid, J.P., Fishburne, R.P., Rogers, R.L., and 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 No. 8-75), NavalAir Station, Memphis, TN.
- Landauer, T. 2011. *Pearson's text complexity measure*. White Paper, Pearson.
- Manning, C.D., Surdeanu, M., Bauer, J, Finkel, J, Bethard, S.J., and McClosky, D. 2014. *The Stanford CoreNLP Natural Language Processing Toolkit*. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MD.
- Milone, M. 2008. *The development of ATOS: The Renaissance readability formula*. Wisconsin Rapids, WI: Renaissance Learning.
- 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, Washington, DC: Council of Chief State School Officers.
- Sheehan, K.M. 2015. *Aligning TextEvaluator scores with the accelerated text complexity guidelines specified in the Common Core State Standards*. ETS Research Report. Princeton, NJ: Educational Testing Service.
- Sheehan, K.M, Flor, M., and Napolitano, D. 2013. *A two-stage approach for generating unbiased estimates of text complexity*. In Proceedings of the 2nd Workshop on Natural Language Processing for Improving Textual Accessibility, Atlanta, GA.
- Sheehan, K.M, Kostin, I, Napolitano, D., and Flor, M. 2014. The TextEvaluator Tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal*, 115(2): 184–209.
- Stenner, A.J., Burdick, H., Sanford, E., and Burdick, D. 2006. How accurate are Lexile text measures? *Journal of Applied Measurement*, 7(3): 307 – 322.
- Toutanova, K, Klein, D., and Manning, C. 2003. *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network*. In Proceedings of the North American Association for Computational Linguistics, Edmonton, AB, Canada.