Readability Assessment Tool for English Texts

Joon Suh Choi and Scott Crossley

July 30, 2021

Readability Assessment Tool for English Texts

Joon Suh Choi[1], Scott A. Crossley[1]

[1]Applied Linguistics & ESL, Georgia State University

Author Note

The authors declare that there no conflicts of interest with respect to this preprint.

Correspondence should be addressed to Joon Suh Choi. Email: jchoi92@gsu.edu

Abstract

This paper introduces and tests the reliability of a new computer application that facilitates the application of eight different extant readability formulas (i.e., statistically derived readability assessment models that are based on linguistic features and different readability criteria). The reliability is tested by comparing the readability criteria from two separate corpora (the Bormuth corpus and the Newsela corpus) with the formula scores derived using the application. The results show that the formula scores exhibit significant and high correlation to the difficulty of the texts, and that the application produces reliable output suitable for use in research and education.

Keywords: readability assessment, readability formulas, text comprehension

Readability Assessment Tool for English Texts

Readability assessment (i.e., labeling and predicting the comprehensibility of text) can be applied to a wide range of tasks in the educational domain, including matching readers to appropriate reading materials and selecting materials for standardized tests. In recognition of such versatility, there has been continuous effort starting from the 1920s to derive accurate measures of text readability for texts written in English, leading to the development and distribution of various readability formulas.

Readability formulas are measures of text difficulty that are statistically derived from different readability criteria (i.e., representations of text difficulty through different measures such as fifth-word deletion cloze test results or intuitive, manual categorization of texts into discrete levels) and linguistic features. There are currently over hundreds of different readability formulas available for use, but there is no single formula that is considered to be the gold-standard without significant weaknesses. Older formulas are problematic because they prioritize ease of manual calculation and thus suffers from weak construct validity, while newer formulas are either impossible (i.e., the statistical models and/or the linguistic features that underlie the formulas are unavailable or not fully reported) or difficult (i.e., the linguistic features adopted in the formulas must be derived by using multiple different natural language processing tools) to replicate and implement. Such shortcomings of extant readability formulas warrant further research with regard to readability assessment.

ARTE (Automatic Readability Tool for English) is a free and easy-to-use tool that facilitates the use of eight different pre-existing formulas for research and practical purposes. ARTE offers batch processing of texts through readability formulas, which is an essential feature that many of the freely available online applications lack. ARTE also serves a purpose different from many other pre-existing natural language processing (NLP) tools that offer analyses pertaining to specific sets of linguistic features (and not holistic assessment of readability) such as CTAP (Chen & Meurers, 2016) or L2SCA (Xiaofei, 2010) in that the purpose of ARTE is to produce specific readability scores as outcome instead of particular individual linguistic features.

In this paper, we test whether ARTE produces reliable and expected outcome by using ARTE to produce formula scores from two separate corpora (i.e., we will process texts through different formulas to obtain readability scores) and comparing the results to the readability criteria of each respective corpus.

Method

Corpora

Bormuth Corpus

Bormuth corpus (Bormuth, 1971) consists of thirty-two texts (instructional materials) published between the years 1960 and 1966. The topics include biology, chemistry, civics, current affairs, economics, geography, history, literature, mathematics, and physics, and the difficulty of the texts range from K1 to college level. The texts comprise 8,482 words and 597 sentences in total.

Two different readability criteria are available for this corpus: one pertaining to L1 English readers and the other pertaining to L2 English readers. For the L1 readability criteria, fifth-word deletion cloze tests (i.e., every fifth word were replaced with underlined blanks) were developed based on the corpus and administered to 285 elementary and high school L1 English students by Bormuth (1971). For the L2 criteria, a similar fifth-word deletion cloze test was developed using thirty-one texts and administered to 200 Japanese students by Greenfield (1999). Each readability criteria represent the perceived difficulty of text with regard to L1 and L2 readers respectively.

Newsela Corpus

Newsela corpus (Xu et al., 2015) is a collection of 1,130 newspaper articles that have each been rewritten four times (each iteration designed to be easier to comprehend than the last) by professional editors at Newsela, an online platform that provides educational resources. In other words, the corpus comprises texts that are categorized into five discrete levels of difficulty (i.e., original, Simp-1, Simp-2, Simp-3, and Simp-4). The difficulty of the simplified texts was grounded using the Lexile Framework, a widely used commercial readability formula that predicts text difficulty based on sentence length and word frequency. For the present study, only one version out of the five different iterations of the same texts was selected to maintain independence between the texts.

Automatic Readability Tool for English

Automatic Readability Tool for English (ARTE) is a free application with user-friendly GUI that can be used to automatically calculate the readability scores of batches of texts using eight different readability formulas (i.e., four traditional formulas and four newer formulas). The four traditional formulas are Flesch Reading Ease formula (Flesch, 1948), Flesch-Kincaid Grade Level formula (Kincaid et al., 1975), the SMOG Readability formula (McLaughlin, 1969), and the Automated Readability Index (ARI; Kincaid et al., 1975). The four newer formulas are the New Dale-Chall Formula (Chall & Dale, 1995), Coh-Metrix L2 Readability Index (CML2RI; Crossley et al., 2008), Crowdsourced Algorithm of Reading Comprehension (CAREC; Crossley et al., 2019), and Crowdsourced Algorithm of Reading Speed (CARES; Crossley et al., 2019).

The seven formulas measuring text difficulty (i.e., all formulas excluding CARES) were selected for analyses in this study.

Table 1: Correlation Coefficients *r* Between Different Readability Formulas and L1 Cloze Test Results

	CTL1	FRE	FKGL	ARI	SMOG	NDC	CAREC	CML2RI
CTL1	1	0.91	-0.94	-0.93	-0.92	-0.82	-0.86	0.93
FRE		1	-0.98	-0.96	-0.98	-0.91	-0.87	0.91
FKGL			1	0.99	0.98	0.89	0.86	-0.93
ARI				1	0.97	0.85	0.84	-0.92
SMOG					1	0.91	0.87	-0.92
NDC						1	0.87	-0.82
CAREC							1	-0.76
CMl2RI								1

(CTL1: Cloze test results L1, FRE: Flesch Reading Ease, FKGL: Flesch-Kincaid Grade Level, ARI: Automated Readability

Statistical Analysis

To illustrate the usage of ARTE, we compare the formula scores of each text in the two different corpora (i.e., readability scores derived using seven different readability formulas) with the readability criteria of each text (i.e., L1/L2 cloze test results for the Bormuth corpus, and difficulty labels for the Newsela corpus). For the first analysis, we conduct correlation analyses between the formula scores and the cloze test results, and conduct Fisher *r*-to-*z* transformation to examine whether the correlations reported for the seven readability formulas significantly differ from one another. For our second analyses, we train ordinal logistic regression classifiers on a subset (60% of the texts) of the Newsela corpus and evaluate the accuracy of the classifiers using

Table 2: Correlation Coefficients *r* Between Different Readability Formulas and L2 Cloze Test Results

	CTL2	FRE	FKGL	ARI	SMOG	NDC	CAREC	CML2RI
CTL2	1	0.83	-0.85	-0.83	-0.83	-0.75	-0.67	0.90
FRE		1	-0.98	-0.96	-0.98	-0.90	-0.86	0.91
FKGL			1	0.99	0.98	0.88	0.85	-0.93
ARI				1	0.96	0.84	0.82	-0.92
SMOG					1	0.90	0.86	-0.92
NDC						1	0.86	-0.81
CAREC							1	-0.76
CML2RI								1

(CTL2: Cloze test results L2, FRE: Flesch Reading Ease, FKGL: Flesch-Kincaid Grade Level, ARI: Automated Readability

Table 3: ***p*-values of Correlation Between Different Readability Formulas and L1 Cloze Test Results**

	FRE	FKGL	ARI	SMOG	NDC	CAREC	CML2RI
FRE		0.424	0.617	0.818	0.159	0.373	0.617
FKGL			0.764	0.569	0.027	0.091	0.764
ARI				0.794	0.056	0.165	1
SMOG					0.099	0.259	0.795
NDC						0.603	0.056
CAREC							0.165
CML2RI							

(FRE: Flesch Reading Ease, FKGL: Flesch-Kincaid Grade Level, ARI: Automated Readability Index, NDC: New Dale-

a separate subset (40%). The precision, recall, accuracy, macro F1 score, and weighted kappa for each model are reported.

Results

Analysis 1

Results of the Pearson correlation analyses show that there are significant correlations between the formula scores (derived using the seven different readability formulas) and the L1/L2 cloze test results (see Table 1 and Table 2, all *p* values were < .001). Results of the Fisher r-to-z transformations show that the differences between the correlations are not statistically significant (see Table 3 and Table 4), meaning that the seven different readability formulas showed similar correlations to the readability criteria.

Table 4: ***p*-values of Correlation Between Different Readability Formulas and L2 Cloze Test Results**

	FRE	FKGL	ARI	SMOG	NDC	CAREC	CML2RI
FRE		0.802	1	1	0.418	0.159	0.289
FKGL			0.802	0.802	0.289	0.095	0.418
ARI				1	0.418	0.159	0.289
SMOG					0.418	0.159	0.289
NDC						0.542	0.615
CAREC							0.013
CML2RI							

(FRE: Flesch Reading Ease, FKGL: Flesch-Kincaid Grade Level, ARI: Automated Readability Index, NDC: New Dale-

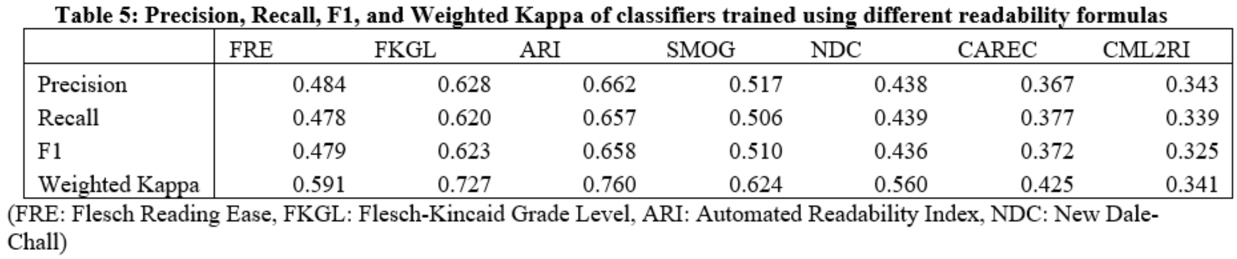

Table 5: Precision, Recall, F1, and Weighted Kappa of classifiers trained using different readability formulas

	FRE	FKGL	ARI	SMOG	NDC	CAREC	CML2RI
Precision	0.484	0.628	0.662	0.517	0.438	0.367	0.343
Recall	0.478	0.620	0.657	0.506	0.439	0.377	0.339
F1	0.479	0.623	0.658	0.510	0.436	0.372	0.325
Weighted Kappa	0.591	0.727	0.760	0.624	0.560	0.425	0.341

(FRE: Flesch Reading Ease, FKGL: Flesch-Kincaid Grade Level, ARI: Automated Readability Index, NDC: New Dale-Chall)

Analysis 2

All seven logistic regression models derived from seven separate formulas scores were statistically significant (*p* < 0.001). The precision, recall, F1 scores, and the weighted kappa for each model are reported in Table 5. The results demonstrate that ARI shows the best performance on all metrics, followed by Flesch-Kincaid Grade Level, SMOG, Flesch Reading Ease, New Dale-Chall, CAREC, and CML2RI. The results were as expected because the Newsela corpus is grounded on the Lexile framework which shares components similar to traditional readability formulas (i.e., they are based on average sentence length and word difficulty). The agreement between the true labels and predicted labels for ARI were excellent and beyond chance according to Fleiss (2013). All other formulas, with the exception of CML2RI, showed a fair-to-good agreement beyond chance between the true labels and predicted labels

Conclusion

This paper introduces ARTE and demonstrates that it produces expected and reliable output by comparing the formula scores and readability criteria derived from two separate corpora. The results showed that there was high correlation between the formula scores and the L1/L2 cloze test results of the Bormuth corpus, and a fair-to-good agreement between most of the formula scores and the discrete difficulty labels of the Newsela corpus. Such results suggest that ART is capable of producing reliable and expected measures of text difficulty suitable for use in both research and practical applications.

We are currently working on porting ARTE to a web environment and adding new readability formulas and features to make it more versatile and accessible. We hope that ARTE will be adopted by researchers and educators for a wide variety of research projects and practical tasks to help match readers with texts that are appropriate to their reading ability.

References

Bormuth, J.R. (1971). *Development of standards of readability: Toward a rational criterion of passage performance* (Report No. 9-0237).

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula.* Brookline Books.

Chen, X., & Meurers, D. (2016, December). CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)* (pp. 113-119).

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly, 42*(3), 475-493.

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: new methods and new models. *Journal of Research in Reading, 42*(3-4), 541-561.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions*. John Wiley & Sons.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology, 32*(3), 221.

Greenfield, G. R. (1999). *Classic readability formulas in an EFL context: Are they valid for Japanese speakers?* (Doctoral dissertation, Temple University).

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.* (Report No. RBR-8-75). Naval Technical Training Command, Millington, TN: Research Branch.

Lu, Xiaofei (2010). Automatic analysis of syntactic complexity in second language writing.

International Journal of Corpus Linguistics, 15(4):474-496.

McLaughlin, G. H. (1969). *Clearing the SMOG*. J Reading.

Xu, W., Callison-Burch, C., & Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics, 3*, 283-297.