EasyChair Preprint

№ 3851

Assessing Readability Formulas: A Comparison of Readability Formula Performance on the Classification of Simplified Texts

Joon Suh Choi and Scott A. Crossley

July 13, 2020

Assessing Readability Formulas: A Comparison of Readability Formula Performance on the Classification of Simplified Texts

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

[1]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 study compares the performance of five different traditional and new readability formulas in the task of classifying simple Wikipedia articles and authentic Wikipedia articles (N = 4,000). Results indicated that a new formula, the Crowdsourced Algorithm of Reading Comprehension (CAREC) performed the best. The traditional readability formula, Flesch-Kincaid Grade Level, also showed reliable performance. The results suggest the linguistic features used in newer readability formulas are capable of reliably representing the difficulty of texts.

Assessing Readability Formulas: A Comparison of Readability Formula Performance on the Classification of Simplified Texts

Learning to read is an important component of learning a language (National Research Council, 1998) and much of learning to read rests on matching texts to learners' reading ability (Crossley, Skalicky, & Dascalu, 2019). Texts are often simplified to better match texts to low level readers (Nation, 2008), and readability formulas are often employed to gauge the difficulty of the adjusted texts. However, the reliability and validity of traditional readability formulas have been challenged since their inception. The formulas have been criticized for their simplicity as a result of adopting only surface-level linguistic features and for their construct validity. Newer readability formulas are often either unavailable to the public or have not been extensively tested for validity and reliability.

The purpose of the current study is to test the reliability of both newer and traditional formulas by examining their predictive power in terms of classifying simplified texts and authentic texts. We predict that newer formulas utilizing more fine-grained linguistic features will perform better in the task of classifying simplified and authentic texts. While such expectation has been verified to an extent (Crossley et al., 2011), research on the validity of newer readability formulas requires replication. Thus, this study focuses on comparing the accuracy of a more diverse group of traditional readability formulas and newer readability formulas using a large corpus of online encyclopedia articles. The research question that guides this study is:

1. What is the classification potential of different readability formulas in differentiating between simplified texts and authentic texts?

Method

Corpus

A set of Wikipedia and Simplified Wikipedia articles (N =3,000) was used to train predictive models for both traditional and newer readability formulas in terms of text status (authentic or simplified). The models were then tested on another set of Simplified and authentic Wikipedia articles (N = 1,000) to measure the formulas' respective predictive capacity on a held-out dataset. Table 1 shows the descriptive statistics for the corpus used in this study.

Table 1. *Descriptive statistics for the Wikipedia and Simple Wikipedia articles*

	Mean number of words per text	SD	Mean number of paragraphs per text	SD
Simple	506.5	719.1	33.83	47.13
Authentic	4146.13	4299.2	188.17	205.88

Readability Formulas

Five readability formulas were selected for the analysis. Two of the formulas were classified as traditional readability formulas, and three were classified as new. To select the traditional readability formulas, a correlation analysis was conducted among six traditional readability formulas: Flesch Reading Ease formula (Flesch, 1948), Flesch-Kincaid Grade Level formula (Kincaid, Fishburne, Rogers, & Chissom, 1975), the SMOG Readability formula (McLaughlin, 1969), the Coleman-Liau formula (Coleman & Liau, 1975), the Gunning-FOG formula (Gunning, 1968), and the Automated Readability Index (Kincaid et al., 1975), using a larger collection of Simple Wikipedia and authentic Wikipedia articles (N = 49,398). Formulas for this study were selected that did not have a high correlation ($r \geq 0.700$) with other formulas (Mukaka, 2012) leaving two traditional readability formulas: the Flesch-Kincaid Grade Level formula, and the Coleman-Liau Index.

We selected three newer readability formulas as well. The criteria for the selection of newer formulas were: 1) the formula must include at least one linguistic feature that is not a

feature normally utilized by traditional readability formulas, 2) the formula must be available to the public, and replicable, and 3) the formulas must control for text length in the indices they contain to ensure that classification accuracy is not the result of length differences between simplified and authentic texts. Three formulas were found that met these criteria: the New Dale-Chall Formula (Chall & Dale, 1995), the Coh-Metrix L2 Reading Index (CMLRI; Crossley et al., 2007), and the Crowdsourced Algorithm of Reading Comprehension (CAREC; Crossley et al., 2019).

Statistical Analysis

A series of logistic regressions with each traditional readability formula as a single predictor variable were conducted to test how well the individual formulas could classify the texts as authentic or simplified encyclopedia articles. A logistic regression analysis was first conducted on a training set of 3,000 articles to produce a model for each readability formula, and the accuracy of the yielded models was then tested on a test set of 1,000 articles. The texts were assigned either a 1 or a 0 to represent whether they were categorized accurately for each respective formula. To test for the differences in classification accuracy, *t*-tests between the model outputs were conducted.

Results

All individual models that included the Flesch-Kincaid Grade Level formula, the Coleman-Liau formula, the New Dale-Chall formula, CMLRI and CAREC were statistically significantly over a baseline model with only the intercept ($p < .001$; see Table 2 for detailed statistics). Flesch-Kincaid Grade Level, Coleman-Liau, and New Dale-Chall performed better at recalling simple texts while showing better precision in classifying authentic texts. CAREC and

CMLRI performed better at recalling authentic texts while showing better precision in classifying simple texts. Table 3 shows the precision and recall of each formula.

Table 2. *Logistic Regression Predicting Decision from Readability Formulas*

Predictor	B	SE	Wald	P	Odds Ratio
Intercept	-9.3	0.33	-28.35	<0.001	<0.001
Flesch-Kincaid Grade	0.88	0.03	28.81	<0.001	2.4
Intercept	-8.53	0.33	-25.5	<0.001	<0.001
Coleman-Liau	0.75	0.03	25.79	<0.001	2.11
Intercept	-10.35	0.46	-22.53	<0.001	<0.001
New Dale Chall	1.01	0.04	22.69	<0.001	2.74
Intercept	-1.53	0.07	-21.96	<0.001	<0.001
Coh-Metrix L2 Reading Index	0.03	0.01	21.77	<0.001	1.03
Intercept	-7.26	0.27	-26.52	<0.001	<0.001
CAREC	3.07	0.12	25.27	<0.001	21.48

Table 3. *Precision and Recall Results*

Formula	Total set	Recall	Precision	F1
Flesch-Kincaid Grade Level	Simple	0.838	0.800	0.819
	Authentic	0.809	0.846	0.827
Coleman-Liau	Simple	0.746	0.743	0.744
	Authentic	0.738	0.750	0.744
New Dale-Chall	Simple	0.760	0.676	0.716
	Authentic	0.708	0.786	0.745
CAREC	Simple	0.835	0.898	0.865
	Authentic	0.890	0.822	0.855
Coh-Metrix L2 Reading Index	Simple	0.738	0.892	0.807
	Authentic	0.864	0.684	0.764

T-tests were conducted on all possible combinations of the results to test for statistical significance between the models (see Table 4 for descriptive statistics). The threshold for significance was set to $p = .01$ to prevent any Type I errors. Results indicate that CAREC performed significantly better than all other formulas except Flesch-Kincaid Grade Level, for which no significant difference at a reduced alpha value was shown $t(1998) = 2.267$, $p < 0.024$; $d = 0.1$. Flesch-Kincaid Grade Level showed significantly better performance when compared to

all formulas with the exception of CAREC and the Coh-Metrix L2 Reading Index, $t(1998) = 1.978$, $p < 0.048$; $d = 0.09$. The results are shown in Table 5.

Table 4. *Descriptive statistics for the accuracy of all formulas*

Readability formula	Mean	SD	N
Flesch-Kincaid Grade Level	0.823	0.381	1000
Coleman-Liau	0.744	0.438	1000
New Dale-Chall	0.731	0.444	1000
CAREC	0.860	0.347	1000
Coh-Metrix L2 Reading Index	0.786	0.409	1000

Table 5. *T-test results for all formulas*

	FKGL		CL		NDC		CAREC		CMLRI	
	t	p	t	p	t	p	t	p	t	p
FKGL			$d = 0.2$		$d = 0.22$		$d = 0.1$		$d = 0.09$	
CL	4.409	<.001			$d = 0.02$		$d = 0.3$		$d = 0.11$	
NDC	4.970	<.001	0.558	0.577			$d = 0.32$		$d = 0.13$	
CAREC	2.267	0.024	6.679	<.001	7.241	<.001			$d = 0.19$	
CMLRI	1.978	0.048	2.428	0.015	2.987	< .01	4.244	<.001		

(FKGL: Flesch-Kincaid Grade Level, CL: Coleman-Liau Index, NDC: New Dale-Chall Formula, CMLRI: Coh-Metrix L2 Reading Index)

Discussion

The results demonstrate that CAREC, a new readability formula, showed the strongest performance in the classification of simplified and authentic texts compared to other readability formulas. The results also demonstrate that Flesch-Kincaid Grade Level and Coh-Metrix L2 Reading Index performed well. Thus, this study provides reliability for at least one traditional readability formula and two newer readability formulas. We find promise in the strong performance of CAREC, a readability derived from a larger corpus adopting more deep-level linguistic features, as an indicator that a more theoretically valid and accurate measure of text difficulty can be derived using larger corpus and more fine-grained linguistic features.

References

National Research Council. (1998). *Preventing reading difficulties in young children*. National Academies Press.

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

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology, 60*(2), 283.

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a foreign language*, *23*(1), 84-101.

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.

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

Gunning, R. (1968). *The Technique of Clear Writing*. Mcgraw-Hill.

Kincaid, J. P., Fishburne Jr, 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*.

McLaughlin, G. H. (1969). SMOG grading–A new readability formula in the journal of reading.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, *24*(3), 69-71.

Nation, I. S. (2008). *Teaching ESL/EFL reading and writing*. Routledge