The CommonLit Ease of Readability (CLEAR) Corpus

Scott Crossley
Georgia State University
Atlanta, GA 30303
scrossley@gsu.edu

Aron Heintz
CommonLit
Washington, DC 20003
aron.heintz@commonlit.org

Joon Choi
Georgia State University
Atlanta, GA 30303
jchoi92@gsu.edu

Jordan Bachelor
Georgia State University
Atlanta, GA 30303
jbatchelor@gsu.edu

Mehrnoush Karimi
Georgia State University
Atlanta, GA 30303
mkarimi3@student.gsu.edu

Agnes Malatinszky
CommonLit
Washington, DC 20003
agnes.malatinszky@commonlit.org

ABSTRACT

In this paper, we introduce the Commonlit Ease of Readability (CLEAR) corpus. The corpus provides researchers within the educational data mining community with a resource from which to develop and test readability metrics and to model text readability. The CLEAR corpus improves on previous readability corpora include size (N = ~5,000 reading excerpts), the breadth of the excerpts available, which cover over 250 years of writing in two different genres, and the readability criterion used (teachers' ratings of text difficulty for their students). This paper discusses the development of the corpus and presents reliability metrics as well as initial analyses of readability.

Keywords

Text readability, corpus linguistics, pairwise comparisons

1. INTRODUCTION

Reading is an essential skill for academic success. One important way to support and scaffold literacy challenges faced by students is to match text difficulty to their reading abilities. Providing students with texts that are accessible and well matched to their abilities helps to ensure that students better understand the text and, over time, can help readers improve their reading skills. Readability formulas, which provide an overview of text difficulty, have shown promise in more accurately benchmarking students with their text difficulty level, allowing students to read texts at target readability levels.

Most educational texts are matched to readers using traditional readability formulas like Flesch-Kincaid Grade Level (FKGL) [19] or commercially available formulas such as Lexile [30] or the Advantage-TASA Open Standard (ATOS) [29]. However, both types of readability formulas are problematic. Traditional readability formulas lack construct and theoretical validity because they are based on weak proxies of word decoding (i.e., characters or syllables per word) and syntactic complexity (i.e., number or words per sentence) and ignore many text features that are important components of reading models including text cohesion and semantics. Additionally, many traditional readability formulas were normed using readers from specific age groups on small corpora of texts taken from specific domains. Commercially available readability formulas are not publicly available, may not have rigorous reliability tests, and may be cost-prohibitive for many schools and districts let alone teachers.

In this paper, we introduce the open-source the CommonLit Ease of Readability (CLEAR) corpus. The corpus is a collaboration between CommonLit, a non-profit education technology organization focused on improving reading, writing, communication, and problem-solving skills, and Georgia State University (GSU) with the end goal of promoting the development of more advanced and open-source readability formulas that government, state, and local agencies can use in testing, materials selection, material creation, and other applications commonly reserved for readability formulas. The formulas that will be derived from the CLEAR corpus will be open-source and ostensibly based on more advanced natural language processing (NLP) features that better reflect the reading process. The accessibility of these formulas and their reliability should lead to immediate uptake by students, teachers, parents, researchers, and others, increasing opportunities for meaningful and deliberate reading experiences. We outline the importance of text readability along with concerns about previous readability formulas below. As well, we present the methods used to develop the CLEAR corpus. We then examine how well traditional and newer readability formulas correlate with the reading criteria reported in the CLEAR corpus and discuss next steps.

2. TEXT READABILITY

Text readability can be defined as the ease with which a text can be read (i.e., processed) and understood in terms of the linguistic features found in that text [9][27]. However, in practice, many readability formulas are more focused on measuring text understanding (e.g., [18]) than text processing.

Text comprehension is generally associated with word sophistication, syntactic complexity, and discourse structures [17][31], three features whose textual elements relate to text complexity. For example, many studies have revealed that word sophistication features such as sound and spelling relationships between words [16][25], word familiarity and frequency [15], and word imageability and concreteness [28] can result in faster word processing and more accurate word decoding. The meaning of words, or semanticity, also plays an important role in text readability, in that readers must be able to recognize words and know their meaning [26]. Therefore, word semanticity and larger text segments can facilitate the linking of common themes and easier processing based on background knowledge and text familiarity [1][23].

Effective readers should also be able to parse syntactic structures within a text to help organize main ideas and assign thematic roles where necessary [13][26]. Two features that allow for quicker syntactic parsing are words or morphemes per t-unit [8] and sentence length [21]. Parsing information in the text helps readers develop larger discourse structures that result in a discourse thread [14]. These structures, which relate to text cohesion, can be partially constructed using linguistic features that link words and concepts within and across syntactic structures [12]. Sensitivity to these cohesion structures allows readers to build relationships between words, sentences, and paragraphs, aiding in the construction of knowledge representations [4][20][23]. Moreover, such sensitivity can help readers understand larger discourse segments in texts [11][26].

Traditional readability formulas tend use only proxy estimates for measuring lexical and syntactic features. Moreover, they disregard the semantic features and discourse structures of texts. For instance, these formulas ignore text features including text cohesion [4][20][23][24] and style, vocabulary, and grammar, which play important roles in text readability [1]. Additionally, the reading criteria used to develop traditional formulas are often based on multiple-choice questions and cloze tests, two methods that may not measure text comprehension accurately [22]. Finally, traditional readability formulas are suspect because they have been normed using readers from specific age groups and using small corpora of texts from specific domains.

Newer formulas, both commercial and academic, generally outperform traditional readability formulas. These formulas rely on more advanced NLP features, although this may not be the case with commercial formulas for which text features within the formulas are proprietary and, thus, not publicly available. Newer formulas come with their own issues though. For instance, commercially available formulas, such as the Lexile framework [30] and the Advantage-TASA Open Standard for Readability (ATOS) formula [29], often lack suitable validation studies. In addition, accessing commercially available formulas may come at a financial cost that is unaffordable for some schools and education technology organizations. Academic formulas such as the Crowdsourced Algorithm of Reading Comprehension (CAREC) [7] have been validated through rigorous empirical studies, are transparent in their underlying features, and are free to the public. However, the datasets on which they have been developed, while much larger than traditional readability formulas, can still be considered as relatively small and specific. The populations the formulas are trained on (i.e., adults) may also not generalize well to other target populations like young students.

3. CURRENT STUDY

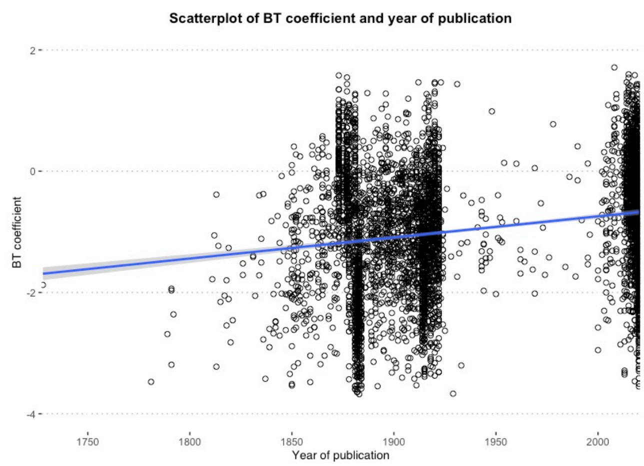
We hope to spur innovation to address many of the concerns noted above in reference to both traditional and newer readability formulas by publicly releasing the CommonLit Ease of Readability (CLEAR) corpus as well as hosting an open-source competition to develop readability formulas based on the CLEAR corpus. We hope that these formulas outperform existing readability formulas and can be used to better match 3rd-12th grade students to texts, thus improving learning outcomes in primary and secondary classrooms.

4. THE CLEAR CORPUS

4.1 Corpus Collection

We collected text excerpts from the CommonLit organization's database, Project Gutenberg, Wikipedia, and dozens of other open digital libraries. Excerpts were selected from the beginning, middle, and end of texts and only one sample was selected per text. Text excerpts were selected to be between 140-200 words, with all excerpts beginning and ending at an idea unit (i.e., we did not cut excerpts in the middle of sentences or ideas). The text excerpts were written between 1791 and 2020, with the majority of excerpts selected between 1875 and 1922 (when copyrights expired) and between 2000 and 2020 (when non-copyright texts were available on the internet). Visualizations of these trends are available in Figure 1.

Figure 1



Excerpts were selected from two genres: informational and literature texts. We started with an initial sample of ~7,600 texts. Each excerpt was read by at least two raters and judged on acceptability. The two major criteria for acceptability were the likelihood of being used in a 3rd-12th grade classroom and whether or not the topic was appropriate. We used Motion Picture Association of America (MPAA) ratings (e.g., G, PG, PG-13) to flag texts by appropriateness. Texts that were flagged as potentially inappropriate were then read by an expert rater and either included or excluded from the corpus. We also conducted automated searches for traumatic terms (e.g., terms related to racism, genocide, or sexual assault). Any excerpt flagged for traumatic terms was also reviewed by an expert rater. Lastly, we limited author representation such that each author had no more than 12 excerpts within the corpus. After removing excerpts based on these criteria, we were left with 4793 excerpts. These excerpts were copy-edited to ensure texts did not contain grammatical, syntactic, and spelling errors. Punctuation was also standardized in the texts, as were line-breaks. Lastly, selected archaic spellings (e.g., to-day, Servia) were replaced with modern spellings (e.g., today, Serbia) and identified British English spellings were converted to American spellings.

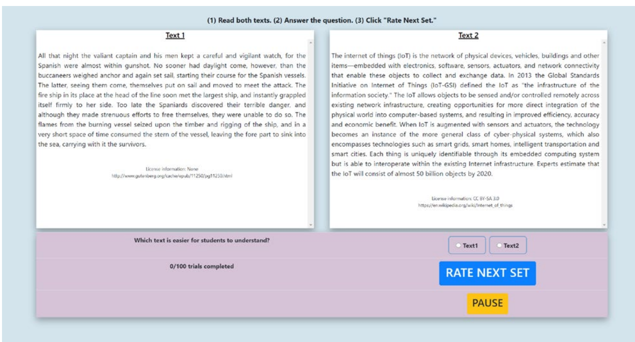
4.2 Human Ratings of Readability

We recruited ~1,800 teachers from the CommonLit teacher pool through an e-mail marketing campaign. Teachers were asked to participate in an online collection experiment. They were expected to read 100 pairs of excerpts and make a judgment for each pair as to which excerpt was easier to understand. Teachers were paid $50 in an Amazon gift card for their participation.

4.3 Data Collection Site

We developed an online data collection website. The basic format of the site was to show two excerpts side by side and ask participants to judge which of the two texts would be easier for a student to understand using a checkbox format. There were two additional buttons on the website. The first moved the participant to the next comparison and the second allowed participants to pause the experiment. The website also included a progress tally to show participants how many comparisons they had made (see Figure 2 for screenshot of pairwise comparison task).

Figure 2



The website first provided participants with informed consent and an overview of the expectations. The website then collected simple demographic information and survey information about reading/writing and television habits. Participants were then given a practice excerpt comparison to familiarize them with the design. After the practice comparison, participants moved forward with the data collection. Excerpts were paired randomly, and excerpts were shown on either the right or left-side panel randomly. The licensing information and the uniform resource locator (URL) for each text were displayed on the bottom side of each panel. Participants were redirected to a break screen after completing every 20 comparisons. The break screen showed how much time (in total and per comparison) the participant had spent on the task. A button allowing the participant to continue to the next comparison appeared after spending one minute on the break screen, meaning that the participants were required to take at least a one-minute break per 20 comparisons. After completing 100 comparisons, the participants were given a completion code that they could redeem for the gift card. The website was written in Python, JavaScript, CSS, and HTML. The website was housed on a cloud server.

4.4 Participant Reliability

Of the ~1,800 participants that initially logged into the experiment, 1,198 completed the entire experiment. However, not all participant data was kept. We removed participants who did not complete the entire experiment. We also removed participants to increase the reliability of the pairwise scores based on deviant patterns and time spent on judgments. In terms of deviant patterns, we removed all participants who selected excerpts in either the right or left panel more than 70% of the time. We also removed participants who had binary patterns of selecting left/right or right/left panels more than 20 times in a row. In terms of time spent on judgments, we removed participants who spent less than 10 seconds on average per comparison and/or spent a median time under 5 seconds. After removing participants based on patterns and time, we were left with data from 1,116 participants. Those participants made 111,347 overall comparison judgments (M = 99.773 judgments per participant). On average, each excerpt was read 46.47 times and participants spent an average of 101.36 seconds per judgment. However, we did not remove participants for taking too long on judgments, especially since pauses were allowed. Thus, our data for time was right skewed.

4.5 Pairwise Rankings for Readability

To calculate pairwise comparison scores for the human judgments of text ease, we used a Bradley-Terry model [3]. A Bradley-Terry model describes the probabilities of the possible outcomes when items are judged against one another in pairs (see Equation 1). The Bradley-Terry model ranks documents by difficulty based on each excerpt's probability to be easier than other excerpts. The model creates a maximum likelihood estimate which iteratively converges towards a unique maximum that defines the ranking of the excerpts (i.e., the easiest texts have the highest probability).

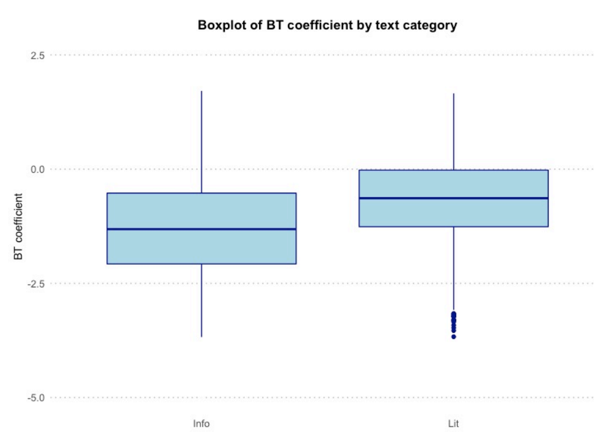
Equation 1: Bradley-Terry Model

$$P(\text{text}_i \text{ more difficult than } \text{text}_j) = \gamma_i / (\gamma_i + \gamma_j)$$

After computation, the Bradley-Terry model provides a coefficient for each text along with a standard error. We examined both coefficients and standard errors for outliers. We found 52 texts that had a coefficient with a standard deviation greater than 2.5 and additional 17 excerpts with a standard error greater than 0.65. These were removed from the final dataset leaving us with a sample size of 4,724. We conducted two additional analyses of the final data set in terms of differences in Bradley-Terry coefficients between informational and literature texts and trends in the coefficients as a function of time of publication for the texts.

As expected, we found significant differences between informational and literature texts such that informational texts were rated significantly more difficult ($t(4723) = -20.95$, $p < .001$), with a moderate effect size ($d = -0.61$). See Figure 2 for a box plot depicting this difference in text categories. In addition, we used a Pearson's correlation test to test whether Bradley-Terry coefficients were correlated with the texts' year of publication, finding a weak correlation, $r(4722) = .20$, $p < .001$). Thus, more recent passages were often rated as simpler than older passages (see Figure 1).

Figure 3



4.6 Pairwise Scoring Validation Checks

To examine convergent validity for the pairwise scores, we examined correlations between the scores and classic and newer readability formulas. The formulas we included were Flesch Reading Ease, Flesch Kincaid Grade Level, the New Dale-Chall, and the Crowdsourced Algorithm of Reading Comprehension [7]. All formulas were calculated using the Automatic Readability Tool for English (ARTE) [6]. ARTE provides free and easy access to a wide range of readability formulas and is available at linguisticanalysistools.org. ARTE automatically calculates different readability formulas for batches of texts (i.e., thousands of texts can be run at a time) and produces readability scores for individual texts in an accessible spreadsheet output. ARTE was developed to help educators and researchers easily process texts and derive different readability metrics allowing them to compare that output and choose formulas that best fit their purpose. The tool is written in Python and is packaged in a user-friendly GUI that is available for use in Windows and Mac operating systems. Correlations for this analysis are reported in Table 1.

Table 1: Correlations between readability formulas and text ease

	FRE	FKGL	NDC	CAREC
Text ease	0.547	-0.517	-0.557	-0.582
FRE		-0.913	-0.829	-0.726
FKGL			0.676	0.579
NDC				0.739

*FRE = Flesch Reading Ease, FKGL = Flesch Kincaide Grade Level, NDC = New Dale Chale, CAREC = Crowdsourced Algorithm of Reading Comprehension

The results indicate strong overlap between the four selected readability formulas and the text ease scores reported by the Bradley-Terry model. The strongest correlations were reported for CAREC while the weakest correlations were reported for FKGL. While strong, the correlations indicate that the readability formulas only predict around 27%-34% of the variance in the reading ease scores. Thus, there are opportunities for improvement in future readability formulas.

5. DISCUSSION

In this paper, we introduced the CommonLit Ease of Readability (CLEAR) corpus. The corpus provides researchers within the educational data mining community with a resource from which to develop and test readability metrics and to model text readability. The CLEAR corpus has a number of improvements over previous readability corpora, which are discussed below.

First, the CLEAR corpus is much larger than any available corpora that provide readability criterion based on human judgments. While there are large corpora that provide leveled texts (e.g., The Newsela corpus), these corpora only provide indications of reading ability based on levels of simplification (i.e., beginning texts as compared to intermediate texts). The corpora do not provide readability criterion for individual texts. Individual reading criteria, like that reported in the CLEAR corpus, allows for the development of linear models of text readability. While there are other corpora that have reading criteria for individual texts, the corpora are much smaller (N = ~20 - 600 texts), and they do not contain the breadth of texts found in the CLEAR corpus. The size of the CLEAR corpus ensures wide sampling and variance such that readability formulas derived from the corpus should be strongly generalizable to new excerpts.

The breadth of excerpts found in the CLEAR corpus is an additional strength. The corpus was curated from the excerpts available on the CommonLit website, all of which have been specially leveled for a particular grade level. The CommonLit texts were supplemented by hand selected excerpts taken from Project Gutenberg, Wikipedia, and dozens of other open digital libraries. The text excerpts were published over a wide range of years (1791-2020) and are representative of two genres commonly found in the K-12 classroom: informational and literary genres. The texts were read by experts to ensure they matched excerpts used in the K-12 classroom and checked for appropriateness using MPAA ratings. All texts were hand edited, so that grammatical, syntactic, and spelling errors were limited, while punctuation was minimally standardized to honor the authors' expression and style.

A final strength is the reading criteria developed for the CLEAR Corpus. Previous studies have developed reading criteria based on cloze tests or multiple-choice tests, both of which may not measure text comprehension accurately [22]. Additionally, while many readability formulas are marketed for K-12 students, their readability criteria are based on a different population of readers. The best example of this is Flesch-Kincaid Grade Level, which was developed using reading tests administered to adult sailors. We bypass these concerns, to a degree, by collecting judgments from schoolteachers about how difficult the excerpts would be for their students to read. This provides greater face validity for our readability criteria, which should translate into greater predictive power for readability formulas developed on the CLEAR corpus.

Lastly, while the purpose of the CLEAR corpus is for the development of readability formulas, the corpus includes meta-data that will allow for interesting and important sub-analyses. These analyses would include investigations into readability differences based on year of publication, genre, author, and standard errors, among many others. The sub-analyses afforded by the CLEAR corpus will allow greater understandings of how variables beyond just the language features in the excerpts influence text readability.

6. FUTURE DIRECTIONS

The next step for the CLEAR corpus is an online data science competition to promote the development of new open-science readability formulas. The competition will be hosted within an online community of data scientists and machine learning engineers who will enter a competition to develop readability formulas using only the reading excerpts and the reported standard errors to predict the Bradley-Terry ease of reading co-efficient scores. Prize money will be offered to increase the likelihood of participation. Once winners from the competition are announced, the winning readability formulas will be included in ARTE so that access to the formulas is readily available to teachers, students, administrators, and researchers. ARTE will also be expanded to include an online interface and a functional API. The online interface will allow end-users to easily upload texts to analyze for readability to better match texts to readers. The API will allow other educational technologies to include text readability formulas in their systems to help select texts for online students.

7. ACKNOWLEDGMENTS

We want to thank Kumar Garg and Schmidt Futures for their advice and support for making this work possible. We also thank other researchers who helped develop the CLEAR corpus and to

the nearly two thousand teacher participants who provided judgments of text readability.

8. REFERENCES

[1] Bailin, A. & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. Language & Communication, 21(3), 285-301. DOI = https://doi.org/10.1016/S0271-5309(01)00005-2.

[2] Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. Educational Psychology Review, 24(1), 63-88. DOI = https://doi.org/10.1007/s10648-011-9181-8.

[3] Bradley, R.A. & Terry, M.E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. Biometrika, 39, 324-345. DOI = https://doi.org/10.2307/2334029.

[4] Britton, B.K. & Gülgöz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. Journal of Educational Psychology, 83, 329-345.

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

[6] Choi, J.S., & Crossley, S.A. (2020) *Machine Readability Applications in Education*. Paper presented at Advances and Opportunities: Machine Learning for Education (NeurIPS 2020).

[7] 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. DOI = https://doi.org/10.1111/1467-9817.12283.

[8] Cunningham, J.W., Spadorcia, S.A., Erickson, K.A., Koppenhaver, D.A., Sturm, J.M., & Yoder, D.E. (2005). Investigating the instructional supportiveness of leveled texts. Reading Research Quarterly, 40(4), 410-427. DOI = https://doi.org/10.1598/RRQ.40.4.2.

[9] Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. Educational research bulletin, 37-54.

[10] Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221-233. DOI = https://doi.org/10.1037/h0057532.

[11] Gernsbacher, M.A. (1990). Language comprehension as structure building. Hillsdale, NJ: Erlbaum.

[12] Givón, T. (1995). Functionalism and grammar. Philadelphia: John Benjamins.

[13] Graesser, A.C., Swamer, S.S., Baggett, W.B. & Sell, M.A. (1996). New models of deep comprehension. In B.K. Britton & A.C. Graesser (Eds.), Models of understanding text, (pp. 1-32). Mahwah, NJ: Erlbaum.

[14] Grimes, J.E. (1975). The thread of discourse. The Hague, Netherlands: Mouton

[15] Howes, D.H. & Solomon, R.L. (1951). Visual duration thresholds as a function of word probability. Journal of Experimental Psychology, 41(6), 401-410. DOI = https://doi.org/10.1037/h0056020.

[16] Juel, C. & Solso, R.L. (1981). The role of orthographic redundancy, versatility and spelling-sound correspondences in word identification. (1981). In M.L. Kamil (Ed.), Directions in reading: Research and instruction, (pp. 74-82). Rochester, NY: National Reading Conference.

[17] Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. Psychological Review, 87, 329-354. DOI = https://doi.org/10.1037/0033-295x.87.4.329.

[18] Kate, R.J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R.J. et al. (2010, August). Learning to predict readability using diverse linguistic features. In Proceedings of the 23rd International Conference on Computational Linguistics, (pp. 546-554). USA: Association for Computational Linguistics

[19] Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Research Branch Report. Millington, TN: Naval Technical Training Command, 1975: 8-75. DOI = https://doi.org/10.21236/ada006655.

[20] Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. Psychological Review, 95, 163-182. DOI = https://doi.org/10.1037/0033-295X.95.2.163.

[21] Klare, G.R. (1984). Readability. In P.D. Pearson, R. Barr, M.L. Kamil, & P. Mosenthal (Eds.), Handbook of reading research (Vol. 1, pp. 681-744). New York: Longman.

[22] Magliano, J.P., Millis, K., Ozuru, Y. & McNamara, D.S. (2007). A multidimensional framework to evaluate reading assessment tools. In D.S. McNamara (Ed.), Reading comprehension strategies: Theories, interventions, and technologies, (pp. 107-136). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

[23] McNamara, D.S. & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. Discourse Processes, 22, 247-288. DOI = https://doi.org/10.1080/01638539609544975.

[24] McNamara, D. S., Kintsch, E., Butler-Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition and Instruction, 14, 1-43. DOI = https://doi.org/10.1207/s1532690xci1401_1.

[25] Mesmer, H.A. (2005). Decodable text and the first grade reader. Reading & Writing Quarterly, 21(1), 61-86. DOI= https://doi.org/10.1080/10573560590523667.

[26] Mesmer, H.A., Cunningham, J.W., & Hiebert, E.H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. Reading Research Quarterly, 47(3), 235-258. DOI = https://doi.org/10.1002/rrq.019

[27] Richards, J. C., Platt, J., & Platt, H. (1992). Longman Dictionary of Language Teaching and Applied Linguistics. London: Longman.

[28] Richardson, J.T.E. (1975). The effect of word imageability in acquired dyslexia. Neuropsychologia, 13(3), 281-288. DOI = https://doi.org/10.1016/0028-3932(75)90004-4.

[29] School Renaissance Inst., Inc. (2000). The ATOS[TM] readability formula for books and how it compares to other formulas. Madison, WI: School Renaissance Inst., Inc.

[30] Smith, D., Stenner, A.J., Horabin, I., Smith, M. (1989). The Lexile scale in theory and practice: Final report. Washington, DC: MetaMetrics.

[31] Snow, C. (Ed.) (2002). Reading for understanding: Toward an R & D program in reading comprehension. Santa Monica, CA: Rand.