Improving Reading Level Evaluations Using Sentence Structure and Word Frequency

Militsa Sotirova

TJHSST Computer Systems Research Lab 2018

Abstract

People have been using readability formulas, or reading level formulas, for a wide variety of uses for many years. These formulas usually depend on factors like the number of characters, words, sentences, syllables per word, etc. there are in a sample of text. I created a formula which relies on the number of clauses in a sentence, which indicates the sentence complexity, as well as the vocabulary difficulty, which is based on the frequency of the words in modern English. In addition, I created a scale to go along with my formula so a user can see where their text falls with respect to other, more well-known texts.

Background

People have been using readability formulas, or reading level formulas, for a wide variety of applications; for example, teachers and administrators use them to determine whether a certain text is too difficult for a group of students, and publishers of military handbooks and pamphlets use them to ensure that the content they are publishing is accessible to consumers (Readability Formulas). The first major use of readability formulas occured during World War II; “the war period made us realize more than ever the importance of reaching large audiences. More people had to fill out tax forms; more people had to be appealed to to buy war bonds; more people had to co-operate in numerous activities to help win the war” (Dale & Chall, 1948). Additionally, scientists can use reading level formulas to determine trends over time; researchers at Carnegie Mellon University's Language Technologies Institute did a study in which they found that most recent political candidates used language at a middle school level, while politicians from the past (e.g. Abraham Lincoln) used language at an 11th grade level (Spice, 2016). Furthermore, with the increased use of the Internet as a source for important medical information, patients need to have access to correct and comprehensible information for their conditions. “The current recommendation is that patient-related information should be written at the seventh-grade reading level at most” (Fozia & Anderson, 2017). This is necessary because this information must be easily understood by a large variety of people with varied educational backgrounds.

The apparent reading level of a body of text determines whether or not we want to read it; people may be discouraged or frustrated from reading a news article with abnormally difficult vocabulary, and it may get tedious to read something that is far too simple (Routman, 2003).

Thus, “it is no accident that many best-selling adult novels are written at the 8th grade level” (Fry, 2002). In addition, Edgar Dale and Jeanne Chall, creators of the Dale-Chall readability formula, said that for someone to “comprehend” a body of text, they need to be able to answer 50% to 75% of questions concerning the material (Dale & Chall, 1948). This fact is especially important because one thing that many elementary and middle school students across the country must complete is a reading comprehension test, which determines the approximate reading level of the student. However, these tests may give false results; students who answer questions correctly for a relatively simple passage receive abnormally high scores, perhaps high school level or even college level, which (in most cases, at least) is not accurate as they would have trouble reading books for that grade level. This is because the tests do not go beyond the actual grade level of the student, and administrators can change this by adopting a system to adapt the passages to get progressively harder to give a more accurate measurement. Furthermore, current formulas, such as the Automated Readability Index (ARI) or the Flesch-Kincaid Index, rely only on basic statistics of a passage, such as the number of characters, words, sentences, syllables, etc (Readability Formulas). Below are the ARI and the Flesch-Kincaid Index formulas in mathematical terms:

$$ARI = 4.71 * (\frac{characters}{words}) + 0.5 * (\frac{words}{sentences}) - 21.43$$

$$Flesch–Kincaid\ Reading\ Level = 0.39 * (\frac{words}{sentences}) + 11.8 * (\frac{syllables}{words}) - 15.59$$

Figure 1. ARI and the Flesch-Kincaid Index

Determining whether the reading level of a body of text is suitable for a specific audience is important, not only for the readers themselves, but also for publishers and writers who want to

appeal to a certain audience. My goal was to create an assessment of reading level beyond measuring simply the number of words and sentences present in the text, which are the metrics that many common formulas use. The ARI, for instance, outputs a number that corresponds to the grade level necessary to understand the text; for example, a sample from *The Outsiders* by S.E. Hinton outputs 7.3, which makes sense as this book is commonly read in middle school. However, the ARI fails when a relatively simple sentence happens to be long. Figure 2 below is a table which shows how the ARI fails when it encounters a sentence from *Winnie-the-Pooh*:

***Winnie-the-Pooh* (194 words, 1 sentence)**	**ARI**
In after-years he liked to think that he had been in Very Great Danger during the Terrible Flood, but the only danger he had really been in was the last half-hour of his imprisonment, when Owl, who had just flown up, sat on a branch of his tree to comfort him, and told him a very long story about an aunt who had once laid a seagull's egg by mistake, and the story went on and on, rather like this sentence, until Piglet who was listening out of his window without much hope, went to sleep quietly and naturally, slipping slowly out of the window towards the water until he was only hanging on by his toes, at which moment, luckily, a sudden loud squawk from Owl, which was really part of the story, being what his aunt said, woke the Piglet up and just gave him time to jerk himself back into safety and say, "How interesting, and did she?" when — well, you can imagine his joy when at last he saw the good ship, Brain of Pooh (Captain, C. Robin; 1st Mate, P. Bear) coming over the sea to rescue him.	45.5

Figure 2. Winnie-the-Pooh sample and ARI Score

"The developers of readability formulas have established certain basic factors of readability: vocabulary load, sentence structure or style, redundancy, and level of human interest in the written material" (Barker & Stokes, 1968). My assessment of reading level is based off of two of these factors: vocabulary load (or difficulty) and sentence structure (or complexity). The level of the vocabulary is a crucial component of a reading level formula because the words present in a body of text can be too difficult for someone to understand it properly. Different reading level formulas use different methods of measuring this, such as the number of characters per word or the number of syllables per word. Sentence structure is also a measure of reading level; I decided to focus on the number of clauses, or divisions in a sentence. For example, "a sentence consisting of both a main clause and a subordinate clause such as 'The woman saw a

man who ate a sandwich' is considered more complex than a coordinate structure as in 'The woman saw a man and ate a sandwich,' because the former comes later in acquisition than the latter" (Martohardjono et al., 2005). In the first sentence, there are two "thoughts," (the woman seeing a man and the man eating a sandwich) but there is only one in the second (the woman seeing the man while eating a sandwich). My calculation and analysis of reading level is made up of a combination of these factors because those were the two which I thought were the most relevant and measurable out of those listed above.

Method

Using Python and the CherryPy Webframe Network, I created a web program into which a user inputs a sample of text, and which outputs the reading level based on an algorithm I created. This assessment is based on the metrics of the number of clauses, or divisions, per sentence and the average word frequency, which correlates to vocabulary difficulty. To count clauses, I used the Web-based L2 Syntactic Complexity Analyzer created by Haiyang Ai. In order to access the frequencies of words in the input, I used the Oxford English Dictionary (OED) API, which uses the New Words Corpus (also known as the New Monitor Corpus). I chose this API because it draws from a corpus of around seven billion words, which come from "newspapers, scientific articles, and individual blogs" (The Oxford New Words Corpus). The pseudocode is shown below, followed by a more thorough explanation of the algorithm that I created:

```
remove punctuation & numbers
remove proper nouns
remove the most common 100 words in the English language #they can artificially lower the score
lemmatize each word & remove repeats #so that "phrase" and "phrases" are counted as the same word

for each word in remaining list: #count each word 4 times, so a hard word doesn't artificially increase the score
    request frequency
    add 1 / log(frequency) to running sum
use the sum to find average frequency → vocabulary difficulty

request number of clauses per sentence #the more clauses in a sentence, the more complicated it is

multiply the number of clauses per sentence by the vocabulary difficulty
```

I use the Natural Language Toolkit (NLTK) word-tokenizer method to split up the user's input into a list of words, or, more specifically, a list of entities, which include words, punctuation, numbers, etc. Then, I remove any words that contain punctuation, which is defined by the string.punctuation constant. I use NLTK's part of speech tagger to eliminate words which are proper nouns, as names like "Sarah" and the names of places will not affect the reading level. In addition, I remove the 100 most common words in the English language, according to *The Reading Teacher's Book Of Lists: Grades K-12*, because their overwhelming presence in a body of text can artificially lower the score; for example, think of how many times the words "the" and "is" appear in any text. Next, I lemmatize the remaining words; this reduces all different forms of a word into the same form. For instance, both "phrases" and "phrasing" become "phrase." I do this so that in the next step I can easily remove repeats of words, lowering the amount of words for which I need to request the frequency. Figure 3 is a picture of my code, which summarizes the process I created to prepare input text.

```
def prepare(self, textStr):
    words = word_tokenize(textStr)
    words = self.removePunct(words)
    words = self.removeProperNouns(words)
    words = self.removeCommonWords(words)
    words = self.lemma(words)
    words = self.removeRepeats(words)
    words = self.remove_nums_and_punct(words)
    return words
```

Figure 3. The method I wrote to prepare input text for further frequency analysis.

For each remaining word, I request its frequency from the OED API, and I take the log of this number so that the frequencies of all the words are more saturated. Then, I calculate the sum of the reciprocals of this value for each word, because the more frequent a word is, the "easier" it will be, as there is a greater chance that a person has encountered it in everyday life. Using this sum, I find the average frequency value across all the words in the text and multiply it by the number of clauses per sentence in the sample, as determined by the Web-based L2 Syntactic Complexity Analyzer. Figure 4 below is the final formula, after adjustments made when I looked at results from various sample inputs.

$$Reading\ Level = \frac{\Sigma \frac{1}{log(freq\ of\ word)}}{number\ of\ words} \times \frac{clauses}{sentences} \times 100$$

Figure 4. Final reading level formula

All of the above descriptions define the back-end of my project, which I wrote using Python. To connect the back-end to the front-end of my project, a user-friendly website, I used CherryPy. CherryPy is "a minimalist Python Web Framework," and I chose to use it over other

frameworks, such as Flask and Django, because it "allows developers to build web applications in much the same way they would build any other object-oriented Python program. This results in smaller source code developed in less time" ("CherryPy"). To accomplish the website portion of my project, I learned the basics of HTML, CSS, and JavaScript, as well as the basics of server connections and web applications.

Analysis

Using the formula in Figure 4, I found the reading levels of samples from a variety of different books. I did this so that I could create a scale, pictured in Figure 5. The purpose of the scale is for the user to be able to consider the reading level that my program outputs with respect to major works. For reference, Figure 5 below is a table of common books that students read in each grade level, as established by the Scholastic book company.

Reading Level (Grade)	**Common Books**
Pre K - K	*Clifford the Big Red Dog* by Norman Bridwell *Koko's Kitten* by Ronald H. Cohn & Dr. Francine Patterson
1 - 2	*Tooth Trouble* by Abby Klein & John McKinley *Nate the Great* by Marjorie Weinman Sharmat & Marc Simont
3 - 5	*A Wrinkle in Time* by Madeleine L'Engle *The Quest of the Cubs* by Kathryn Lasky
6 - 8	*The Bronze Key* by Cassandra Clare & Holly Black *The Sword of Summer* by Rick Riordan
9 - 12	*Rebecca* by Daphne du Maurier *The Hitchhiker's Guide to the Galaxy* by Douglas Adams

Figure 5. Standard reading levels for common books

For my scale, I created three "levels" for this: the first is meant to represent elementary school texts, the second is middle school texts, and the third is high school/college texts. After running samples from nine different books (three for each level) through my program, I saw that the first level has a score of about 10, the second has a score of about 15, and the third has a score of about 20. For the first level, I used samples from *Winnie-the-Pooh* by A.A. Milne, *Green Eggs and Ham* by Dr. Seuss, and *The Giving Tree* by Shel Silverstein; for the second level, I used samples from *The Giver* by Lois Lowry, *The Outsiders* by S.E. Hinton, and *The Hunger Games* by Suzanne Collins; and for the third level, I used samples from *The Age of Innocence* by Edith Wharton, *Moby Dick; Or, The Whale* by Herman Melville, and *Crime and Punishment* by Fyodor Dostoevsky. I chose these books based on benchmarks established by the Scholastic book company as well as my own past experiences in school. I used the scores for these samples as reference points, so the user has an idea of where the output for their input falls on the spectrum. Of course, this is not perfect, but the more text a sample has, or the more samples from a large text the user inputs, the better the output will represent the true reading level.

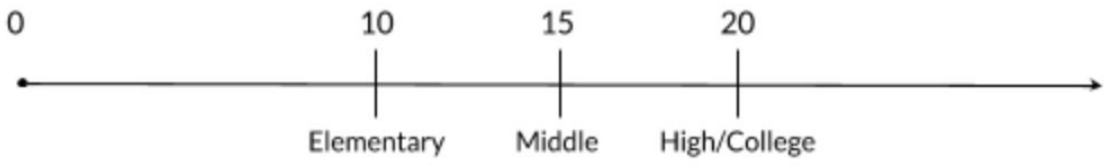

Figure 5. Scale

The *Winnie-the-Pooh* sample, which I used in the Background section of this paper, has a score of 15.4 by this formula, accurately representing the slight increase in complexity from the rest of the book, which has an average of around 13 without that specific sample.

Figures 6, 7, and 8 below are graphs which show, in more detail, the data I obtained when I created the scale.

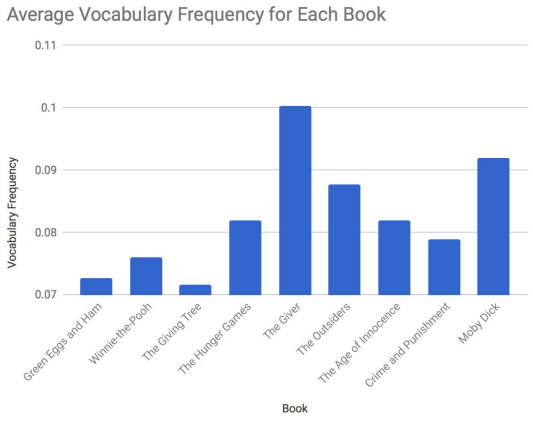

Figure 6. Average Vocabulary Frequency

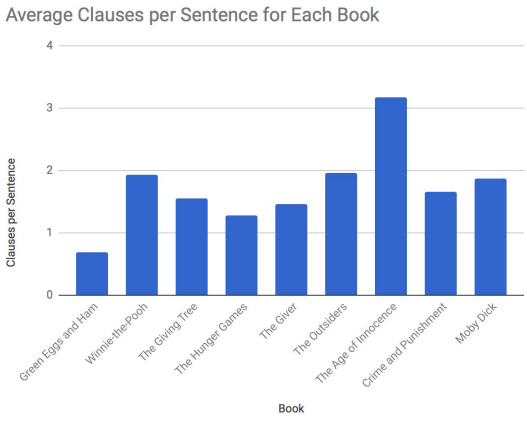

Figure 7. Average Clauses per Sentence

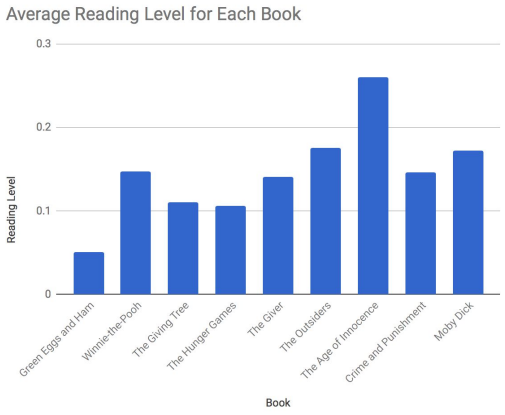

Figure 8. Average Reading Level for Each Book

It's interesting to note that individually, the two metrics of word frequency (Figure 6) and clauses per sentence (Figure 7) increase as the samples get "harder, but when multiplied together (Figure 8), they produce a stronger, expected upward trend among the three different levels.

As I was collecting data and creating the scale, it became clear that the more text a sample has, the more accurate the final reading level score will be. For example, the following sample from the *Age of Innocence* has a reading score of 50.7:

> Therefore, whenever anything happened that Mrs. Archer wanted to know about, she asked Mr. Jackson to dine; and as she honoured few people with her invitations, and as she and her daughter Janey were an excellent audience, Mr. Jackson usually came himself instead of sending his sister. If he could have dictated all the conditions, he would have chosen the evenings when Newland was out; not because the young man was uncongenial to him (the two got on capitally at their club) but because the old anecdotist sometimes felt, on Newland's part, a tendency to weigh his evidence that the ladies of the family never showed.

The incorrectly high score is the result of the presence of 13 clauses within 106 words, some of which I do not count because they are on the list of the most common 100 words or because they are pronouns; therefore, the clauses measure overwhelms the word difficulty measure. Furthermore, the average reading level for The Age of Innocence I obtained from the three samples I used is 25. This shows that, in general, more data is required for an accurate representation of the reading level of the input text.

Conclusion

I created a new formula to determine the reading level of a body of text based on the average frequency of the words present in the text in modern English as well as the number of clauses per sentence in the text. This is different from the ARI, or other previous reading level measurements and formulas, because it does not solely rely on simple characteristics of the text which do not necessarily indicate the reading level. Instead, it relies on properties which serve as key indices for complexity.

Further Research

For further research on this idea of calculating readability, I suggest looking into topics through topic modeling; a higher-level text might have ideas of politics, mechanics, or complicated philosophy while a lower-level text might have ideas of children, family, and morals. In addition, many texts include pictures; e.g. drawings or graphs. I suggest researching whether the presence of images indicates reading level; the presence of graphs perhaps represents a higher reading level than the presence of a colorful illustration of a dog.

Another aspect of the English language that could contribute to the difficulty of a text is the presence of idioms; a lower-level reader might not be familiar with certain well-known phrases that add meaning to the text. My formula deals only with the individual words themselves, so this would be an extension to what I have started, as the researcher could consider a list of English idioms and give the text a higher score for a greater number of them present in the text. My formula could also be extended to other languages in general; however, a different frequency source would be necessary, which may be difficult to find.

Acknowledgements

I would like to thank the Computer Systems Research Lab Director Dr. Peter Gabor for guiding and supporting me through the course of my project. I am also grateful to my peers for their encouragement and assistance for various parts of my research.

Bibliography

Ai, Haiyang and Lu, Xiaofei (2010). *A web-based system for automatic measurement of lexical complexity.*Paper presented at the 27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10). Amherst, MA. June 8-12.

Barker, D. G., and W. W. Stokes. "A Simplification of the Revised Lorge Readability Formula." *The Journal of Educational Research*, vol. 61, no. 9, 1968, pp. 398–400. *JSTOR*, JSTOR, www.jstor.org/stable/27532091.

"CherryPy." *CherryPy - A Minimalist Python Web Framework*, cherrypy.org/.

Collins, Suzanne. (2008). *The Hunger Games*. New York, New York: Scholastic Press.

Dale, Edgar, and Jeanne S. Chall. "A Formula for Predicting Readability." *Educational Research Bulletin*, vol. 27, no. 1, 1948, pp. 11–28. *JSTOR*, JSTOR, www.jstor.org/stable/1473169.

Dostoevsky, Fyodor. (1992). *Crime and Punishment* (R. Pevear & L. Volokhonsky, Trans.). New York, New York: Random House. (Original work published 1866)

Fozia Saeed, Ian Anderson, Evaluating the Quality and Readability of Internet Information on Meningiomas, World Neurosurgery, Volume 97, January 2017, Pages 312-316, ISSN 1878-8750, http://doi.org/10.1016/j.wneu.2016.10.001.

Fry, E. (2002). Readability versus leveling. *The Reading Teacher, 56*(3), 286-291.

Fry, E. B., & Dress, J. E. (2006). *The Reading Teacher's Book Of Lists: Grades K-12* (5th ed.). San Francisco, CA: Jossey-Bass.

Hinton, S. E. (2006) *The Outsiders*. New York, New York: Penguin House

Martohardjono, Gita. et al. (2005). T*he Role of Syntax in Reading Comprehension: A Study of Bilingual Readers.*

Melville, Herman. (2001) *Moby Dick; Or, The Whale*. Urbana, Illinois: Project Gutenberg. Retrieved April 18, 2018, from https://www.gutenberg.org/ebooks/2701.

Milne, A. A. (1926) *Winnie-the-Pooh*. New York, New York: Dutton Children's Books.

Lowry, Lois. (1993) *The Giver*. New York, New York: Houghton Mifflin Harcourt Publishing Company.

Lu, Xiaofei (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal, 96*(2), 190-208.

Readability Formulas. (n.d.). Retrieved from http://www.readabilityformulas.com/

Routman, R. (2003). *Reading essentials: The specifics you need to teach reading well*. Portsmouth, NH: Heinemann.

Scholastic. (n.d.). Retrieved from http://www.scholastic.com/home/

Seuss, . (1960). *Green Eggs and Ham*. New York, New York: Random House.

Silverstein, Shel. (1964). *The Giving Tree*. New York, New York: HarperCollins Publishers.

Spice, B. (2016, March 16). Most Presidential Candidates Speak at Grade 6-8 Level - News - Carnegie Mellon University. Retrieved from https://www.cmu.edu/news/stories/archives/2016/march/speechifying.html

The Oxford New Words Corpus (New Moni... | Oxford Dictionaries. (n.d.). Retrieved from https://en.oxforddictionaries.com/explore/oxford-new-words-corpus

Wharton, Edith. (1920). *The Age of Innocence*. https://itunes.apple.com/us/book/the-age-of-innocence/id498914762?mt=11