Improving Reading Level Evaluations Using Sentence Structure and Word Frequency

TJHSST Computer Systems Research Lab 2018
Militsa Sotirova

Background:

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). Therefore, determining whether the reading level of a body of text is suitable for a specific audience is important. 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 the Automated Readability Index (ARI) uses. The ARI outputs a number that corresponds to the grade level necessary to understand the text, and this becomes a problem when a simple sentence happens to be long. The table below shows how the ARI fails when it encounters a sentence from *Winnie-the-Pooh*:

From *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

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 (pseudocode shown below). 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 (or the New Monitor Corpus). I chose this API because it draws from a corpus of around seven billion words.

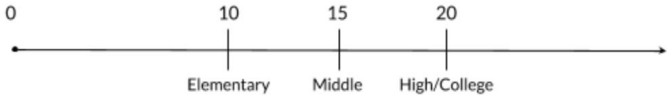
Pseudocode:

```
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
```

Scale:

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. 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. These scores are used as reference points, so the user has an idea of where the output for their input falls on the spectrum.

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



Conclusions:

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.