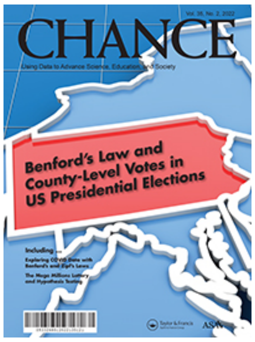


CHANCE

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucha20

The Past, Problems, and Potential of Readability Analysis

Nicholas A. Lines

To cite this article: Nicholas A. Lines (2022) The Past, Problems, and Potential of Readability Analysis, CHANCE, 35:2, 16-24, DOI: 10.1080/09332480.2022.2066411

To link to this article: https://doi.org/10.1080/09332480.2022.2066411

Published online: 04 May 2022.

The Past, Problems, and Potential of Readability Analysis

Nicholas A. Lines

Have you ever wondered how to make your technical writing easier to read? This seemingly simple question is remarkably relevant to many current events and both private and public sector decisions. Consider Washington DC's Minor Consent for Vaccinations Amendment Act of 2020 (D.C. Law 23-193) is the subject of two lawsuits filed in July 2021 and has been a topic of particular interest during the COVID-19 pandemic. The law in question allows minors 11 years or older to seek and receive vaccinations without parental consent, provided the youth can give informed consent by reading an age-appropriate vaccine information sheet. The law also stipulates the DC Department of Health is responsible for designing information sheets matching the needs of the minors in question. Much depends on whether these and other similar vaccine information guides can be objectively evaluated to determine their suitability for target audiences. The big question is how to do this.

All code to reproduce this analysis is available at *github.com/linesn/readability_analysis.*

Readability analysis can be used to approach this problem, just as it has been applied to other important issues such as the following:

- How can a blog or web-based news site increase readership?
- Are US soldiers fit for critical military positions requiring the ability to correctly analyze intelligence reports?
- Was a vital document translated well enough to preserve its technical meaning and level of accessibility?

While literacy rates have climbed to unprecedented heights worldwide during the past 50 years and average linguistic complexity continues to decline in published and informal media, individual reading and writing skills show high variance, even in developed countries with consistent school curriculum, as does the accessibility of critical written information. The need for tools objectively evaluating writing complexity has never been higher. Such tools are plentiful and fill an established role in professional editing. In fact, much of the text you encounter every day has already been analyzed using one or more statistical readability measures.

The central question asked by readability research is what makes text easier or harder to consume and whether these features can be objectively tracked to analyze reading difficulty with minimal human effort. Readability, as a problem in computational linguistics, is housed in the intersection of statistical modeling, theoretical linguistics, and psychological theory, next door to stylometry. Along with its role in making text accessible, it is inextricably linked to the pedagogical challenges of promoting child and adult literacy. In the US alone, readability (typically under the alias "plain writing") has been the subject of three executive orders, a 2010 act of congress, and countless studies funded by many hundreds of millions of dollars of government research grants. If readability is an unfamiliar subject to you, this may indicate these governmental efforts have successfully provided functional solutions relieving common readability issues.

However, while it may maintain a low profile, readability is not an outdated or solved problem, and its controversies lie close to the surface. Standard readability tools are often inconsistent, outdated, misused, or poorly matched to modern tasks. And as language evolves, so must our analysis.

The Past: The Enduring History of Readability Formulae

The story of the readability problem goes back as far as humans have critically thought about the efficacy of their language, but several good, abbreviated histories exist, usually picking up in the 1920s.

These histories show scoping the problem is far from trivial. Every writer, translator, and editor is faced with several rather independent challenges to making text intelligible. William S. Gray and Bernice Leary broke these into four camps in their foundational text

What Makes a Book Readable in 1935. These categories are **content** (including the ideas and organization), **style** (including diction and syntax), **structure** (the visual ordering and sectioning), and **design** (fonts, book cover, figures, etc.).

According to Gray and Leary, content is the most influential of the four, and superior content can cover a multitude of stylistic sins. We might say content is the destination and the other three elements form the pathway to it. A sufficiently interesting destination will still draw visitors despite a rugged path, but even the best-paved paths are ignored if the destination is dull.

Nevertheless, partly out of mathematical convenience, partly due to historical computational challenges, and partly to fill the objective analytic gap in typical writing advice, the focus of most readability research has been restricted to easy-to-count stylistic features. At first, this was done by hand, though modified typewriters were eventually devised to count words, sentences, and other variables, lending themselves to simple statistical analysis.

Between 1950 and 1980, most readability research followed the following simple pattern:

1. Select a predictive subset of stylistic features indicating text complexity (e.g., average word length and average number of words in a sentence).

2. Gather sample texts and present them in cloze tests (see sidebar) to various grade levels of readers to obtain a "ground truth" label for each text's grade level or reading difficulty, producing an outcome vector.

3. Compute the values of all features of interest for the sample texts, which yields predictor vectors.

4. Perform multiple linear regressions of the complexity on the predictors. This provides a formula mapping the document feature space to the real line, usually suggesting larger numbers represent more complex texts.

5. Split the range of formula outputs into bins corresponding to a grade level and compare the results with other readability formulae.

Many hundreds of formulae were proposed and analyzed in this fashion. A handful endured, mostly those depending solely on the average length of sentences, the average length of words, and the number of 'difficult' words used. One early example called Flesch Reading Ease is explored in Figure 1, which shows the decision boundaries for grade classification using this formula.

These regression-based formulae were challenged and debated from 1970 until recently, but rarely bested or improved. For example, Microsoft Word still measures readability in English with the Flesch-Kincaid formula published in 1975, the same year Microsoft was founded. Flesch-type scores are standard in other word processors, grammar tools, and publishing software. These and several other popular formulae in use today are summarized in Table 1.

The bulk of readability research in English was produced either to meet training and literacy demands for the US military and federal agencies or to guide

CLOZE TESTS

Cloze tests were introduced by Wilson Taylor in 1953. A cloze test presents sample readers with a text in which every n^{th} word is deleted (typically n=5) and replaced with a blank, which the subject fills in based on context. The process is similar to filling in *Mad Libs*. Ideally, all n possible tests are produced and presented to subsets of the respondents. The following example is part of a US Navy instruction manual used in a cloze test by J. Peter Kincaid and his coauthors when they created the Flesch-Kincaid formula. The underlined words are the deleted items readers had to fill in.

"TEMPORARY REPAIRS Temporary repairs are usually <u>made</u> by securing some type <u>of</u> patch over the damaged <u>section</u> of pipe. The material <u>used</u> for the patch depends <u>upon</u> the type of piping <u>that</u> is being repaired."

This form of analysis is preferred over multiple-choice exams based on the text content because cloze tests are simple to apply and avoid the subjectivity of exam design. Sometimes, these cloze tests are presented at the same time as a standard reading test (e.g., a Gates-McGinitie Reading Test) to identify the readers' "true" reading level, yielding real-valued labels like "Grade 5.9" rather than integers like "Grade 5," which is important for regression.

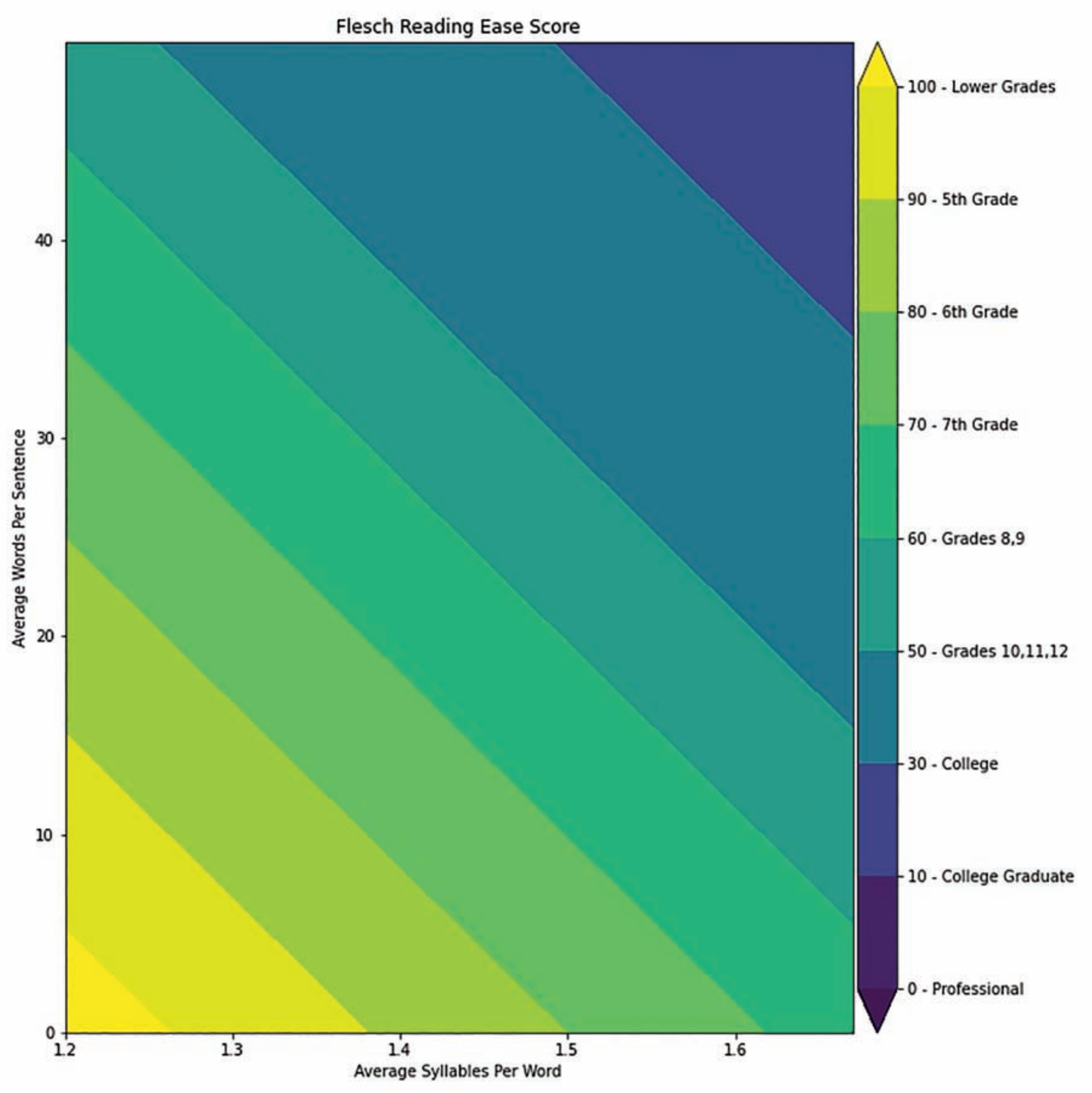
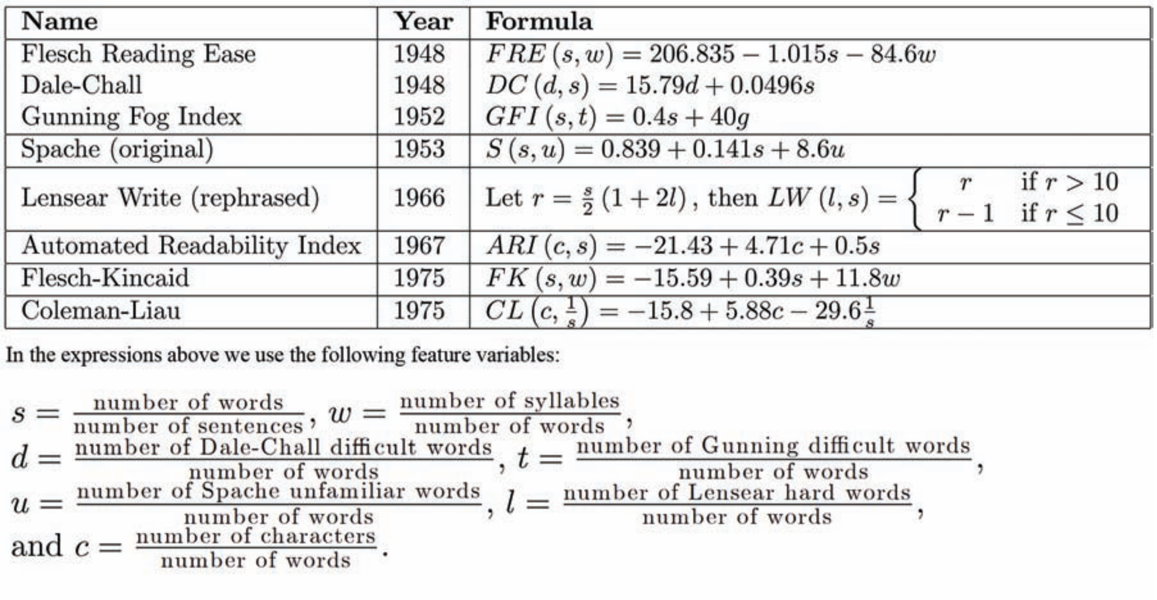


Figure 1. The decision boundaries for grading text using Flesch Reading Ease. Each shaded region corresponds to a single "grade" classification.

public education. Unfortunately, related efforts have often progressed without enough collaboration and interdisciplinary interaction. For example, the challenge of gauging (foreign) language proficiency in civil servants in the United States developed without significant coordination with ongoing research in English document complexity analysis. The federal government did not reach a unified approach to the former problem until the 1980s, when the Interagency Language Roundtable and standardized testing were introduced. The tightly interconnected problems of assessing the complexity of text and the ability of a particular person to consume text at a given complexity level are rarely treated in a holistic manner.

Current research in readability usually falls in one of three categories: application of standard readability formulae to literature in various domains; adaptation of (principally English) readability formulae to other languages; and the pursuit of better readability classifiers through machine learning. The latter group does

Table 1—Eight Common Readability Forumulae



Name	Year	Formula
Flesch Reading Ease	1948	$FRE(s,w) = 206.835 - 1.015s - 84.6w$
Dale-Chall	1948	$DC(d,s) = 15.79d + 0.0496s$
Gunning Fog Index	1952	$GFI(s,t) = 0.4s + 40g$
Spache (original)	1953	$S(s,u) = 0.839 + 0.141s + 8.6u$
Lensear Write (rephrased)	1966	Let $r = \frac{s}{2}(1+2l)$, then $LW(l,s) = \begin{cases} r & \text{if } r > 10 \\ r-1 & \text{if } r \leq 10 \end{cases}$
Automated Readability Index	1967	$ARI(c,s) = -21.43 + 4.71c + 0.5s$
Flesch-Kincaid	1975	$FK(s,w) = -15.59 + 0.39s + 11.8w$
Coleman-Liau	1975	$CL(c, \frac{1}{s}) = -15.8 + 5.88c - 29.6\frac{1}{s}$

In the expressions above we use the following feature variables:

$$s = \frac{\text{number of words}}{\text{number of sentences}},\ w = \frac{\text{number of syllables}}{\text{number of words}},$$
$$d = \frac{\text{number of Dale-Chall difficult words}}{\text{number of words}},\ t = \frac{\text{number of Gunning difficult words}}{\text{number of words}},$$
$$u = \frac{\text{number of Spache unfamiliar words}}{\text{number of words}},\ l = \frac{\text{number of Lensear hard words}}{\text{number of words}},$$
$$\text{and } c = \frac{\text{number of characters}}{\text{number of words}}.$$

show promise, providing steep improvements in classification tasks, but these new classifiers remain largely esoteric and have yet to be widely applied.

The Problems: Limits and Vicissitudes of Readability Testing

Despite their longevity, the traditional readability formulae come with the following troublesome baggage:

1. **Style is not enough.** The first and most prominent statements to this effect were given by the same researchers who created the style-based formulae, who rightly feared these would be applied blindly and content would be ignored. Rudolf Flesch, whose Reading Ease formula is probably the oldest readability tool still in use, originally published his formula alongside an 'interest' scoring function with the intention analysts would balance a text's complexity with the attraction of its content. It is not hard to generate garbage texts that are absolutely unreadable but still produce excellent formula scores. In fact, monosyllabic grunts would be optimal for most formulae. The tricky work of folding in analysis of coherence, cohesion, and other critical content-relevant mechanics has yet to be performed.

2. **Style-based readability is, at best, a vague concept and moving goalpost. At worst, it is established in a circular manner.** A readability formula range is mapped to grade levels and texts are evaluated and distributed to corresponding graded students in schools; when future students are surveyed to score new texts, they perpetuate the previous arbitrary definition and the cycle continues. While there are developmental phenomena possibly causing this process to converge to tools that are useful in grade school, the application to adult learners is not as clear. Furthermore, readability doesn't indicate how 'good' a text is. Ernest Hemingway may score more readable than F. Scott Fitzgerald, but that doesn't tell you much about the quality of their writing, only the complexity. Such a fact could be invoked to claim Fitzgerald was confusing or Hemingway was unsophisticated with equal merit.

3. **Readability formulae notoriously disagree.** It is not simply that they produce different grade-level predictions—this could be explained by the difference in their grade-labeling methodology, which is indeed inconsistent. The formulae may, in fact, be uncorrelated, as is the case in Table 2, in which we find the Coleman-Liau score does

Table 2—Correlation of Seven Popular Readability Formulae Across the Blog Authorship Corpus

	Dale-Chall	ARI	Coleman-Liau	Spache	Lensear Write	Flesch	Flesch-Kincaid
Dale-Chall	1.00	0.93	0.13	0.95	0.93	-0.94	0.93
ARI	0.93	1.00	0.11	1.00	1.00	-0.99	1.00
Coleman-Liau	0.13	0.11	1.00	0.09	0.09	-0.20	0.11
Spache	0.95	1.00	0.09	1.00	0.99	-0.99	1.00
Lensear Write	0.93	1.00	0.09	0.99	1.00	-0.99	1.00
Flesch	-0.94	-0.99	-0.20	-0.99	-0.99	1.00	-0.99
Flesch-Kincaid	0.93	1.00	0.11	1.00	1.00	-0.99	1.00

Table 3—Grading Inconsistencies Between Popular Readability Formulae

Text Sample	Dale-Chall	ARI	Coleman-Liau	Spache	Linsear Write	Flesch	Flesch-Kincaid
Shakespeare	12	14	10	8	17	9	12
Wiki Simple	13	11	11	7	11	12	9
Wiki Standard	14	14	13	8	17	13	13

Three sample texts are provided, beginning with the "All the World's a Stage" monologue uttered by Jacques in *As You Like It,* Act II, Scene VII. The second and third texts are the encyclopedia entries on William Shakespeare taken from the Simplified English and standard Wikipedia, respectively. The grades shown are the maximum grade predicted for each text by each formula. Grade 13 is considered college level, Grade 17 is considered graduate level, etc.

not strongly correlate with Flesch, Automated Readability Index, or other common formulae in the Blog Authorship Corpus (BAC). The formulae emphasize different features. This is not necessarily a bad thing, especially if they are being used in concert, but it can lead the unwary user into trouble. For example, it may seem probable Shakespeare's writing would be rated more complex (i.e., less readable) than, say, a Wikipedia article about Shakespeare. Yet as we see in Table 3, three typical readability tools grade one of Shakespeare's monologues at a lower, more readable grade than the simplified or standard Wikipedia articles, while four others disagree with this order.

4. **Word-difficulty is hard to assess and track.** The Dale-Chall formula and later corrections such as the Spache formula declared a list of words that were 'easy,' (e.g., through a survey of 4th graders). The problem with this approach is popular diction changes. Although Edgar Dale and Jeanne Chall updated their original 1953 list in 1995, the effect is still painfully obvious when perusing the corrected list. Speaking only of words beginning with "A," it comes as no surprise "automobile" and

Table 4—Correlation of Author Age and Sex (Female=0, Male=1) with Readability Formulae Values in the BAC

	Dale-Chall	ARI	Coleman-Liau	Spache	Lensear Write	Flesch	Flesch-Kincaid
Sex	0.03	0.01	0.21	0.01	0.01	-0.03	0.01
Age	-0.08	-0.04	0.26	-0.06	-0.05	0.02	-0.04

The sex row represents point-biserial correlation, and the age row represents Pearson correlation.

Table 5—Point-Biserial Correlation of Sex (Female=0, Male=1) to Blog Readability Scores for Three Self-Declared Vocation Groups in the BAC

	Dale-Chall	ARI	Coleman-Liau	Spache	Lensear Write	Flesch	Flesch-Kincaid
Accounting	0.03	-0.01	0.40	-0.01	-0.01	-0.04	-0.00
Investment Banking	0.17	-0.07	0.64	-0.05	-0.10	-0.08	-0.07
Maritime	0.33	0.31	0.52	0.30	0.25	-0.41	0.31

Rather than explaining away the relationship between sex and readability, grouping bloggers by vocation shows some professions exhibit more extreme sex bias in readability than others. Similar results are found for Pearson correlation of age to readability in these groups.

"airfield" have drastically declined in popularity. According to Google's n-grams data, which tracks term frequency in published English, 26 percent of the Spache 'easy' words and 31 percent of the Dale-Chall 'easy' words have declined in popularity since each list was constructed. Other formulae, such as the Flesch variants, express word complexity using syllable counts, which, aside from being language specific, are not consistent. Even simply counting word length in characters, like the Coleman-Liau index does, is highly dependent on tokenization strategy. The current generation of students has no difficulty skipping over hyperlinks in text, but these tend to overwhelm word-length measurements, meaning some texts kindergartners can parse are overscored by readability formulae. Inversely, some extremely small tokens such as emoji can represent complex concepts, including indicating sarcasm or mixed emotions.

5. **Some readability formulae are significantly correlated with demographics.** For example, readability measures focusing on word complexity tend to correlate with sex, age, and profession. In the BAC, for example, Coleman-Liau readability is weakly correlated with both sex and age, as shown in Table 4. Stylometry has previously shown men are slightly more likely to use hyperlinks and significantly more likely to use longer words, which is confirmed in the BAC posts. These features (and the readability measures dependent on them) are also strongly influenced by career field. Strangely, though, the age and sex biases for word complexity in the BAC are not explained by demographic preference for careers and, in fact, within the same career field, these biases can be much more pronounced. Table 5 shows more complex Coleman-Liau scores are more closely associated with male blog authors within the fields

LEARNING SIMPLE VS. STANDARD WIKIPEDIA CLASSIFICATIONS

Consider this example showing how quickly we can make readability tools more predictive than the classic formulae using only elementary machine learning practices. Wikipedia is an online encyclopedia that has been cloned in many languages—including 'simplified English.' While some loose standards are suggested, the meaning of "simplified" is largely decided by the Wikipedia contributors and editors, making it a fairly organic data source for simple language.

Suppose we wish to use the readability formulae to classify whether an article represents simple English or standard English according to these standards, using the Wikipedia data described earlier. Figure 2 shows the two populations have considerable overlap in their typical Dale-Chall readability scores, which is problematic if we wish to naïvely use these empirical distributions to assign the article to the most probable binary class. We confirm this intuition by applying 10-fold cross validation to a naïve Bayes classifier with the Dale-Chall score the only decision variable, resulting in an accuracy score of just 0.66.

Alternatively, we can use point-biserial correlation to identify text features most predictive of simple/standard labels, including typical sentence and word complexity features and less common features like number of commas per 100 words and proportion of personal pronouns per 100 words. Using these features in a naïve Bayes classifier, we achieve a 10-fold cross validation accuracy of 0.75. This is a modest but significant improvement, not only besting Dale-Chall's value as a discriminator but also every other readability formula mentioned in this article!

of accounting, investment banking, and maritime work than with female authors in the same fields. This is more of a feature than a bug, but it must be controlled for in practical situations. For example, two job candidates with similar educational and professional backgrounds and equivalent experience but different demographics may produce differently scoring text.

The Potential: Where Do We Go from Here?

Many of the concerns raised above naturally segue into the following rich possibilities in readability research and utilization:

1. **Readability and stylometry can benefit each other.** This is a rather unsurprising claim since readability analysis consists largely of combining individual stylometric features into a measure of text complexity, producing readability scores that are simply newly engineered stylometric features. At a high level, documents with disparate readability were probably composed by different authors. Furthermore, readability measures biased toward particular demographics can be used as a stylistic feature to identify authors or author groups. Thus, the interrelationship of readability and stylometry analysis indicates progress in either study may well advance both.

2. **Machine learning offers a chance to improve readability scores.** We have come a long way from counting typewriter keystrokes and all relevant computational barriers have been lifted. As hinted at before, some academic progress has been made toward more accurate grade-level classification as more sophisticated machinery than linear regression has been brought to bear. Hitherto, such work was neglected, perhaps out of a desire to retain the alluring transparency of the earlier readability formulae. However, we have already observed how dangerous it is for authors or editors to optimize text through the lens of any one formula. The Fry graph (1968) is an interesting step in the right direction: It directly describes readability grade levels as classification decision regions in the product space of word complexity and sentence complexity, eschewing a typical formulaic representation. An obvious extension could, and should, be made to analyze grade-labeled text data in an *n*-dimensional stylistic feature space and perform dimensionality reduction, cluster analysis, and related tasks that are already well-practiced within the machine learning community. This

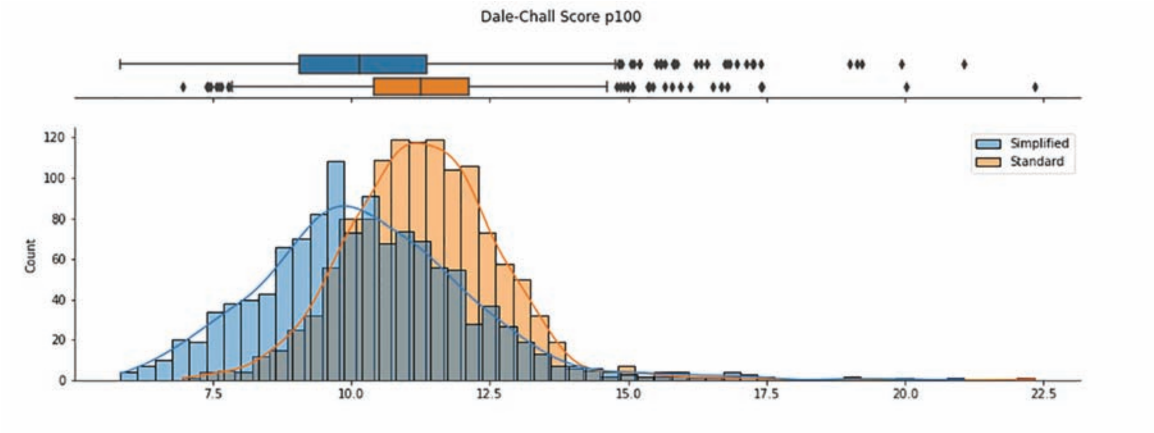


Figure 2. Histograms representing the Dale-Chall scores for the Wikipedia data set. For consistency, these are computed over only the first 100 words of each article. A boxplot is provided overhead, and a smoothed kernel estimate overlays each histogram.

genuine transparency could advance readability from its confines as a post-mortem analysis to a dynamic toolbox that can provide suggestions and corrections in a word processor environment to help authors write to their intended audience during first drafts.

3. **Readability can and should march beyond the bounds of style.** The rise of neural text generation software has led to valuable work in text coherence evaluation, which compliments stylistic readability analysis. There is also a good argument for including organizational evaluation, using change-point analysis to recognize the number of subject-changes in short texts or topic modeling to evaluate the efficiency of organization in longer texts. In some applications, it may also be convenient to tie in sentiment analysis to check the tone of a text, vocabulary entropy analysis to check sophistication of diction, or other content-oriented tests.

4. **Bigger and better data is available for the taking.** Modern data collection and experimental design can now be applied to readability studies. Automated cloze test production and evaluation using mechanical-Turk-style participants, students, or other groups is fairly inexpensive. Data fusion methods now exist to combine and contrast various historical and modern data sets. Vision-tracking tools also offer new ways to check text complexity at the word level.

5. Perhaps most importantly, **market forces and academic needs point in favor of readability advances.** Take, for example, the self-publishing industry spearheaded by Amazon's Kindle Direct Publishing that has unintentionally resulted in a strange economy in which ghostwriters assemble poor-quality texts (sometimes even blank books), which are then sold to 'authors' who self-publish through Amazon and often try to game the system to produce unmerited profits. Similarly, the past two decades have seen a bizarre plague of computer-generated nonsense articles infiltrating scientific publications and even computer-generated patent applications. More amusingly, the world seems to be filled with poorly translated instruction manuals; advertisements; government publications; and other high-readership, low-quality documents. Not only can all these challenges potentially be addressed via readability analysis, but there is also big money on the line motivating their resolution. Providing better discriminators may lead to more optimized fake content, but the asymptotic result of this 'readability arms race' would be, at worst, computer-generated material that is indistinguishable from human-generated text, which would be preferable over the empty or nonsensical fake content currently produced.

Conclusion

It is not hard to imagine a world in the near future where readability analysis is as universally applied as spell-checking and predictive text. On the other hand, the state-of-the-art in readability has changed very little over the past fifty years, despite its obvious faults. Much will depend on the various stakeholders who stand to benefit from this applied research in the fields of publication, academia and education, and government; will they be content to use increasingly fractured and unrelated solutions to their individual readability problems, or will they work together to produce a unified theory and practice? Although our world now revolves more around fused multimedia data than it does columns of newspaper print, most analysis of such multifaceted data relies on a solid foundation of elementary text and image processing tools, suggesting the readability problem will continue to haunt us no matter how many new channels of information are opened and layered. Meanwhile, conditions are ripe for significant progress to be made in many areas of readability research. The potential benefits and applications of these improvements are exciting to contemplate. 🅲

About the Author

Nicholas A. Lines is an applied research mathematician for the US Department of Defense. His research is focused on data science and text mining. He is pursuing a master's degree in applied and computational mathematics at Johns Hopkins University.

Further Reading

Schler, Jonathan; Koppel, Moshe; Argamon, Shlomo; and Pennebaker, James W. 2006. Effects of age and gender on blogging. In AAAI spring symposium: Computational approaches to analyzing weblogs, Vol. 6, pp. 199–205.

Gray, William Scott, and Leary, Bernice Elizabeth. 1935. *What makes a book readable*.

Herzog, M. An overview of the history of the ILR language proficiency skill level descriptions and scale. Interagency Language Roundtable. *www.govtilr.org/Skills/IRL%20Scale%20History.htm*.

Savoy, Jacques. 2020. *Machine learning methods for stylometry*. Springer International Publishing.