An Analytical Study of the Flesch-Kincaid Readability Formulae to Explain Their Robustness over Time

Yo Ehara
Tokyo Gakugei University
ehara@u-gakugei.ac.jp

Abstract

The Flesch–Kincaid formulae are classic and frequently utilized as English-specific readability metrics, even in recent assessments of large language models. These formulas combine the average sentence length in words with the average word length in syllables. Despite their simplicity, these formulas have been extensively used for decades, suggesting a cognitive rationale for their robustness. This study conducts a theoretical analysis of these formulae, examining the factors that contribute to their continued robustness over time. Notably, unlike previous research, we showed that these formulas can be interpreted as the average number of syllables per sentence. While the vocabulary inventory may expand as the grade level rises, the syllable inventory remains constant across different grades and ages. This stability is a key factor for their robustness over time. In our evaluation experiment, we confirm the validity of our theoretical framework using the British National Corpus (BNC).

1 Introduction

The Flesch–Kincaid formulas, specifically the Flesch–Kincaid Grade levels (FKGL) (Kincaid et al., 1980) and Flesch Reading Ease (FRE) (Flesch, 1948), are widely used to evaluate the readability of English texts, including those produced by large language models (Tanprasert and Kauchak, 2021; Imperial and Tayyar Madabushi, 2023; Kew et al., 2023). This popularity stems from the ease of interpreting the FKGL scores and the fact that neither method depends on word lists, which can be challenging to maintain.

One reason for the long-standing acceptance of the Flesch–Kincaid formulas (Kincaid et al., 1980) is their robustness. Unlike the formulas dependent on word lists, which are challenging to maintain and quickly outdated by new terms like "smartphones," these do not suffer from obsolescence. What makes these equations consistently reliable? We hypothesized that they must be based on the fundamental aspects of human cognition, an idea that drives our research. We demonstrate in later sections that these formulas are grounded in cognitive characteristics.

1.1 FKGL

Our emphasis is on FKGL, as the same rationale applies to FRE, and we aim to standardize the notation, where a higher value indicates greater difficulty.

$$\begin{aligned}\text{FKGL} &= 0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) \\ &+ 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59 \quad (1)\end{aligned}$$

The rationale for the FKGL is as follows. The number of words in a sentence, which corresponds to the first term, can act as an indicator of sentence complexity. Nonetheless, the average number of words alone does not fully capture the sentence difficulty. Even a brief sentence can be difficult for students if it includes words that are unfamiliar or sophisticated for their educational level. Consequently, the difficulty of vocabulary within a sentence must also be considered, giving rise to the second term, which calibrates the first term.

Nonetheless, this calibration appears to be overly heuristic and lacks theoretical assurance that the score can be excessively calibrated. This insight encourages the development of improved calibration techniques using larger annotated datasets and considering numerous linguistically rich attributes, thereby sacrificing the robustness of these formulas over time. A key recent automatic readability assessment study was Imperial (2021), and other studies were surveyed in Vajjala (2021).

This study addresses these research questions in relation to Equation 1. Each question is indicated by an *RQ*.

RQ1 *Why does a linear combination of the two measures, average number of words and average number of syllables in a sentence, work well?* This type of linear combination can be represented as the product of the average syllable count per sentence and M, which has a narrow range provided that the coefficients are appropriately chosen. We argue that FKGL utilizes the average syllable count per sentence to determine difficulty. We demonstrated that FKGL adopts this structure.

RQ2 *Is there any possibility of overcalibration? That is, is it possible that the average number of syllables in a word takes too large a value?* As indicated above, the maximum FKGL value can be obtained by determining the maximum value of M. This aided in establishing the upper bound.

RQ3 *What is the cognitive rationale behind FKGL?* As individuals grow, their vocabularies expand. Even sentences with only a few words, on average, may include challenging terms. Thus, the average word count per sentence does not necessarily reflect the text complexity and should be reconsidered. However, their phonetic repertoires did not increase with age. In other words, various recognizable syllables remained constant over time. Therefore, sentences with a higher average syllable count are undoubtedly more complex than those with fewer syllables. Indeed, because the Flesch–Kincaid Grade Level (FKGL) corresponds to a school year, we can derive the annual increase in the average number of syllables per sentence.

These new findings were not observed in previous FKGL studies and are an important contribution to this study.

2 Analyzing FKGL

We repeat Equation 1 as follows.

$$\begin{aligned}\text{FKGL} &= 0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) \\ &+ 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59 \quad (2)\end{aligned}$$

In this formula, the number of words per sentence and syllables per word appear. Currently, we focus on the number of words in each sentence. In computational linguistics, it is common to consider a sentence as a sequence of words and assume that there is always a word at the end of the sentence (EOS) that does not explicitly appear but is always present at the end of the sentence. Subsequently, the number of EOSs matches the total number of sentences; thus, the probability of EOS occurrence can be expressed by the following equation: For simplicity, we define this probability as p_{sw}, where s represents a sentence, and w represents a word. Thus, the number of words per sentence can be understood as the reciprocal of the probability of occurrence of a word that indicates the sentence boundary, as follows:

$$p_{\text{sw}} \equiv \frac{\text{total sentences}}{\text{total words}} \quad (3)$$

Similarly, the number of syllables per word can be considered a syllable sequence. To avoid confusion with the word 'sentence,' we will use the letter 'l' for syllables and express this probability as p_{wl}.

$$p_{\text{wl}} \equiv \frac{\text{total words}}{\text{total syllables}} \quad (4)$$

Furthermore, by setting the constants $a = 0.39$, $b = 11.8$, and $c = -15.59$, FKGL can be rewritten as follows:

$$\begin{aligned}\text{FKGL} &= \frac{a}{p_{\text{sw}}} + \frac{b}{p_{\text{wl}}} + c \\ &= \frac{1}{p_{\text{sw}} p_{\text{wl}}}\left(a p_{\text{wl}} + b p_{\text{sw}}\right) + c \quad (5)\end{aligned}$$

Here, we introduce the number of syllables per sentence p_{sl}.

$$\begin{aligned}p_{\text{sl}} &\equiv \frac{\text{total sentences}}{\text{total syllables}} \\ &= \frac{\text{total sentences}}{\text{total words}}\frac{\text{total words}}{\text{total syllables}} \\ &= p_{\text{sw}} p_{\text{wl}} \quad (6)\end{aligned}$$

Subsequently, eqrefeq: fkglabstis rewritten as follows:

$$\text{FKGL} - c = \frac{1}{p_{\text{sl}}}\left(a p_{\text{wl}} + b p_{\text{sw}}\right) \quad (7)$$

The right-hand side of Eq. eq:fkglabst can be decomposed into the first term $1/p_{\text{sl}}$ and second term $a p_{\text{wl}} + b p_{\text{sw}}$. Note that up to Equation 7, we only performed simple formula transformations, and no approximations were made. In the following subsection, we discuss the research questions predicted by Equation 7. We verify these research questions in the following sections.

2.1 Answer to research questions

The first research question is "*Why does a linear combination of the two measures, average number of words, and average number of syllables in a sentence work well?*". This can be partly explained by Equation 7. In Equation 7, FKGL is essentially the product of $\frac{1}{p_{\mathrm{sl}}}$, which is the average number of syllables in a sentence, and M, which is defined as In the experiments, we demonstrate that Equation 8 does not range significantly for FKGL using a general corpus.

$$M = (ap_{\mathrm{wl}} + bp_{\mathrm{sw}}) \tag{8}$$

The second research question was "*Is there any possibility of overcalibration? That is, is it possible that the average number of syllables in a word takes too large a value?*". Here, we can easily see that Equation 8 is bounded because p_{wl} and p_{sw} are probability values. Therefore, we can easily see that $0 \leq M \leq a + b$. Hence, combined with Equation 7, we can derive the following bound for Equation 1.

$$c \leq \mathrm{FKGL} \leq \frac{1}{p_{\mathrm{sl}}}(a + b) + c \tag{9}$$

In Equation 9, note that c is a negative value, namely $c = -15.59$, in the case of FKGL, whereas a and b are positive values. Hence, the FKGL is bounded by the number of syllables in a sentence. Hence, even if the average number of syllables in a word is excessively large, the FKGL is bound by the average number of syllables in a sentence. To the best of our knowledge, no previous study has addressed this theoretical bound. Hence, this is a novel result and is one of our contributions.

The third research question is "*What is the cognitive rationale behind FKGL?*". $\frac{1}{p_{\mathrm{sl}}}$ is the average number of syllables in the sentence. The average number of syllables in a sentence differs greatly from the average number of words. This is because the average number of acceptable words in a sentence changes according to the grade. Intuitively, we can see that acceptable vocabulary increases as the grade level increases. Teaching materials were created to increase vocabulary for each grade level. This indicates that the complexity of a text cannot be measured using the average number of words in a sentence alone. It is necessary to predict the acceptable vocabulary according to the learner's grade level and incorporate this into planning. It seems unlikely that the complex process of calculating text complexity involving both the average number of words in a sentence and changes in receptive vocabulary can be performed using a simple formula in the original equation Equation 1. This motivated the development of more advanced methods by considering the FKGL as a traditional heuristic. However, even advanced methods based on language models in recent years are models that view language as a sequence of words. For this reason, to estimate the complexity of a text for a particular grade, it is also necessary to estimate the vocabulary for that grade. Therefore, even if an advanced language model is used, it is still necessary to make predictions that consider both the average number of words in a sentence and the vocabulary used in that sentence.

However, the derived equation, Equation 7, provides a completely different perspective. This indicates that the FKGL can be considered as the average number of syllables in a sentence. It is assumed that the number of words in a learner's vocabulary, or vocabulary inventory, will increase as they progress through school. However, the number of syllables that can be recognized, or the phonetic inventory, will remain the same, even as they progress through school. The phonetic inventory is specific to a language, and once a person has acquired their native language, the number of phonemes in the phonetic inventory of native speakers of that language remains stable. In addition, owing to the arbitrariness of words, there is no need to use specific sounds to express specific difficulties. Because the size of the acceptable phonetic inventory is constant, an increase in the average number of syllables in a sentence certainly represents an increase in sentence complexity.

Furthermore, unlike vocabulary, the phonetic inventory is also very robust over time. While many words, like "smartphones," have become familiar in recent decades, virtually no languages have experienced a sudden increase or decrease in the number of phonemes over this period.

Unlike words, the average syllable count of a sentence does not model semantic complexity. Therefore, if a sentence is given with a low average syllable count but high semantic difficulty, this formula is likely to yield incorrect results. Hence, it is impossible to determine the difficulty level of a poem with a syllable-count limit such as a haiku. Intuitively, such studies are rare. Practically, practitioners and educators need to be careful when applying formulae to such limited types of text.

2.2 FKGL-derived increase in the number of syllables per year

If we use the equation in Equation 7, we can see that the increase in the number of syllables per year is also modeled from FKGL. In Equation Equation 7, we focus on the FKGL for a particular school year. Let us then consider the FKGL for the school year one year above FKGL+1. For FKGL+1, we assume that M remains constant.

For FKGL+1, let M remain the same and let p change from p_{sl} to p'_{sl}. Subsequently, the following equation holds:

$$\text{FKGL+1} - c = \frac{1}{p'_{\text{sl}}} M \tag{10}$$

$$\text{FKGL} - c = \frac{1}{p_{\text{sl}}} M \tag{11}$$

$$\tag{12}$$

By subtracting equation Equation 10 from equation Equation 11, we obtain

$$\left(\frac{1}{p'_{\text{sl}}} - \frac{1}{p_{\text{sl}}} \right) = \frac{1}{M} \tag{13}$$

The expression $frac1p_{\text{sl}}$ indicates the average number of syllables per sentence in a specified year, whereas $\frac{1}{p'_{\text{sl}}}$ signifies the average one year later. This highlights the increase in the average number of syllables per sentence over a year. In addition, it is evident that $\frac{1}{M}$ denotes an increase in the average number of syllables per sentence over a year.

3 Experiments

3.1 Setting

Based on this, we now describe our experiments. The British National Corpus (BNC) was used in this experiment. We used the Python readability library to determine the average number of words and syllables in the sentences.

3.2 Histogram of FKGL

First, we present a histogram of FKGL in the BNC. Figure 1 shows the histogram. The histogram exhibits a bell-shaped curve.

3.3 Histogram of the average number of syllables in a sentence

Next, we present our key findings: in Equation 7, we recognized that FKGL could be reformulated and that the primary complexity of the input text is represented by the average number of syllables per sentence, labeled as $\frac{1}{p_{\text{sl}}}$. We determined the average number of syllables in the texts from the BNC corpus using the readability library to compute this metric for each text, $\frac{1}{p_{\text{sl}}}$, and subsequently created a histogram of the results. The histogram in Figure 2 shows the average number of syllables per text on the horizontal axis and the percentage on the vertical axis. We can observe that Figure 2 also exhibits a bell-shaped distribution similar to Figure 1, indicating that the complexity of the text is captured by the average number of syllables per sentence, as anticipated.

3.4 Scatterplot of FKGL against the average number of syllables per sentence

Following this, Figure 3 displays a scatter plot depicting the relationship between FKGL and the average number of syllables per sentence. Figure 3 highlights a distinct correlation between FKGL and the average syllables per sentence, reinforcing the idea that the average syllables per sentence is crucial in FKGL for representing text complexity.

3.5 Checking that M does not change significantly

We postulate that M remains constant in Equation 8. To verify this result, we present a histogram of M in Figure 4. The horizontal axis represents the values of M and the vertical axis represents the percentage. The peak for M clusters is approximately 1. According to Equation 7, because M is the sole factor multiplied by the average syllable count per sentence, the average syllable count per sentence is almost directly utilized in the FKGL. Indeed, nearly 60% of M fall within the ranges of 0.7 and 1.0. In addition, we observed an extended tail, indicating that high M values were uncommon.

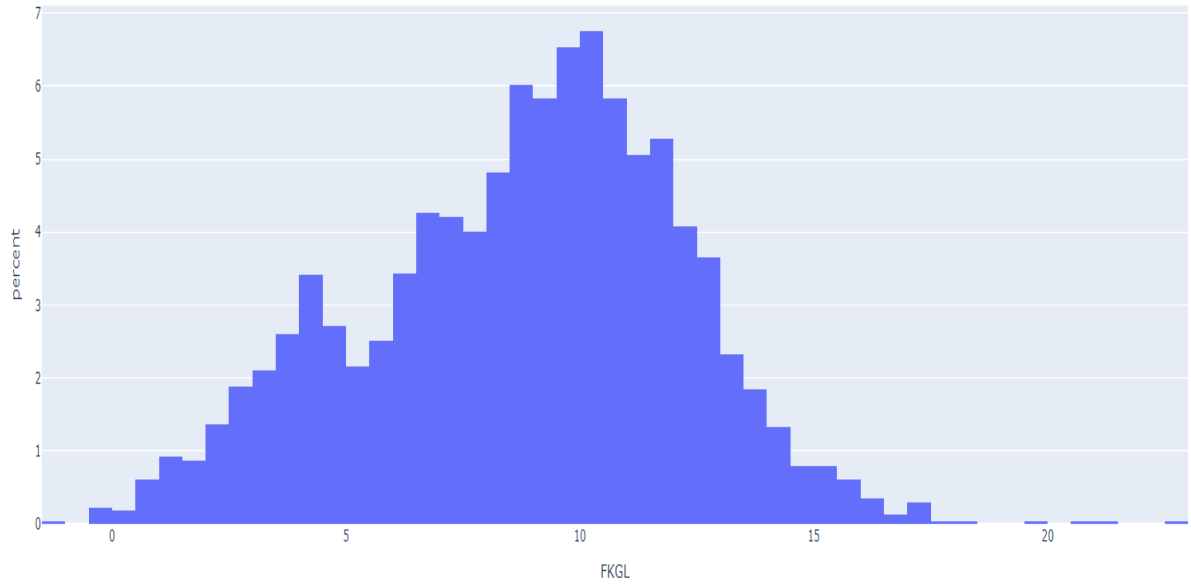

Figure 1: Histogram of FKGL in BNC.

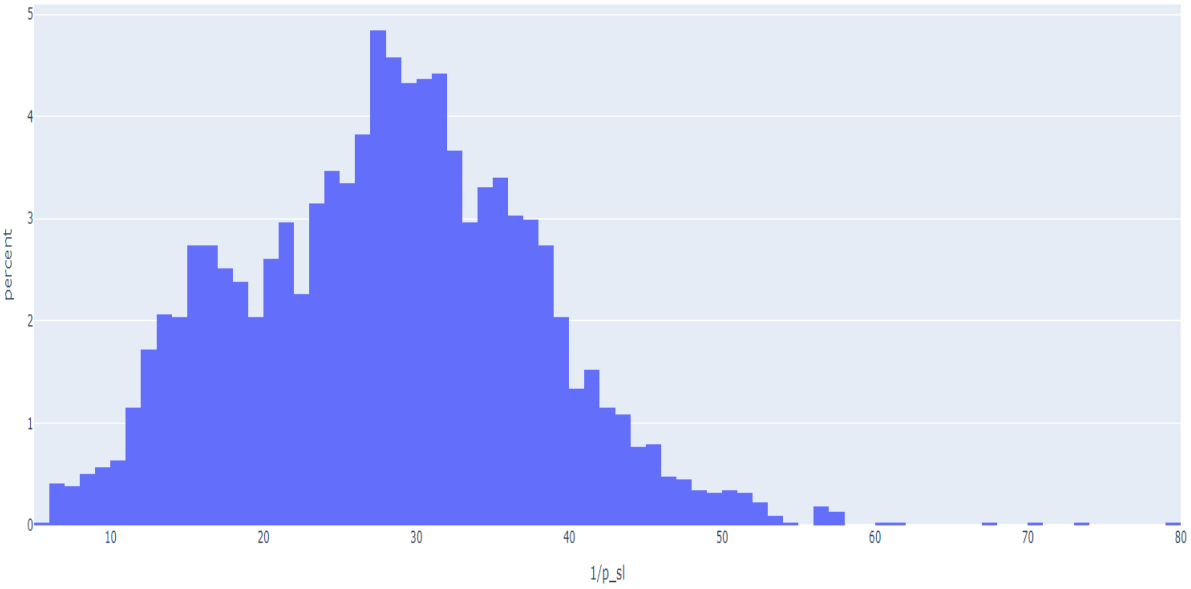

Figure 2: Histogram of $1/p_{\mathrm{sl}}$ in BNC.

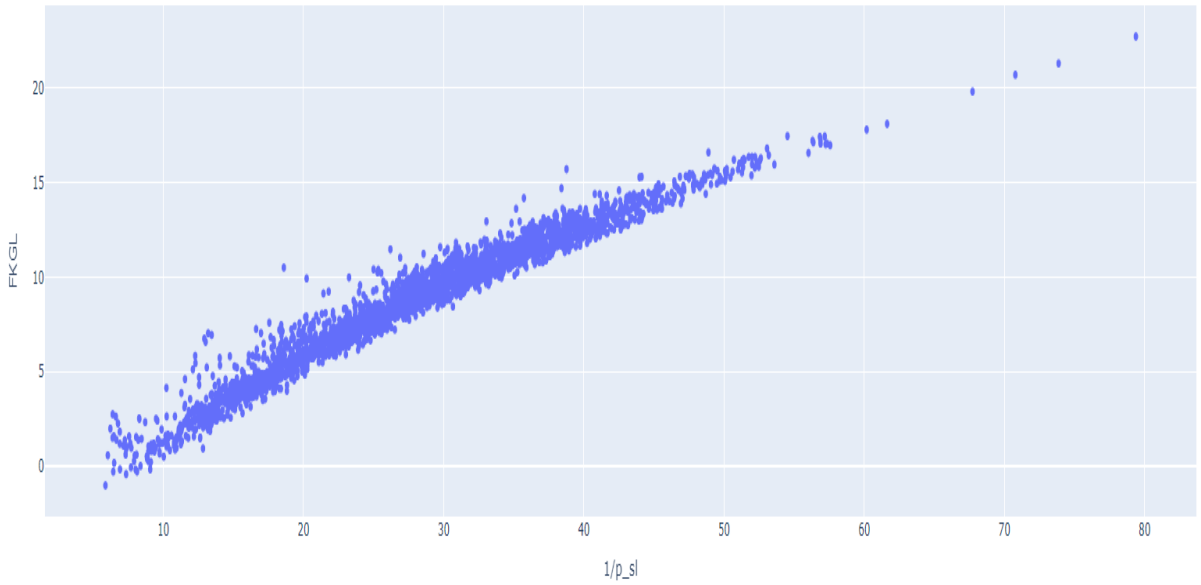

Figure 3: FKGL against $1/p_{\text{sl}}$ in BNC.

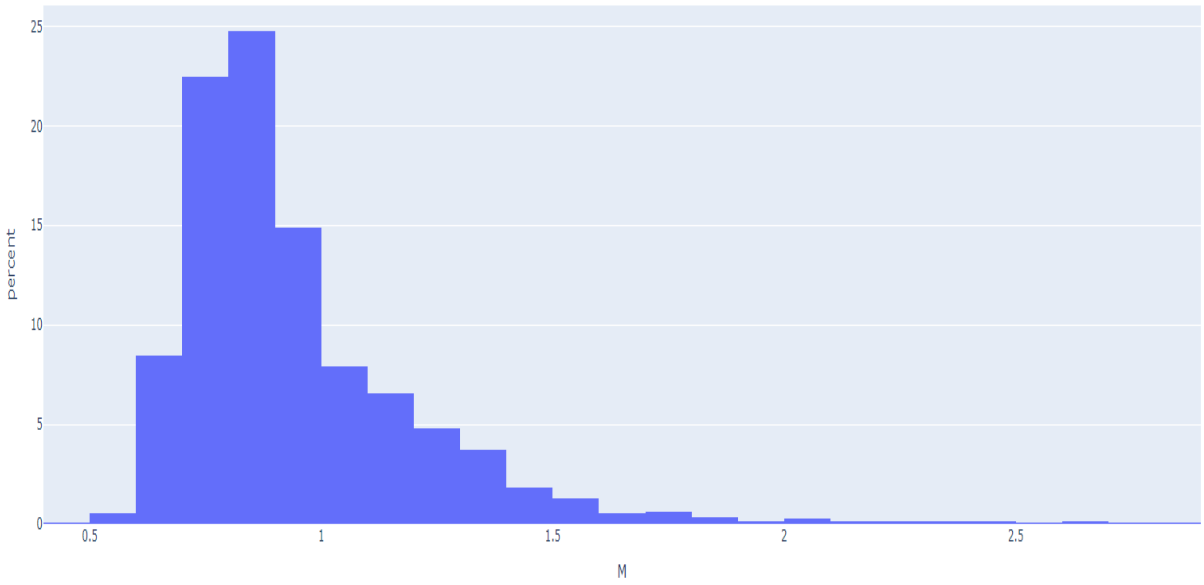

Figure 4: Histogram of M in BNC.

Figure 5: M and text domains (category). The addition of the horizontal and vertical values is M.

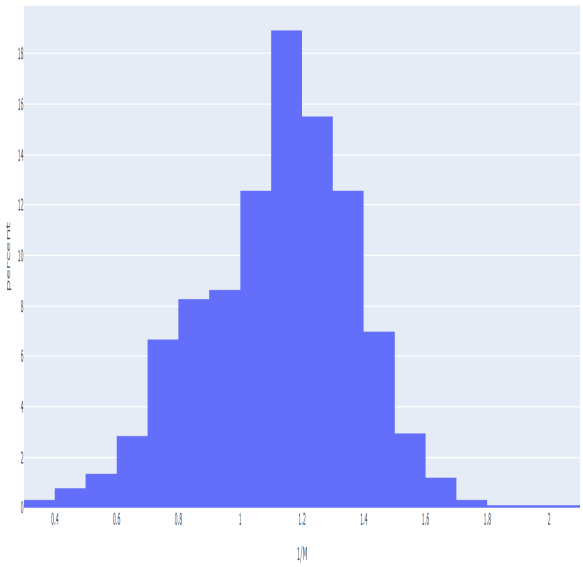

Figure 6: Histogram of $1/M$, which corresponds to the gain in the average number of syllables in a sentence within a year predicted by FKGL.

3.6 Domain Analysis of M

One of the primary traits of the BNC is its general nature, which implies that the corpus comprises various texts sourced from numerous topics. The genres of BNC texts are termed "*domains*," and approximately 4-out-of-3 out of these texts are tagged with domains [1]. In Figure 5, a scatterplot of bp_{sw} versus ap_{wl} in Equation 7 is presented. By integrating the horizontal and vertical axes, we derived the M factor as previously described. In Figure 5, it is shown that a domain confines text to a restricted area. Therefore, when a text's domain is fixed, the value of M remains more consistent and does not fluctuate significantly, rendering $\frac{1}{p_{\text{sl}}}$, the average number of syllables per sentence in Equation 7, the sole factor influencing text complexity.

3.7 The histogram of $1/M$

According to Equation 13, $1/M$ can be interpreted as the annual increase in the average number of syllables per sentence. We derived $1/M$ in the BNC and present its histogram in Figure 6. Interestingly, Figure 6 illustrates the distribution of the annual increase in the average number of syllables per sentence in the BNC, showing a peak at 1.2 and ranging between 0.4 and 2.0. To the best of our knowledge, this specific increase in text complexity, as evidenced by a measurable statistic via FKGL, has not been previously addressed. This is a significant finding of this study.

[1]We excluded the texts without domains from the entire experiments.

4 Discussions and Related Work

In this study, we focused exclusively on FKGL. However, as shown in Equation 7, the same logic applies to other readability formulas that are linear combinations of the average number of words per sentence and the average number of syllables per word. A well-known example of such a formula is FRE, which typically ranges from 0 to 100 for most texts.

$$\begin{aligned}\text{Reading Ease} &= 206.835 \\ &- 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) \\ &- 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right) \quad (14)\end{aligned}$$

According to Equation 14, by defining $a = -1.015$, $b = -84.6$, and $c = 206.835$ in Equation 5, we can obtain Equation 14. It is evident from the signs of a and b that higher FRE values indicate greater ease of readability; this is reasonable, given that FRE measures easiness while FKGL measures grade level, which correlates with difficulty. To prevent confusion arising from the contrasting natures of FRE and FKGL, we have focused exclusively on FKGL in this paper.

4.1 Applicability to Other Formulas

In this study, we focus on FKGL and FRE, or the Flesch–Kincaid formulas. This is because these formulas are notable examples of formulas consisting of only a linear combination of the average number of words in a sentence and the average number of syllables in a word. To the best of our knowledge, no other widely known formulas have this form.

However, some formulas have very *similar* forms. For example, the automated readability index (ARI) (Smith and Senter, 1967) (Equation 15) consists of a linear combination of the average number of words in a sentence and the average number of *characters* as follows: Using the same argument used in this study, ARI can be regarded as a measure simply based on the average number of *characters* in a *sentence* weighted by genre.

$$\begin{aligned} \text{ARI} &= -21.43 \\ &+ 0.5\left(\frac{\text{total words}}{\text{total sentences}}\right) \\ &+ 4.71\left(\frac{\text{total characters}}{\text{total words}}\right) \end{aligned} \quad (15)$$

As with ARI, the Coleman-Liau index (Coleman and Liau, 1975) consists of a linear combination of the average number of words in a sentence and the average number of characters in a word. However, the Coleman-Liau index requires an average of over 100 sentences and words. However, the same argument that we have addressed in this study also applies to the Coleman-Liau index. Other than these formulae, our reasoning should generally hold for formulas that are a linear combination of the average sentence length and per-word statistics.

Regarding cognitive implications, this study reveals that the FKGL and FRE formulas can be simply regarded as the number of syllables in a sentence weighted by genre. However, the relationship between the number of syllables in a sentence and reading comprehension remains unclear, and we show that this is an important open question. In addition, although the BNC is an excellent general corpus, it does not cover all text genres. The relationship between the number of syllables in a sentence and text genre is one of open questions. For second language learning, recent personalized readability studies (Ehara, 2022a,c,b; Liu et al., 2023) are also important for studying such relationships. Also, regarding FKGL and FRE, Ehara (2023) previously addressed that the reciprocal of a probability can be seen as a perplexity of tokens denoting delimiters of sentences or words.

5 Conclusions

This study makes significant contributions to the literature by examining the Flesch–Kincaid readability formulas, specifically FKGL and FRE. Unlike previous automatic readability assessment studies (Si and Callan, 2001; Collins-Thompson and Callan, 2005; Pitler and Nenkova, 2008; Vajjala, 2021; Martinc et al., 2021; Crossley et al., 2023), we demonstrate that the average number of syllables per sentence is a crucial determinant of text complexity in these formulas. Because readers' phonetic inventories are generally stable, our findings explain the enduring robustness of these formulas from a cognitive perspective.

Future research should focus on creating new robust readability formulas based on the average number of syllables in other languages. Although these formulas are widely used, their English specificity is a major limitation. Although FKGL has been adapted for some European languages, developing a readability formula for Asian languages remains challenging because of their distinct writing systems. Nevertheless, our analysis focused on syllables per sentence, a metric that is easily transferable to Asian languages. Based on our findings, we believe that an FKGL-equivalent readability formula can be developed for other Asian languages, which makes comparing readability between different languages possible.

Ethical Considerations

As our analysis relies on mathematical transformations, and our experiments utilize the BNC, a widely recognized and publicly accessible general English corpus, we believe that this study does not require any special ethical considerations.

Limitations

Although the BNC is a widely utilized general corpus and the corpus-linguistic analysis derived from it is broadly accepted, we acknowledge that our experiments relied on a single specific corpus. While we expect that experiments using general corpora would yield similar results across other general corpora, we did not conduct experiments using other corpora in this paper.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 22K12287 and by JST, PRESTO Grant Number JPMJPR2363.

References

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Kevyn Collins-Thompson and Jamie Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462.

Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2023. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55(2):491–507.

Yo Ehara. 2022a. No meaning left unlearned: Predicting learners' knowledge of atypical meanings of words from vocabulary tests for their typical meanings. In *Proc. of Educational Data Mining (short paper)*.

Yo Ehara. 2022b. Selecting reading texts suitable for incidental vocabulary learning by considering the estimated distribution of acquired vocabulary. In *Proc. of Educational Data Mining (poster paper)*.

Yo Ehara. 2022c. Uncertainty-aware personalized readability assessment framework for second language learners. *Journal of Information Processing*, 30:352–360.

Yo Ehara. 2023. A novel interpretation of classical readability metrics: Revisiting the. language model underpinning the flesch-kincaid index. In *Proc. of ICCE (Work-in-Progress Poster)*.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233. Place: US Publisher: American Psychological Association.

Joseph Marvin Imperial. 2021. Bert embeddings for automatic readability assessment. *arXiv preprint arXiv:2106.07935*.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore. Association for Computational Linguistics.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

J Peter Kincaid et al. 1980. Development and test of a computer readability editing system (cres). final report, june 1978 through december 1979.

Yuliang Liu, Zhiwei Jiang, Yafeng Yin, Cong Wang, Sheng Chen, Zhaoling Chen, and Qing Gu. 2023. Unsupervised readability assessment via learning from weak readability signals. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1324–1334.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, 47(1):141–179.

Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 574–576, New York, NY, USA. Association for Computing Machinery.

Edgar A. Smith and R.J. Senter. 1967. *Automated readability index*, volume 66. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air ....

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14.

Sowmya Vajjala. 2021. Trends, limitations and open challenges in automatic readability assessment research. *arXiv preprint arXiv:2105.00973*.