A comparison of text analysis tools: Levels of agreement and disagreement

Mart Christine Johnston
martchristinevito2017@gmail.com
Takushoku University

Abstract

This paper explores agreement levels among text analysis tools and factors influencing text difficulty using 75 Grade 2 EIKEN texts. EIKEN is a Japanese proficiency test for high school graduates. Text analysis included word count, average sentence length, CEFR, CEFR-J levels (determined by Text Inspector and CVLA, respectively), Flesch Reading Ease Score, Flesch-Kincaid Grade Level, Lexile score, and total coverage (from AntWordProfiler) using the New General Service List (NGSL). Average sentence lengths had stronger correlations with other indices than word counts. Flesch-Kincaid Grade Level and Lexile reading levels correlated moderately due to reliance on sentence length. According to both tools, the texts are generally deemed appropriate for grades 8-9 in the US education system. Considering that text levels for second language classrooms are typically several levels lower than those in the US education system and recognizing that EIKEN Grade 2 is intended for high school graduates in Japan, it can be inferred that the texts align with expectations for high school graduates in Japan, equivalent to grades 8-9 in the US education system. CEFR and CEFR-J levels, although their text level assignments were similar, had only moderate correlations, reflecting metric differences between Text Inspector and CVLA. AntWordProfiler's total coverage showed weak correlations, focusing solely on word frequency. The results from this study show clear discrepancies in text difficulty depending on the type of measure used and call for varied approaches to text analysis.

Keywords: text analysis, EIKEN, text difficulty, text measures, Text Inspector, CVLA, AntWordProfiler, NGSL

The aim of reading in one's first (L1) and second (L2) languages is similar: to find the meaning of the text (Nuttall, 1996). However, languages differ in many ways, such as orthographies, phonologies, and morphologies, which can affect how L2 readers process L2 texts (Grabe and Yamashita, 2022). Some ways to address this L2 reading issue are creating graded readers (Waring, n.d.) and developing basal reading programs (Ocampo, 1997). Although young children who are native speakers of English, for example, may benefit from graded readers as they begin to learn how to decode, L2 readers may struggle more with comprehension than L1 readers. As such, graded readers are designed to control some variables of text difficulty, such as vocabulary items, sentence structures, and even text length, to avoid overestimation of the types of texts and difficulty levels L2 readers can process.

Literature Review

While a wide variety of metrics for text difficulty are available, it is not clear if they all classify texts in the same way. Hermosa (2002) argues that different formulas are expected to obtain different scores when other variables are considered. Student perceptions were used by Holster et al. (2017) and Arai (2022) to prove that correlation of indices to difficulty varies significantly.

To test such claims, it would be of benefit to run the same texts through multiple schemes used to classify to see if the ratings they give are consistent. Online text analysis tools have made it easier to determine readability and assess difficulty in vocabulary use and sentence structures. This study seeks to evaluate the level of agreement among a number of online tools available and identify factors influencing text difficulty using EIKEN Grade 2 texts as its sample.

Measures of Text Difficulty

This literature review starts with an overview of EIKEN and its relationship with the CEFR, or Common European Framework of Reference for Languages. It is followed by the introduction of CEFR and CEFR-

J levels based on metrics provided by online analysis tools, namely Text Inspector (n.d.) and CEFR-based Vocabulary Level Analyzer ver. 2.0 or CVLA (Uchida, n.d.). Other readability formulas to be the discussed are Flesch Reading Ease (FRE), Flesch Kincaid Grade Level (FKGL), Lexile, and token coverage.

EIKEN and CEFR

Dunlea and Matsudaira (2009) determined student performance on Pre-1 and Grade 1 EIKEN tests with the abilities described at each level in the CEFR. Their results indicated that students who passed the Grade 1 test exhibited strong performance described at the CEFR C1 level, while those who passed the Grade Pre-1 test were at the CEFR B2 level. It is then assumed that EIKEN Grade 2 passers correspond to B1 or one level below. However, this assumption warrants further research for confirmation. The present study, however, does not examine test scores or test-takers' abilities, nor does it refer to the Council of Europe's Manual. Instead, it will use Text Inspector to analyze EIKEN Grade 2 passages. Developed by a professor of Applied Linguistics, Stephen Bax, this online tool was chosen for this research as it relies on the English Vocabulary Profile or EVP (English Profile, n.d.), which categorizes words according to difficulty (Text Inspector, n.d.).

Text Inspector's Scorecard (CEFR)

Text Inspector (n.d.) is an online analysis tool that can assign CEFR levels. As there are debates on CEFR reliability (Runnels, 2014; Hong et al., 2020), Text Inspector does not completely rely on the Framework. The CEFR level that Text Inspector assigns is based on over 200 metrics such as readability, lexical diversity, and lexical sophistication. To determine lexical sophistication, corpora such as the British National Corpus and the Corpus of Contemporary American English, and word reference tools such as the Academic Word List and the English Vocabulary Profile (EVP), were used. The EVP, for example, assigns CEFR levels from A1 to C2 to single and multi-word items in a text. However, Text Inspector utilizes only EVP's word list at the word level. CEFR A1 is subdivided into two sub-levels, and there are two additional levels, D1 and D2, for texts of a higher academic level. The estimated CEFR levels provided by Text Inspector can also be compared with the estimated CEFR-J levels given by CVLA, a text analysis tool tailored to English education in Japan.

CVLA (CEFR-J)

CEFR-based Vocabulary Level Analyzer ver.2, or CVLA, created by Prof. Uchida Satoru, is a free online tool that assigns CEFR-J levels to texts (Uchida, n.d.). This was designed specifically for Japanese students, as they tend to fall within different sections of the lower levels of CEFR. It relies on Tono's (2022) CEFR-J wordlist, developed from a corpus that contains items from primary and secondary school textbooks from China, Korea, and Taiwan (MEXT, 2012; Tono, n.d.). While the wordlist covers levels A1 to B2, CVLA extracts its C-level words from the EVP, which is also used as a reference by Text Inspector (n.d.). Factors including readability index, verbs per sentence, average word difficulty, and the ratio of B-level content words to A-level content words contribute to the text's CEFR-J level assignment. The text's average scores using these metrics are totaled and converted into one of 12 levels, from Pre-A1 to C2. Despite focusing on single-word items, CVLA's other indices compensate for this limitation. Therefore, this study will employ CVLA to analyze EIKEN Grade 2 texts.

Flesch Reading Ease and Flesch Kincaid Grade Level

Flesch Reading Ease (FRE) and Flesch Kincaid Grade Level (FKGL) metrics help distinguish between the comprehension ease of different texts (Wallace, 1992, p. 77). FRE and FKGL rely on word and

sentence lengths, remaining traditional yet widely used tools (Flesch, n.d.). FKGL, calculated by (.39 x words/sentences) + (11.8 x syllables/words), correlates with US class grades, indicating higher scores for more complex texts (Kincaid et al., 1975). FRE, 206.845 - (1.015 x words/sentences) - (84.6 x syllables/words), rates from extremely difficult to very easy (Flesch, 1948). Because of the similar variables present in their respective formulas, both tools show an almost perfect linear correlation (Štajner et al., 2012).

FRE scores can be categorized from extremely difficult to very easy, as shown in Table 1. The rating scale ranges from 0 to 100, where 0 indicates text that is nearly impossible to read, while 100 signifies very easy readability (Flesch, 1948, p. 230). The corresponding grade level is also provided for each readability score. As mentioned earlier, FRE scores are dependent on the sentence, word, and syllable counts.

Table 1
Flesch Reading Scores Translated to School Grades

Readability Scores	Description	Grade Levels (US system)
90-100	Very Easy	5th grade
80-90	Easy	6th grade
70-80	Fairly Easy	7th grade
60-70	Plain English	8th and 9th grade
50-60	Fairly Difficult	10th to 12th grade (high school)
30-50	Difficult	college
0-30	Very Difficult	college graduate

Regarding FKGL, scores are compared to the grade levels in the US education system. Text Inspector (n.d.) states that an FKGL score below 12 signifies easy readability for the general public, while a score below 8 suggests very easy text comprehension. Grade-level adjustments are therefore suggested for second/foreign language learners (Linguapress, n.d.). For example, some studies show that Japanese high school reading materials are equivalent to US grades 5-9 (Browne, 1998; Chujo & Hasegawa, 2004; Kukita & Fukuda, 2015), while Sugiura et al. (2020) found that high school textbook units can cover larger grade levels from 4th to 11th. EIKEN Grade 2 tests target high school graduates, indicating that texts should align similarly with FKGL levels below Grades 11-12 in the US system. Because of these comparisons, FKGL and FRE are appropriate for EIKEN Grade 2 analysis, whose reliability can also be cross-checked with Lexile scores.

Lexile Reading

EIKEN used MetaMetrics' online Lexile® Analyzer (2016) to measure the complexity of the test forms (reading sections) administered in 2013 and 2014 for all levels. The tool employs a Lexile framework based on word frequency and sentence length. According to MetaMetrics (2024), a score of 200L or below indicates a beginner level, while a score of 1200 or above suggests advanced proficiency. EIKEN Grade 2 tests were between 1000L and 1020L, indicating higher difficulty (MetaMetrics, 2016). Similar to FKGL and FRE, the Lexile Framework is designed primarily in the context of the US education system (MetaMetrics, 2022). For instance, Grade 8 US students typically encounter books with 1010L scores, rising to 1185L by year-end and 1300L in college. Given that the EIKEN Grade 2 texts are designed for high school graduates in Japan, this study will interpret Lexile grade levels similar to the approach in analyzing FKGL scores. It is anticipated that the grade levels corresponding to Lexile scores will not surpass Grade 11 or 12, equivalent to Grades 8-9 in the American school system.

Total coverage

Total coverage is dependent on word frequency. Word frequency affects readers' word recognition speed and text comprehension (Haberlandt & Graesser, 1985). Text comprehension becomes more difficult with more low-frequency or unknown words (Nation, 2013). Word frequency lists such as the NGSL (Browne et al., 2013) can be beneficial in material development. These lists are developed from corpora, such as the COCA (Corpus of Contemporary American English), which can rank words based on frequency.

Laufer (1989, p. 321) found that learners need to know at least 95% of the text's word tokens for comprehension. Hirsh and Nation (1992) later stated that a reader needs around 98% token coverage to read for pleasure, which Hu and Nation (2000, p. 419) reaffirmed in a later study. This study determines whether 95% of the vocabulary found in EIKEN Grade 2 texts is covered by the NGSL.

Correlations between readability indices

As mentioned earlier, this study aims to determine the correlations between the variables and formulas used in text analysis. As previous studies only focus on finding relationships among traditional readability formulas such as FRE and FKGL, this study will investigate the correlations among the following text analysis variables and formulas: word count, average sentence length, CEFR and CEFR-J levels (determined by Text Inspector and CVLA, respectively), Flesch Reading Ease Score, Flesch-Kincaid Grade Level, Lexile score, and total coverage (determined by AntWordProfiler and using the NGSL). It is hypothesized that indices sharing similar text measures will exhibit stronger correlation scores. Thus, the research question is: Will there be agreements among text analysis tools in terms of the results?

Method

Fifteen EIKEN Grade 2 reading tests collected from the EIKEN website (2023) were used in this study. Each reading test contains six parts labelled 1, 2A, 2B, 3A, 3B, and 3C. The actual labels in the test were not used in this study. Section 2A will be named Section A; 2B will be named Section B, and so forth. As mentioned earlier, the first section, with 20 short texts, is excluded from the analysis.

The gaps in the first and second gap-fill passages were completed with the omitted phrases or words. The answer choices not selected were removed. The comprehension questions for the fourth and fifth texts were not analyzed either. For every C text (an email), other parts of the email message, such as the subject and sender’s email address, were removed, and only the body text was analyzed. This was done to avoid a false count of the total number of sentences. The title of every text was also removed from the analysis. This study refers to the passages by their year, number, and letter, such as 2018-1A (section A of the first test administered in 2018).

Text Inspector was used to determine the following metrics: word counts and average sentence lengths, Flesch Reading Ease (FRE) scores, Flesch Kincaid Grade levels, and scorecards or CEFR levels (based on the metrics set by this online analysis tool). To analyze texts, one can either copy and paste them into the provided box or upload a .txt document. Then a mode is chosen for analysis (reading, writing, and listening). For this study, I selected the reading mode for text analysis.

Regarding the CEFR-J level of the texts, I used CVLA: CEFR-based Vocabulary Level Analyzer (ver. 2.0). Users can simply copy and paste the text for analysis. In contrast to Text Inspector, it has only two modes for text analysis: reading and listening. To avoid confusion, the average scores for calculating four textual features (readability index, verbs per sentence, average word difficulty, and the ratio of B-level content words to A-level content words) for each CEFR-J level displayed in the results were derived from previous studies (Uchida and Negishi, 2018). The "Input" section presents the scores obtained from the text being analyzed.

Lexile® Analyzer (2024) was employed to obtain the Lexile grades of the texts. In addition to the range scores and grade levels, it recommends books aligning with the estimated scores. Due to limited space, I recorded only the equivalent grade levels in the results section and moved the estimated Lexile range of the texts to *Appendix A.*

To generate vocabulary profiles for the texts, I used a free software tool called AntWordProfiler 2.0.1 (Anthony, 2022). NGSL (1000, 2000, 2800) and AWL (960) (Browne et al., 2013) were imported into the software as reference lists. The program calculates the percentage of the text covered by each list. Texts can either be copied and pasted on the given input screen or encoded as .txt format.

During the analysis, the difficulty levels of the five text sections of each test were compared. Texts with the lowest scores were generally considered the easiest (except for Flesch Reading Ease). In addition, Pearson's correlation scores were determined using JASP software to evaluate the level of agreement among all text measures. Non-numerical data, e.g., CEFR-J levels, were changed to numerical values for analysis. Later, I used these data for further discussion.

Results

This section will demonstrate the results using the text analysis tools introduced in previous sections. Table 2 presents Pearson's correlation coefficients between various text analysis measures. Abbreviations include WC (word count), ASL (average sentence length), FKGL (Flesch Kincaid Grade Level), LE (Lexile Grade Level), FRE (Flesch Reading Ease), FRG (FRE equivalent grade level), TC (total coverage), CEFR (CEFR levels by Text Inspector), and CEFRJ (CEFR-J levels by CVLA).

Please refer to *Appendix B* as it presents the 75 texts alongside their word counts and average sentence lengths, as determined by Text Inspector. The fourth and fifth columns display the grade levels assigned to the texts by Text Inspector (FKGL) and Lexile Analyzer, respectively. The FRE scores, also analyzed by Text Inspector, are provided along with their corresponding grade-level equivalents in the sixth and seventh columns. The eighth column indicates the total coverages resulting from analyzing the texts using AntWordProfiler. Finally, the ninth and last columns show the CEFR-J and CEFR levels determined by CVLA and Text Inspector, respectively.

Table 2
Pearson's correlation scores between the different text analysis indices

	WC	ASL	FKGL	LE	FRE	FRG	TC	CEFR
ASL	.29*	-						
FKGL	.49***	.69***	-					
LE	.38***	.68***	.72***	-				
FRE	-.49***	-.43***	-.95***	-.61***	-			
FRG	.34**	.35**	.86***	.5***	-.92***	-		
TC	-.26*	.16	-.05	-.17	.14	-.06***	-	
CEFR	.44***	.36**	.65***	.45***	-.66***	.60***	-.13***	-
CEFRJ	.33**	.38***	.64***	.43***	-.64***	.57***	-.03***	.64***

**p < .05, ** p <.01, *** p < .001*

Discussion

To recall, the research question is: Will there be agreements among text analysis tools in terms of the results? First, this section will discuss the results displayed in *Appendix B.* Variables such as word count and sentence length are examined for their impact on text difficulty. The study compares text analysis measures that share similar variables: Lexile and FKGL; CEFR levels assigned by Text Inspector and CEFR-J levels by CVLA. FRE and total coverage are also compared because Strauss et al. (2007) found a strong correlation between word length and word frequency, and both tools assess the readability of texts. The section concludes with interpretations of the correlation scores (Table 2) among text analysis indices.

Ranges of word counts and sentence lengths

The word counts (column 2 of *Appendix B)* for texts A and B range between 255 to 276, while the email texts or C texts have the lowest word count, ranging from 201 to 240. The longest texts are D and E, with word counts ranging from 334 to 373. However, average sentence lengths do not have similar patterns to determine which texts receive the highest or lowest scores (column 3). The range of average sentence lengths of the texts are 13.00 and 22.99. Longer sentences can make a text more difficult, as they may contain dependent clauses, creating compound, complex, or complex-compound sentences. However, a longer sentence could also provide explanations or extra information to help the reader. Nevertheless, the longest sentences with scores close to 22.99 were predominantly from texts A, B, D, and E. Only a few C texts (18.0-19.0) fall into this range. This suggests that average sentence length is not consistently reliable in determining text difficulty. Moreover, based on the results, word count and average sentence length do not consistently align.

Ranges of Lexile Grade Levels and Flesch Kincaid Grade Levels

Determining the FKGL (column 4 of *Appendix B*) involves factors such as word length, sentence length, and syllable count of a text. Twenty-six texts scored between 8.00 and 8.99, suggesting suitability for average students in American Grades 8 to 9. The average FKGL of all 75 texts is 8.63, closely correlating with the FRE scores (columns 6 and 7). Individual text FKGLs ranged from 5.41 to 11.82, covering up to six academic year levels. However, none exceed the 11.82-grade level assignment, with texts 2019-1E, 2019-3A, and 2022-3E identified as the most challenging and designated for American Grade 11 students. The texts span a broader range of reading levels, from Grades 3 to 12 using the Lexile Analyzer. Lexile scores, determined by sentence length and word frequency (MetaMetrics, 2022), show that longer sentences and fewer familiar words correlate with higher difficulty levels. Among them, 46 texts are considered readable for Grades 5 to 12 in the US, with Lexile scores ranging from 1010 to 1200L (Please refer to *Appendix A*), while 27 are suitable for Grades 3 to 7, ranging from 810 to 1000L (column 5). Notably, 2019-3A and 2019-3B stand out with Lexile scores of 1210 to 1400, intended for Grades 10 to 12 students in the US system. According to MetaMetrics (2024), US high school textbooks typically have around 1100L, while postsecondary texts usually have 1300L. The EIKEN Grade 2 texts' average Lexile scores fall between 943.33 and 1133.33L, generally suitable for students in Grades 5 to 10 in the US. These students are expected to read books with an average score of 943.33L at the beginning of the year, increasing to 1133.33L by the end.

Comparison reveals that both FKGL and Lexile Analyzer generally assess EIKEN texts within Grades 11 and 12 in the US system, with only a few reaching those levels. Both measures concur that texts 2019-2C, 2019-3C, and 2022-3B are the easiest, with FKGLs ranging from 5.41 to 5.78 and Lexile levels averaging Grades 4-5. Texts 2019-3A and 2019-3B pose the greatest difficulty according to the Lexile Analyzer, placing them in Grades 10-12, which aligns with FKGLs of 11.82 and 10.09 respectively. These results

suggest that both indices utilize sentence length as a factor in scoring, potentially resulting in similar grade level assignments for texts.

Ranges of the Flesch Reading Ease scores and total coverage percentages

As previously discussed, FRE calculation involves three aspects: sentence count, word/token count, and syllable count. Out of 75 texts, 44 scored between 60 and 60.99, suitable for Grades 8-9 in the US system, akin to the Flesch Reading average of 8.63. Column 6 of *Appendix B* shows FRE score ranges across sections, with equivalent grade levels in Column 7. C texts, typically emails, are the easiest, with scores between 62.17 and 83.47, fitting Grade 6-9 readers. The word count may have played a significant role in this outcome, as each email text contains between 201 and 240 words. On the other hand, longer texts (A, B, D, E) suit students from 7th grade to college. Sections A and E may be more challenging. Only 2019-1E and 2019-3A were rated college-level. However, none of the texts were rated as readable for college graduates (0-30 FRE scores), suggesting that the texts might have been simplified or controlled to remain below a certain difficulty level.

The majority of the text reading scores align with 8th and 9th grades in the US education system, supporting EIKEN's assertion that Grade 2 texts are intended for high school graduates. This assumption stems from the belief that English books created in non-English speaking countries may be comparatively easier than those produced for native English speakers. Therefore, disparities in English materials between the US and Japan may span three to four grade levels. This is evident in the Flesch-Kincaid Grade and Lexile grade levels.

While Flesch Reading Ease scores are based on word, syllable, and sentence counts, total coverages are based on word frequency. Integrated with the NGSL, AntWordProfiler forms the vocabulary profiles of texts (Column 8). Only 27 out of 75 texts achieved 95% total coverage, suggesting many contain unfamiliar words, especially in science and healthcare topics. The results further show that 8 out of 15 text C or email texts are more readable due to higher high-frequency word usage. There were five texts in each of sections A, B, and E, and four texts in section D, all of which achieved 95% or higher in total coverage. Therefore, 41 texts across these categories do not exhibit readability solely based on total coverage. Additionally, comparing readability based on grade levels and word frequency may lead to the assumption that texts intended for Grade 8-9 students (US education system) generally have coverages below 95% and include more low-frequency or unfamiliar words, an aspect teachers should consider.

Ranges of the Text Inspector's Scorecard (CEFR) and CVLA- CEFR-J levels

Columns 9 and 10 of *Appendix B* show CEFR and CEFR-J levels determined by Text Inspector and CVLA. Both tools consider factors that include readability, word count, and sentence length, referring to corpora and word reference lists such as the BNC and EVP to categorize words into different difficulty levels. CVLA utilizes its own corpora for A and B levels but references EVP for C vocabulary.

According to both Text Inspector and CVLA, C texts are generally less challenging, not surpassing B2 and B2.1. CVLA assigns subcategories A2.1 and A2.2 to align with Text Inspector's A2+. Most texts labeled A2.1 and A2.2 by CVLA were also rated easiest by Text Inspector. Both analysis tools also agreed that 2019-2C was the easiest text. If CVLA A2.2 level will be considered, both tools also agreed that 2022-1C and 2022-3B can be categorized as very easy. For A and B texts, one per section received A levels from both tools. Specifically, 2022-3A got A2+ from both, while 2022-3B was A2+ by Text Inspector and A2.2 by CVLA.

The majority of the A, B, D, and E texts fell into CEFR/CEFR-J levels B or C. Notably, four texts—2018-3A, 2019-3A, 2019-3B, and 2022-3E—were labeled very difficult by both Text Inspector (B2+) and

CVLA (C1), suggesting alignment between B2+ and C1. CVLA, designed for English education in Japan, classified several texts at a C1 level, suggesting they may be overly challenging for Japanese learners. Therefore, CVLA deems what Text Inspector finds moderately difficult as very difficult. Despite this, no texts were categorized as C2, typically associated with news articles and academic papers. This trend mirrors Text Inspector's findings, which similarly did not identify any EIKEN Grade 2 texts at levels C1, C2, or D. Most texts were rated B level with Text Inspector assigning B2 to 22 texts, B1+ to 21 texts, and B1 to 17 texts, while CVLA rated 27 texts as B2.1, 10 as B2.2, 9 as B1.2, and 8 as B1.1. Factors such as word count likely influence difficulty, as longer texts (A, B, D, and E) were assigned higher CEFR or CEFR-J levels. Educators can note the presence of challenging vocabulary items in these texts. Wordlists such as the CEFR-J wordlist (Tono, 2022) serve as valuable references for identifying vocabulary items belonging to the B levels.

Correlations among Indices

The correlation between FKGL and word count is moderate (0.488), indicating that texts with more words tend to have higher FKGL scores. Conversely, there is a moderate negative correlation of -0.485 between word count and FRE, indicating that longer texts generally have lower FRE scores, suggesting increased difficulty. The Lexile scores and FKGL demonstrate strong to very strong negative correlations with FRE (-0.608 and -0.949 respectively), indicating that higher grade levels correspond to lower FRE scores, suggesting more challenging readability. Additionally, stronger positive relationships exist between FKGL and Lexile (0.722), FKGL and ASL (0.694), and Lexile and ASL (0.677), surpassing other indices in correlation strength. These scores shed light on why FKGL, Lexile, and FRE assigned similar grade levels to texts. However, the ASL's correlation with FRE is only moderate at -0.432. Despite this, ASL correlates better with other text measures than word counts, contrary to the earlier assumption that word count is the primary predictor of text difficulty.

Total coverage (TC) displays weak correlations with other text measures. Despite the assumed strong correlation between word length and frequency (Strauss et al., 2007), total coverage lacks even a moderate correlation with FRE (0.136) or Lexile scores (-0.169). This weak correlation likely stems from AntWordProfiler's sole focus on word frequency, neglecting sentence structure. Consequently, the previous assumption that texts for US Grade 8-9 students typically have coverages below 95% and include more low-frequency items remains inconclusive.

The analysis of CEFR levels by Text Inspector reveals strong correlations with other text measures. For instance, a strong linear relationship (0.654) exists between CEFR and FKGL, likely influenced by Text Inspector's comprehensive assessment of texts using 200 metrics, including readability. Additionally, CEFR shows a strong correlation with CEFR-J (0.639), possibly because both tools utilize wordlists to categorize vocabulary items. It is worth noting that CEFR-J (CVLA) and CEFR (Text Inspector) exhibit almost identical correlation scores with other measures. To recall, CVLA employs factors such as readability index, verbs per sentence, average word difficulty, and the ratio of B-level to A-level content words for text analysis. However, the correlation score between CEFR and CEFR-J is only at 0.639. This is unexpected, as one might anticipate a very strong correlation similar to that of FRE and FKGL (-0.949). The correlation between Text Inspector and CVLA could be attributed to the differing factors used in their text analyses. Conversely, FRE and FKGL, despite employing distinct formulas, share common measurement factors such as word, syllable, and sentence counts, which could explain their very strong correlation.

Limitations and Future Research

The outcomes of this study were informed by a purely quantitative approach to text analysis. While this quantitative analysis is valuable, teachers should also incorporate qualitative analysis to strengthen their judgment in selecting materials for proficiency test preparation. The approach in this research also had a number of limitations, for example, excluding comprehension questions, answer choices, and specific text elements may have influenced the study outcomes as assessed by text analysis tools. Not understanding the questions would of course present difficulty issues for test takers. Furthermore, this research focused on a small subset of EIKEN tests. Furthermore, it focused on EIKEN tests at a specific level. Analyzing texts from multiple EIKEN levels (e.g., EIKEN 1, pre-1, pre-2, 3, 4, and 5) alongside EIKEN 2 would offer a broader perspective on how text analysis tools perform across varying proficiency levels.

Another possibility for future research would be a separate and focused study on multi-word items present in texts that can affect text difficulty. This would help to avoid the limitations present when many of the analysis tools focus mainly on word count and variety. Additionally, a subsequent study where the findings could be validated by testing students using the same subject texts used. Careful analysis of student performance would provide important insight into the text difficulty.

Conclusion

The analysis conducted in this study provides valuable insights into the readability and difficulty levels of Grade 2 EIKEN texts intended for Japanese high school graduates, addressing the research question of whether there are agreements among various text analysis tools regarding the assessment of text difficulty. Word counts varied significantly among sections, with the email passages emerging as the easiest due to their lower word counts, while the longer texts presented the most challenging reading sections. While average sentence length does not follow such clear patterns, examining correlation scores reveals that ASL has a stronger linear relationship with other indices such as Flesch Kincaid and Lexile grades. Notably, there were results from obtaining FKGL and Lexile reading levels that were similar, most probably because both tools depend on sentence length. It is important to note that only a few texts reached grades 11 and above (US system). Most texts align with grades 8-9 in the US education system, as determined by both tools. Given that text levels for second language classrooms are typically several levels lower than those in the US education system and considering that EIKEN Grade 2 is designed for high school graduates in Japan, it can be inferred that the texts meet the expected proficiency levels for high school graduates in Japan, equivalent to grades 8-9 in the US education system. Texts used for language learning purposes are generally expected to be easier to read than those used in native language education. Regarding FKGL and FRE, they are considered to have the strongest correlation at -0.949, likely due to the similarities of textual factors being measured: syllable, sentence, and word counts, despite using different formulas. Similarly, the CEFR and CEFR-J levels of the texts (assigned by Text Inspector and CVLA respectively) were generally comparable, with email passages deemed the easiest and longer texts considered difficult, though the correlation is only 0.639. This discrepancy may be attributed to the differing factors utilized in their text analyses. Despite this, both tools exhibited almost identical correlation scores with other measures. Looking at the majority of the results of the texts B2 (Text Inspector) and B2.1 (CVLA) scores, teachers may consider finding texts with such levels when preparing students. Using these tools, the research also found out that there were vocabulary items that the Text Inspector found moderately difficult, whereas CVLA found them very difficult for Japanese learners. Regarding the analysis of total coverages, it was revealed that total coverage has weak correlations with other text measures. While total coverage focuses solely on word frequency, it does not consider other important factors such as sentence structure, leading to its limited utility in assessing text difficulty. Overall, the findings of this study highlight the need for educators and curriculum developers to consider

multiple factors when evaluating text difficulty. Hermosa (2002) suggests educators consider student variables, such as their responses using techniques like the cloze procedure, and encourages assessments and opinions provided by teachers regarding the learning materials. By integrating various text analysis tools and methodologies, educators can make more informed decisions when selecting texts for language learning purposes, ultimately enhancing the effectiveness of language instruction.

Acknowledgements

I would like to thank Professors Jonathan Rees of the University of Birmingham and Jeffrey Stewart of Tokyo University of Science/Takushoku University, as well as all the reviewers of the JALT Testing and Evaluation SIG, for their guidance and advice when writing this paper.

Declaration of competing interests:

M. Johnston has declared no competing interests.

References

Anthony, L. (2022). *AntWordProfiler* (Version 2.0.1*)* [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Arai, Y. (2022). Exploring perceived difficulty of graded reader texts. *Reading in a Foreign Language, 34*(2), 249-270. http://hdl.handle.net/10125/67425

Browne, C. (1998). Japanese high school textbooks: How readable are they? *Temple University Japan Working Papers in Applied Linguistics, 12,* 1-13.

Browne, C., Culligan, B., & Phillips, J. (2013). The new general service list: A core vocabulary for EFL students and teachers. *JALTs The Language Teacher, 34*(7), 13-15.

Chujo, K., & Hasegawa, S. (2004). Goi no cover ritsu to readability kara mita daigaku eigo nyushi mondai no nanido [Assessing Japanese college qualification tests using JSH text coverage and readability indices]. Nihon University Student Faculty of Engineering Research Report B, 37, 45-55. http://www.cit.nihon-u.ac.jp/laboratorydata/kenkyu/publication/journal_b/b37.5.pdfCVLA: CEFR-based Vocabulary Level Analyzer (ver. 2.0). (2023). *Home*. https://cvla.langedu.jp/

Dunlea, J., & Matsudaira, T. (2009). Investigating the relationship between the EIKEN tests and the CEFR. *Linking to the CEFR levels: Research perspectives,* 103-110.

EIKEN. (2023). Research. https://www.eiken.or.jp/eiken/en/recognition/

English Profile. (n.d.). *Wordlists.* https://www.englishprofile.org/wordlists

Flesch, R. (n.d.). *How to Write Plain English.* University of Canterbury: Management, Marketing, and Entrepreneurship. https://web.archive.org/web/20160712094308/http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml

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

Grabe, W., & Yamashita, J. (2022). Cambridge Applied Linguistics. Cambridge University Press.

Haberlandt, K. and Graesser, A. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General, 114*(3), 357-374. https://doi.org/10.1037/0096-3445.114.3.357

Hermosa, N. (2002). *The Psychology of Reading.* University of the Philippines -Open University.

Hirsh, D. and Nation, I.S.P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language 8(*2), 689-696.

Holster, T. A., Lake, J. W., & Pellowe, W. R. (2017). Measuring and predicting graded reader difficulty. *Reading in a Foreign Language, 29*(2), 218–244. https://nflrc.hawaii.edu/rfl/item/377

Hong, J. F., Tseng, H. C., Peng, C. Y., & Sung, Y. T. (2020). Linguistic Feature Analysis of CEFR Labeling Reliability and Validity in Language Textbooks. Journal of Technology & Chinese Language Teaching, 11(1).

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430.

Kukita, D., & Fukuda, M. (2015). On the Readability of the English Technical Writing Test: With Special Reference to the Textbooks Used in Technical High Schools and Colleges of Technology in Japan. *Bulletin of Miyazaki Municipal University Faculty of Humanities, 22*(1), 251-260.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans to thinking machines* (pp. 316–323). Multilingual Matters.

Linguapress. (n.d.). *Flesch-Kincaid Readability Test.* Linguapress. https://linguapress.com/teachers/flesch-kincaid.htm

MetaMetrics. (2016). *Lexile® Measurement of Tests: EIKEN & Test of English for Academic Purposes.* MetaMetrics.https://metametrics.s3.amazonaws.com/public/dynamic/international/pdfs/EikenText_Measurement_Report_Digital.pdf

MetaMetrics. (2022). *Lexile® Framework for Reading: Development and Validity Evidence.* MetaMetrics. https://hubsupport.lexile.com/Images/Lexile%20Framework%20for%20Reading%20Validity%20Evidence_2022.pdf

MetaMetrics. (2024). *The Lexile Framework of Reading.* https://hub.lexile.com/analyzer

MEXT: Ministry of Education, Culture, Sports, Science and Technology. (2012*). CEFR-based framework for ELT in Japan.* https://www.mext.go.jp/b_menu/shingi/chousa/shotou/092/shiryo/__icsFiles/afieldfile/2012/09/24/1325972_2_1.pdf

Nation, I. S. P. (2013). *Teaching & learning vocabulary.* Heinle Cengage Learning.

Nuttall, C. (1996). *Teaching reading skills in a foreign language*. Heinemann.

Ocampo, S. (1997). *Trends in Reading Instruction.* University of the Philippines- Open University.

Runnels, J. (2014). An exploratory reliability and content analysis of the CEFR-Japan's A-level can-do statements. *JALT journal, 36*(1), 69-89.

Štajner, S., Evans, R., Orasan, C., & Mitkov, R. (2012). What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility* (pp. 14-22).

Strauss, U., Grzybek, P., & Altmann, G. (2007). Word length and word frequency (pp. 277-294). Springer Netherlands.

Sugiura, R., Imai, N., Hamilton, M., Dean, E., & Ashcroft, R. (2020). Input and Output in Japanese High School Government-Approved English Textbooks. *Journal of Higher Education Tokai University, 21*, 1-16.

Text Inspector. (n.d.). *Analyse*. https://textinspector.com

Tono, Y. (2022). *CEFR-J wordlist.* http://www.cefr-j.org/download.html

Tono, Y. (n.d.). *Using corpora for reference level descriptions of the CEFR and the CEFR-J* https://languages-cultures.uq.edu.au/files/28901/Using-corpora-for-reference-level-descriptions.pdf

Uchida, S. (n.d.) CVLA: CEFR-based Vocabulary Level Analyzer (ver. 2.0).

Uchida, S., & Negishi, M. (2018). Assigning CEFR-J levels to English texts based on textual features. In Y. Tono & H. Isahara (Eds.), *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference* (pp. 463-467).

Wallace, C. (1992). *Reading.* Oxford University Press.

Waring, R. (n.d.). Writing a graded reader. Retrieved June 29, 2024, from https://www.er-central.com/authors/writing-a-graded-reader/writing-graded-readers-rob-waring/#google_vignette

Appendix A

Lexile Ranges of Texts Per Test

Texts	Lexile Range Min	Lexile Range Max
2018-1	810.00	1200.00
2018-2	810.00	1200.00
2018-3	810.00	1200.00
2019-1	1010.00	1200.00
2019-2	810.00	1200.00
2019-3	810.00	1400.00
2020-1	810.00	1200.00
2020-2	810.00	1200.00
2020-3	810.00	1200.00
2021-1	810.00	1200.00
2021-2	810.00	1200.00
2021-3	810.00	1200.00
2022-1	810.00	1200.00
2022-2	810.00	1200.00
2022-3	810.00	1200.00

Appendix B

Text Scores Derived from Index Calculations

1	2	3	4	5	6	7	8	9	10
Texts	Text Inspector	Text Inspector	Text Inspector	Lexile Analyzer	Text Inspector	Flesch	AntWord Profiler	CVLA	Text Inspector
	Word count	Average sentence length	Flesch Kincaid grade levels	Lexile grades	Flesch Reading Ease scores	FRE grade levels	Total coverage	CEFR-J levels	CEFR levels
2018-1A	270	15.88	8.26	3-7	64.13	8-9	96.3	B2.1	B2
2018-1B	274	15.22	7.79	3-7	66.34	8-9	93.0	B1.2	B1
2018-1C	201	18.27	7.03	3-7	77.17	7	98.6	B1.2	B1
2018-1D	344	18.11	8.93	5-12	63.28	8-9	92.1	B2.1	B1+
2018-1E	373	16.2	8.45	5-12	63.36	8-9	93.7	B2.2	B2+
2018-2A	265	15.59	10.08	5-12	50.55	10-12	83.5	B1.2	B2
2018-2B	260	18.64	9.85	5-12	57.51	10-12	94.7	C1	B2
2018-2C	221	14.73	6.12	3-7	77.42	7	95.1	A2.2	B1
2018-2D	354	16.86	7.62	3-7	70.47	7	96.1	B2.1	B1
2018-2E	356	18.74	10.25	5-12	54.98	10-12	86.3	C1	B1+

1	2	3	4	5	6	7	8	9	10
Texts	Text Inspector	Text Inspector	Text Inspector	Lexile Analyzer	Text Inspector	Flesch	AntWord Profiler	CVLA	Text Inspector
	Word count	Average sentence length	Flesch Kincaid grade levels	Lexile grades	Flesch Reading Ease scores	FRE grade levels	Total coverage	CEFR-J levels	CEFR levels
2018-3A	264	18.86	10.18	5-12	55.67	10-12	94.0	C1	B2+
2018-3B	274	16.12	7.97	5-12	66.66	8-9	90.5	B2.1	B1
2018-3C	227	17.46	7.13	3-7	75.07	7	94.0	B2.1	B1
2018-3D	362	17.24	8.34	5-12	65.94	8-9	93.1	B2.1	B1+
2018-3E	372	16.91	10.45	5-12	50.26	10-12	92.0	B2.1	B1+
2019-1A	267	15.71	7.82	5-12	67.00	8-9	87.5	B2.1	B2+
2019-1B	274	17.12	8.23	5-12	66.57	8-9	92.2	B2.1	B1
2019-1C	226	18.83	8.25	5-12	69.43	8-9	96.3	B1.2	B1
2019-1D	334	16.70	9.01	5-12	60.20	8-9	94.1	B2.1	B1+
2019-1E	365	18.25	11.25	5-12	46.92	college	97.3	C1	B2
2019-2A	276	18.40	9.37	5-12	60.65	8-9	94.9	B2.1	B1+
2019-2B	255	18.21	8.73	5-12	64.93	8-9	90.1	B2.1	B1+
2019-2C	229	15.27	5.41	3-7	83.47	6	95.9	A2.1	A2+
2019-2D	358	17.90	9.16	5-12	61.29	8-9	96.1	B1.1	B1+
2019-2E	361	18.05	10.54	5-12	*51.65*	10-12	93.6	C1	B2
2019-3A	267	19.07	11.82	10-12	44.26	college	94.7	C1	B2+

1	2	3	4	5	6	7	8	9	10
Texts	Text Inspector	Text Inspector	Text Inspector	Lexile Analyzer	Text Inspector	Flesch	AntWord Profiler	CVLA	Text Inspector
	Word count	Average sentence length	Flesch Kincaid grade levels	Lexile grades	Flesch Reading Ease scores	FRE grade levels	Total coverage	CEFR-J levels	CEFR levels
2019-3B	271	18.07	10.09	10-12	54.89	10-12	91.6	C1	B2+
2019-3C	230	14.38	5.46	3-7	81.53	6	90.7	B2.1	A2+
2019-3D	359	18.89	10.22	5-12	55.45	10-12	96.4	B2.2	B2
2019-3E	356	22.25	10.36	5-12	60.44	8-9	98.0	B2.2	B2
2020-1A	260	16.25	8.31	3-7	64.42	8-9	89.9	B1.1	B1
2020-1B	261	17.40	9.33	5-12	59.19	10-12	94.7	B2.2	B1+
2020-1C	213	16.38	8.03	3-7	66.68	8-9	97.7	B2.1	B1
2020-1D	362	18.10	10.18	5-12	54.32	10-12	95.7	B1.2	B1+
2020-1E	356	18.74	8.85	5-12	64.96	8-9	96.0	B1.2	B2
2020-2A	267	16.69	9.39	5-12	57.45	10-12	92.2	B2.2	B2
2020-2B	265	18.93	10.27	5-12	55.14	10-12	95.9	C1	B2
2020-2C	210	16.15	7.79	3-7	67.97	8-9	93.3	B2.1	B1
2020-2D	367	17.48	8.46	5-12	65.54	8-9	91.8	B1.1	B2
2020-2E	357	14.88	9.15	3-7	*55.95*	10-12	88.6	B2.1	B2+

1	2	3	4	5	6	7	8	9	10
Texts	Text Inspector	Text Inspector	Text Inspector	Lexile Analyzer	Text Inspector	Flesch	AntWord Profiler	CVLA	Text Inspector
	Word count	Average sentence length	Flesch Kincaid grade levels	Lexile grades	Flesch reading ease scores	FRE grade levels	Total coverage	CEFR-J levels	CEFR levels
2020-3A	268	14.11	7.61	3-7	65.62	8-9	92.5	B2.1	B2
2020-3B	266	17.73	8.54	5-12	65.43	8-9	97.4	C1	B2
2020-3C	222	15.86	7.23	3-7	71.46	7	94.2	A2.1	B1
2020-3D	349	14.54	7.56	3-7	66.75	8-9	90.2	B2.2	B2
2020-3E	343	17.15	10.50	5-12	50.32	10-12	90.6	C1	B2
2021-1A	275	15.28	9.08	3-7	57.20	10-12	96.5	B2.1	B1+
2021-1B	260	16.25	8.77	5-12	61.16	8-9	92.3	B2.1	B1+
2021-1C	214	19.45	9.42	5-12	62.17	8-9	97.2	B2.1	B1+
2021-1D	363	18.15	8.65	5-12	65.36	8-9	91.5	B1.2	B1+
2021-1E	362	16.45	8.89	5-12	60.66	8-9	93.9	C1	B1+
2021-2A	265	18.93	8.85	5-12	65.35	8-9	92.1	B1.1	B1
2021-2B	271	15.06	7.26	3-7	69.80	8-9	98.9	B1.2	B1
2021-2C	221	13.81	6.94	3-7	69.93	8-9	95.3	A2.1	B1
2021-2D	354	16.09	8.72	5-12	61.21	8-9	85.5	B2.2	B1+
2021-2E	359	14.36	8.22	3-7	61.71	8-9	95.5	C1	B2

1	2	3	4	5	6	7	8	9	10
Texts	Text Inspector	Text Inspector	Text Inspector	Lexile Analyzer	Text Inspector	Flesch	AntWord Profiler	CVLA	Text Inspector
	Word count	Average sentence length	Flesch Kincaid grade levels	Lexile grades	Flesch Reading Ease scores	FRE grade levels	Total coverage	CEFR-J levels	CEFR levels
2021-3A	274	17.12	8.62	5-12	63.79	8-9	89.7	B2.1	B1
2021-3B	270	19.29	9.02	5-12	64.75	8-9	94.0	B2.2	B2+
2021-3C	234	15.60	7.14	3-7	71.69	7	92.7	B1.1	A2+
2021-3D	362	18.10	8.78	5-12	64.37	8-9	93.2	B1.2	B1+
2021-3E	358	15.57	9.33	3-7	55.87	10-12	86.0	B2.1	B2
2022-1A	263	16.44	8.81	5-12	61.16	8-9	90.8	B2.2	B1+
2022-1B	272	14.32	6.91	3-7	71.00	7	88.9	B2.1	B2
2022-1C	234	15.60	6.58	3-7	75.67	7	97.4	A2.2	A2+
2022-1D	359	19.94	9.58	5-12	61.93	8-9	87.7	B2.1	B2+
2022-1E	352	18.53	9.44	5-12	60.41	8-9	94.5	C1	B2
2022-2A	260	15.29	8.39	3-7	62.13	8-9	97.5	C1	B2
2022-2B	276	17.25	8.62	5-12	63.96	8-9	92.6	B1.1	B1
2022-2C	221	18.42	8.63	5-12	66.03	8-9	93.2	B2.1	B1+
2022-2D	362	17.24	8.60	5-12	64.07	8-9	83.6	B2.2	B1+
2022-2E	365	16.59	8.18	5-12	65.99	8-9	88.2	B1.1	B2
2022-3A	269	14.16	6.21	3-7	75.79	7	93.0	B1.1	A2+
2022-3B	262	13.10	5.78	3-7	76.97	7	91.9	A2.2	A2+

1	2	3	4	5	6	7	8	9	10
Texts	Text Inspector	Text Inspector	Text Inspector	Lexile Analyzer	Text Inspector	Flesch	AntWord Profiler	CVLA	Text Inspector
	Word count	Average sentence length	Flesch Kincaid grade levels	Lexile grades	Flesch Reading Ease scores	FRE grade levels	Total coverage	CEFR-J levels	CEFR levels
2022-3C	240	16.00	7.61	3-7	68.98	8-9	93.7	B2.1	B2
2022-3D	361	21.24	10.08	5-12	60.61	8-9	88.5	B2.1	B1+
2022-3E	362	22.62	11.06	5-12	56.04	10-12	96.1	C1	B2+