Evaluating the consistency of automated CEFR analyzers: a study of English language text classification

Piyathat Siripol, Seongha Rhee, Suthathip Thirakunkovit, Aphiwit Liang-Itsara
Faculty of Liberal Arts, Mahidol University, Nakhon Pathom, Thailand

Article Info

Article history:

Received Nov 5, 2024
Revised Apr 3, 2025
Accepted May 9, 2025

Keywords:

Analyzer tools
Automated text analysis
CEFR
Text difficulty classification
Web based analysis

ABSTRACT

With the increasing use of web-based tools for text analysis, there is a growing reliance on automated systems to assess text difficulty and classify texts to the Common European Framework of Reference for Languages (CEFR). However, inconsistencies in these tools' outputs could undermine their effectiveness for language learners and researchers. This study investigates the consistency of five widely used automated CEFR analyzer tools, including ChatGPT, by analyzing 20 English descriptive texts at CEFR levels B1 and B2. A quantitative approach was employed to compare the CEFR classifications generated by these tools. The results reveal significant inconsistencies across the tools, raising concerns about the reliability of automated CEFR alignment. Additionally, the content and genre of texts appeared to influence the CEFR classification, suggesting that certain factors beyond the tools' algorithms may affect their accuracy. These findings have important implications for language educators, curriculum designers, and researchers who rely on automated CEFR tools for text selection, grading, and analysis. The study highlights the limitations of automated CEFR classification systems and calls for a more qualitative approach to text difficulty alignment analysis. Future research recommendation is discussed and call for more focus on refining these tools and exploring additional factors that may impact their effectiveness in text difficulty measurement and CEFR alignment.

Corresponding Author:

Aphiwit Liang-Itsara
Faculty of Liberal Arts, Mahidol University
999 4th Rd. Monthon Phutthamonthon District, Nakhon Pathom 73170, Thailand
Email: aphiwit.lia@mahidol.edu

1. INTRODUCTION

The purpose of this study is to explore the consistency of automated text analyzers that claim to align text difficulty to various proficiency levels. The measurement of English text difficulty has evolved over the past centuries, beginning with readability formulas proposed in the 20th century such as the Flesch-Kincaid and the Gunning-Fox Index [1], [2]. These formulas primarily rely on quantitative calculations of linguistic features, including syntactic and semantic complexity, vocabulary familiarity, syllable counts, and the number of words per sentence. While these readability formulas have been extensively critiqued and tested for validity, they continue to shape perceptions of text difficulty [3], [4].

Advancements in natural language processing (NLP) and computational linguistics have significantly improved these classic readability formulas, introducing innovative techniques and models for text difficulty analysis [4]. For instance, theories of text comprehension have been integrated into new readability formulas, such as those introduced by Crossley *et al.* [4], which employ advanced NLP tools to produce more promising results compared to traditional methods. As a result, automated online instruments have become increasingly accessible to language teachers, text producers, and researchers, enabling them to assess text difficulty efficiently. Although users may have varying criteria for selecting these tools, they share a common purpose: benchmarking texts to appropriate difficulty levels to inform language teaching, assessment, and skill development. This process, known as material leveling, aims to align texts to specific difficulty levels for targeted use [5].

The Common European Framework of Reference for Languages (CEFR): learning, teaching, assessment serves as a guideline to provide coherence and standardization in material leveling. Initially developed to standardize textbooks, curricula, and examinations in Europe, the CEFR framework classifies language proficiency into basic (A1-A2), independent (B1-B2), and proficient (C1-C2) users [6]. Its straightforward descriptors and clear communicative goals have facilitated its adoption worldwide as a foundation for language teaching and assessment [7].

Educational institutions and government organizations have recognized and embraced CEFR as a guideline for expected language proficiency outcomes. For instance, the Commission on Higher Education Standards (CHES) under the Office of the Permanent Secretary of Thailand set out CEFR policy in 2024, specifying that undergraduate students should achieve at least B2 proficiency, while graduate students should attain C1 before graduation. This policy has driven many higher education institutions to align their language teaching and assessment practices with CEFR descriptors [8]–[10]. However, the application of CEFR remains challenging for test developers, material creators, and educators, as its descriptors are open to varying interpretations. Moreover, CEFR does not prescribe specific teaching methodologies or assessments, leaving educators to adapt and integrate its framework into their unique contexts. For example, CEFR-leveled texts often require modification to suit the needs of language learners [11].

Given the critical role of text classification or material leveling in language education, educators must ensure that selected texts align accurately with learners' proficiency levels [12]. These challenges highlight the importance of exploring automated text classification tools that claim to map with specific CEFR descriptors, as reliable tools are essential for effectively supporting language learning and material selection. The novelty of this paper lies in its focus on the role of automated CEFR alignment tools at the text level, an area that has received limited attention compared to broader CEFR alignment practices. By uncovering the variability in CEFR classifications across different tools, the study highlights the need for caution in employing computational methods, ensuring they align with the specific purposes of CEFR classification. Furthermore, it emphasizes the importance of integrating human judgment to enhance the reliability of text difficulty assessments. These findings provide valuable insights for educators, policymakers, and developers seeking to improve the effectiveness of CEFR alignment tools.

2. LITERATURE REVIEW

2.1. Text classification based on CEFR guideline

Text classification under CEFR guidelines has gained increasing importance in recent years [13]. Evaluating and selecting appropriate texts remains challenging for language teachers. Traditional text classification approaches often rely on shallow linguistic features, which do not fully reflect the actual level of text difficulty. Consequently, various methods have been developed to enhance and simplify text classification based on difficulty levels.

Online resources offering limitless access to authentic texts have also prompted the integration of model text recommendation systems into classrooms to better engage language learners [14]. Research by Escobar-Acevedo *et al*. [14] suggested employing automatic classifiers based on CEFR guidelines to streamline the selection of texts from authentic sources, such as online platforms [14]. However, attempts to classify texts using automated tools have often yielded inconsistent results [15].

Recent advancements include the use of machine learning techniques such as the bidirectional encoder representations from transformers (BERT) model for CEFR text classification [13]. Studies demonstrate that automated text-level detection is effective, with the quality of classification largely dependent on the text corpus used. For instance, Schmalz and Brutti [16] proposed an alternative approach using a BERT-based model, reporting success in automated CEFR text classification supported by significant numerical data [16]. However, the practical application of these models in language classrooms remains challenging. BERT-based systems require extensive manual data labeling for training, as the pre-trained model is based on plain text from sources like the English Wikipedia and the Brown Corpus. This process is labor-intensive and may be impractical for teachers with varying levels of technical literacy [17].

Other studies propose combining quantitative and qualitative approaches for text classification under CEFR guidelines. Natova [15] introduced three quantitative tools: i) The Lexile framework for reading; ii) The Flesch reading ease index; and iii) The lexical profiler in Text Inspector. For qualitative assessment, manual analysis incorporating human judgment of text meaning, structure, grammatical complexity, and cognitive demand was employed. While Natovas findings [15] indicate that such tools hold potential, the classification levels heavily appeared to be influenced by qualitative criteria. This raises questions about the consistency and precision of automated tools in leveling or benchmarking texts to CEFR levels. Quantitative assessments, as shown by Natova [15], illustrate the variability of CEFR classifications across different automated tools, potentially leading to inaccurate interpretations of text difficulty. The inconsistency stems from the varying frameworks and algorithms underlying these tools, necessitating further exploration. This gap in the literature highlights concerns for educators, material developers, and text writers [18].

2.2. Automated CEFR text classification tools

Currently, numerous automated tools are available for teachers and material developers to benchmark text difficulty against CEFR levels. These tools facilitate quick analysis and provide results that guide material development and classroom practices. Several tools, such as Cathoven AI, Text Inspector, and Pearson's Global Scale of English (GSE), are widely utilized as validation tools in language research.

2.2.1. Cathoven AI

Cathoven AI's CEFR Checker evaluates texts by analyzing linguistic factors such as vocabulary, grammar, sentence structure, and phrases. A unique feature is its ability to provide decimal-level precision for CEFR classifications. For instance, a text might be classified as A2.7, indicating it is close to B1. However, the tool does not provide algorithmic transparency in explaining how these decimal values are derived and may leave users without insight into the underlying algorithms. Despite this limitation, Cathoven AI allows users to explore detailed CEFR reports for vocabulary, tense, clause, and sentence-level classifications. Studies have employed Cathoven AI for text difficulty analysis and CEFR benchmarking, highlighting its potential applications [19], [20].

2.2.2. Global Scale of English

GSE provides a granular proficiency scale from 10 to 90, aligning with CEFR levels. Initially used for the Pearson Test of English Academic (PTE Academic), it was later validated against standardized tests such as international English language testing system or IELTS and test of English as a foreign language or TOEFL. The GSE toolkit offers features for text analysis, including CEFR equivalency points, readability measures (e.g., ARI, Flesch-Kincaid, and SMOG), and highlighted vocabulary exceeding current CEFR levels. This tool has been particularly effective for assessing text difficulty with consistency and fine-grained insights into proficiency levels.

2.2.3. Text Inspector

Text Inspector is a web-based tool that employs a corpus-based approach, utilizing resources such as the English Vocabulary Profile (EVP), Academic Word List, and British National Corpus (BNC). It calculates metrics like type-token ratio, sentence length, and syllable count to determine text difficulty. A notable feature is its Lexical Profile, which aggregates various metrics to assign a CEFR level. Despite its comprehensive capabilities, the reliability of results depends on the number and combination of metrics used [21].

2.2.4. Oxford Text Checker

Oxford Text Checker focuses on vocabulary profiling by examining text content against extensive corpora, including the Oxford 3000 and Oxford 5000-word lists. Unlike other tools, it does not provide an overall CEFR level but offers percentages of words corresponding to different CEFR levels. This feature helps users estimate the proficiency level of a text, although its application is limited to vocabulary-based analyses. The inclusion of wordlists into each wordlist group, such as Oxford 3000, has undergone three core criteria: frequency, range, and familiarity. The details are summarized in Table 1.

Table 1. Summary of Oxford Text Checker wordlist

Category of wordlist	Detail of wordlist	CEFR
Oxford 3000	A list of the 3,000 core words that English learners need to know	A1-B2
Oxford 5000	An expanded wordlist from the core wordlist for advanced English learners	B2-C1
Oxford phrase list	650 common phrases (idioms, phrasal verb, collocations, common complementation patterns, multi-headwords, alternative forms of head words) for learners	A1-C1
Oxford phrasal academic lexicon	Important words to learn for writing and speaking academic English from Oxford Corpus of Academic English (OCAE) and British Academic Spoken English (BASE)	Not applicable

Oxford Text Checker highlights words that are from different levels of CEFR. With the GSE toolkit, words are not color-coded for CEFR level compared to the Oxford Text Checker. The advantage of the Oxford Text Checker is that it provides percentages of words that are in the list from A1 to C1. Using this tool, the user will have an overview of statistics that most words in the text are that of A2 or B2 level.

2.2.5. Road to grammar

Road to grammar (RTG) is an anonymous online website created by a language teacher with more than 20 years of experience teaching English as a second and foreign language in Asian countries. The website offers ESL materials, games, quizzes, and resources for ESL teachers. An online tool called 'text analyzer' (http://www.roadtogrammar.com/textanalysis/) is also available for ESL teachers to use to determine the appropriate level of text difficulty for different levels of students' proficiency according to the CEFR. The detail of the tool, however, is quite limited compared to other text analyzer instruments, explaining briefly that the analyzer focuses on the complexity of language. The submitted text will be compared to a list of the 10,000 most used words in English together with the position on the list and the average word and sentence length. The web page does not contain the type of corpus used or a reference as to which word list is a reference point. With an easy-to-use interface, however, users can insert a text and simply get a rating of the CEFR A2 to C2.

2.2.6. ChatGPT

ChatGPT, a sophisticated language model developed by OpenAI, has also been used to classify texts by CEFR level. Leveraging advanced NLP techniques, ChatGPT evaluates linguistic features like vocabulary sophistication, grammatical complexity, and sentence structure. Unlike traditional tools, it uses transformer-based architectures trained on diverse multilingual corpora, enabling it to provide nuanced assessments. ChatGPT's inclusion in this study highlights the potential of NLP-based tools to complement traditional methods.

While existing automated tools are widely used, there is limited research that examines their consistency in generating CEFR classifications. Differences in algorithms and frameworks may result in variable outputs, posing challenges for educators and researchers. For instance, tools like Text Inspector rely on lexical profiling, while ChatGPT employs advanced machine-learning techniques. This study addresses concerns about whether current CEFR analyzer programs provide consistent results when classifying leveled texts, aiming to inform their effective use in language teaching and assessment.

3. METHOD

The current study employed a quantitative data analysis approach by identifying congruence across text analyzer tools. Each tool was designed specifically for language analysis purposes. Details of the sample texts and data analysis are provided in the following sections.

3.1. Sample texts and texts recruitment

This study includes 20 text samples from https://lingua.com that offer reading texts of various CEFR levels in more than ten languages, including English. The website was chosen as it claims to offer authentic reading materials designed by experienced English educators to support vocabulary recognition and variation of sentence structure reflected by difficulty levels (novice, intermediate, and advanced). CEFR of A1 to C1 is used to indicate the difficulty levels of basic users, independent users, and proficient users. On Lingua.com, readers can evaluate their progress and understanding through multiple-choice quizzes offered at the end of reading materials. In this study, the English language texts were collected in a group of 10 CEFR B1 texts and 10 CEFR B2 texts. There are 20 freely available texts with 66 additional texts for subscribed members for B1, and there are two freely available texts with 33 additional texts for subscribed members for B2. A descriptive genre was used for this study. The criteria for text recruitment for this research study are shown in Table 2.

Table 2. Criteria for text recruitment

Criteria	Detail
Accessibility and copyright	An article must be accessible to public online offering both a basic (free) account and/or paid account options Copyright of an article does not prohibit readers/users to analyze the discourse content
Text source	An article must be from the same source
CEFR level indication	An article that has clear indicator of B1 or B2
Minimum word count	An article that contains a word count of at least 150 words
Text structure and genre	Descriptive texts

3.2. CEFR analyzer instruments

To arrive at the CEFR result for each text, CEFR B1 and B2 texts indicated by the publisher/author were computed into five CEFR text analyzer programs and one large language model tool (ChatGPT 3.5). Five widely used automated CEFR text analyzers chosen for this study were Cathoven AI, text analyzer (RTG), GSE (Pearson), Text Inspector and Oxford Text Checker (Oxford Learner Dictionary). These online tools are accessible online.

Cathoven AI provides many automated text analysis features such as a readability analyzer and CEFR analyzer. In the CEFR analyzer, a user can generate text output results with a general CEFR level breakdown from vocabulary, verb form, and clause. Moreover, the tool could also provide level mapping of provided text to the American Council on the Teaching of Foreign Languages (ACTFL), British Council, IELTS, and Cambridge Scale. The vocabulary breakdown includes vocabulary knowledge demands. The Oxford 5000 wordlist was used in this study as it is an expanded core wordlist for advanced learners of English, including the 2000 words for learners at B2-C1.

Overall, ten texts of B1 and ten texts of B2 were computed for CEFR level. This accounts for a sum of twenty texts to be automatically analyzed. As most CEFR analyzer tools provide only quantitative results based on different metrics, the descriptive account of a text is often limited. This is because classification approaches of CEFR texts could be from a corpus-based approach or a complex mathematical model readability assessment [22]. Hence, ChatGPT 3.5 as one of the advanced language models was added since it has been trained on a vast corpus of texts and could provide complex mathematical models such as Flesch Kincaid score to arrive at a particular CEFR level. Therefore, it was included as one of the automated tools to analyze text and classify texts into CEFR level. The prompt given to the ChatGPT 3.5 was “analyze a text given and classify it into CEFR level.”

3.3. Data analysis

The analysis involves an evaluation of the consistency and accuracy of the CEFR text classification tools by using descriptive statistical measures for each tool. For each of the 20 text samples, the CEFR levels assigned by the six web-based tools (ChatGPT 3.5, Cathoven AI, RTG Text Analyzer, GSE, Text Inspector, and Oxford Text Checker) are recorded. Ten texts used for descriptive statistical analysis are based on the availability of text from the same source of Lingua.com. To facilitate numerical analysis, CEFR levels are assigned numerical values: A1 as 1, A2 as 2, B1 as 3, B2 as 4, C1 as 5, and C2 as 6. Using descriptive statistics is essential in this process as the mean, median, mode, and standard deviation (SD) of the assigned CEFR levels are calculated for each tool to summarize datasets into basic measures, making it easier to compare and interpret the performance of each tool. The mean provides an average proficiency level assigned by each tool which could offer a benchmark for comparison. The median indicates the central tendency of the assigned levels, and the mode can give the most frequently assigned levels, which can show where tools most commonly agree. SD measures the variation or dispersion of the assigned CEFR levels, indicating the consistency of each tool's classifications. It must be noted, however, that the study does not intend to apply purely descriptive statistics as a major analysis method since: i) the number of texts is limited to only 10 texts and ii) each automated tool uses a different algorithm and framework to analyze the texts automatically. As a result, the generated CEFR level and text content are used to qualitatively analyze the result.

4. RESULTS AND DISCUSSION

Table 3 and Figure 1 present the CEFR levels assigned by each tool for the ten B1 texts. The mean CEFR level assigned by ChatGPT 3.5 was 3.20, indicating a B1 level, with a median of 3.00 and a mode of 3.00, showing a consistent classification within the B1 range. The SD of 0.42 suggests that it has low variability in the classifications by Chat GPT3.5.

Cathoven AI version 2.0 assigned a mean CEFR level of 3.5 (B1+), suggesting that most texts were classified within the upper-intermediate range of independent users. The median and mode were both recorded at 3.00 (B1), indicating that the most frequent classification remained at the lower end of the B1 level. The SD of 0.70 reveals a higher degree of variation in the assigned levels, suggesting less consistency in comparison to other analyzers. This variability implies that Cathoven AI may classify texts with more fluctuation, potentially influencing its reliability for precise CEFR alignment.

The GSE Text Analyzer (Pearson) consistently assigned a mean CEFR level of 4.05 (B2), indicating that most texts were classified at the upper-intermediate level. Both the median and mode were recorded at 4.00 (B2), confirming that the majority of classifications remained stable at this proficiency level. The SD of 0.15 suggests minimal variation, demonstrating that the tool produced highly consistent results across different texts. This consistency indicates that the GSE Text Analyzer provides a reliable benchmark for B2-level text classification with little fluctuation in its assessments.

Text Inspector assigned a mean CEFR level of 5.50 (C1+), indicating that most analyzed texts were classified within the advanced proficiency range. Both the median and mode were also recorded at 5.50 (C1+), showing that the most frequent and central classifications remained at the same level. The SD of 0.37 suggests that while there was some variation in the assigned levels, the classifications did not fluctuate drastically. Overall, this result indicates that the Text Inspector consistently rated the texts at an advanced level with only minor deviations in classification.

The Text Analyzer (RTG) assigned a mean CEFR level of 4.70 (B2+), suggesting that the majority of texts were classified within the upper-intermediate to advanced range. The median and mode were both recorded at 5.00 (C1), indicating that the most common classification placed texts at the proficient level. The SD of 0.48 reflects moderate variability, meaning that while some texts were rated lower or higher, the tool maintained a relatively stable classification trend. This pattern suggests that RTG tends to assign CEFR levels in a range between B2+ and C1, demonstrating a degree of consistency but with noticeable fluctuations.

The Oxford Text Checker provided only word percentage data rather than assigning an overall CEFR level to the texts. The results indicated that the majority of words in the analyzed texts were classified at the A1 level, reflecting basic vocabulary usage. Additionally, words at the B1 level accounted for only 5% to 8% of the total word count, suggesting limited lexical complexity. This data implies that the Oxford Text Checker tends to classify texts at a lower proficiency level based primarily on word frequency rather than broader linguistic features.

Table 3. Automated CEFR analyzer for B1 texts

Title of text	Total word count	ChatGPT 3.5	Cathoven AI Version 2.0 (general level)	GSE Text Analyzer (Pearson)	Text Inspector (overall lexical profile)	Text Analyzer (RTG)	Oxford Text Checker (Oxford Learner Dictionary) Wordlist Oxford 5000
The Empire State Building	247	3 (B1)	3 (B1)	4 (B2)	5.5 (C1+)	5 (C1)	A1 (71%) (With B1 at 5%)
San Francisco	260	4 (B2)	3 (B1)	4 (B2)	5 (C1)	4 (B2)	A1 (71%) (With B1 at 4%)
Chicago	236	4 (B2)	3 (B1)	4 (B2)	5 (C1+)	5 (C1)	A1 (66%) (With B1 at 6%)
The Statue of Liberty	248	3 (B1)	4 (B2)	4 (B2)	6 (C2)	5 (C1)	A1 (69%) (With B1 at 8%)
The Grand Canyon	252	3 (B1)	4 (B2)	4 (B2)	5.5 (C1+)	4 (B2)	A1 (69%) (With B1 at 8%)
Los Angeles	219	3 (B1)	4 (B1)	4.5 (B2+)	5.5 (C1+)	5 (C1)	A1 (61%) (With B1 at 6%)
Miami	253	3 (B1)	3 (B1)	4 (B2)	5.5 (C1+)	5 (C1)	A1 (70%) (With B1 at 2%)
Washington, D.C.	192	3 (B1)	3 (B1)	4 (B2)	5.5 (C1+)	4 (B2)	A1 (65%) (With B1 at 2%)
Yellowstone National Park	268	3 (B1)	3 (B1)	4 (B2)	6 (C2)	5 (C1)	A1 (69%) (With B1 at 4%)
Valentine's Day	171	3 (B2)	5 (C1)	4 (B2)	6 (C2)	5 (C1)	A1 (63%) (With B1 at 4%)
Mean (CEFR level)	-	3.20 (B1)	3.5 (B1+)	4.05 (B2)	5.50 (C1+)	4.70 (B2+)	-
Median (CEFR level)	-	3.00 (B1)	3.00 (B1)	4.00 (B2)	5.50 (C1+)	5.00 (C1)	-
Mode (CEFR level)	-	3.00 (B1)	3.00 (B1)	4.00 (B2)	5.50 (C1+)	5.00 (C1)	-
SD	-	0.42	0.70	0.15	0.37	0.48	-

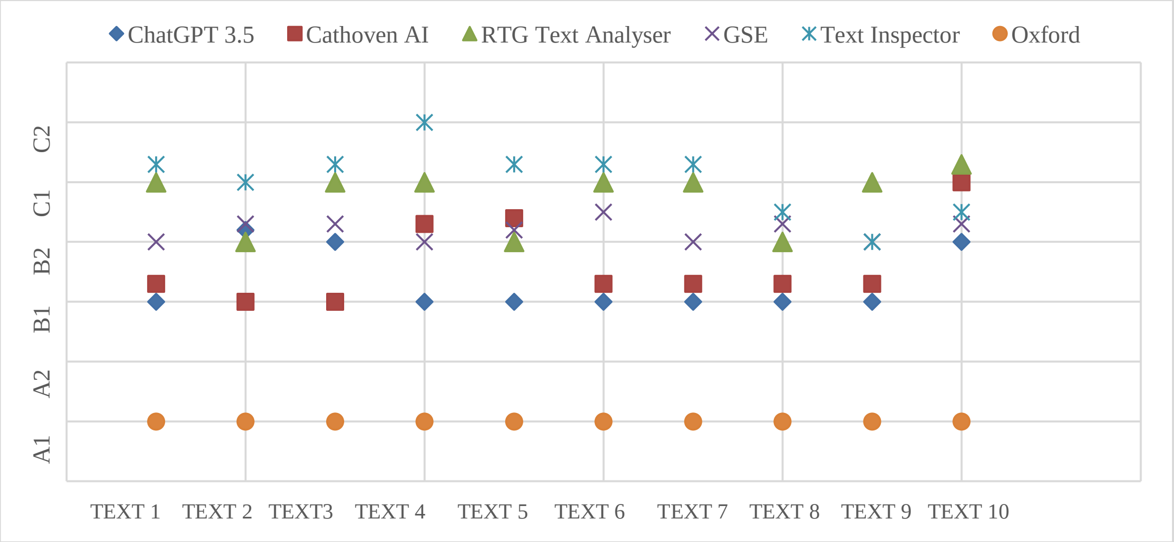

Figure 1. Automated CEFR analyzer mapping for B1 texts

Table 4 and Figure 2 show the CEFR levels assigned by each tool for the ten B2 texts. ChatGPT 3.5 had a mean CEFR level of 3.50 (B1/B2), with a median of 3.50 (B1/B2) and a mode of 3.00 (B1). The SD was 0.53, indicating moderate variability.

Cathoven AI Version 2.0 assigned a mean CEFR level of 3.40 (B1), indicating that most texts were classified within the lower-intermediate range. The median and mode were both recorded at 3.00 (B1), suggesting that the majority of classifications clustered around the B1 level. The SD of 0.52 reflects moderate variability, meaning that some texts were rated slightly above or below this level. This degree of fluctuation suggests that while Cathoven AI generally aligns with the B1 classification, its assessments may not always be entirely consistent.

The GSE Text Analyzer assigned a mean CEFR level of 4.05 (B2), suggesting that most texts were consistently classified within the upper-intermediate proficiency range. Both the median and mode were recorded at 4.00 (B2), confirming that the majority of classifications aligned closely with this level. The SD of 0.15 indicates very low variability, meaning that the tool produced highly stable and reliable results. This consistency suggests that the GSE Text Analyzer is a dependable resource for assessing text difficulty at the B2 level with minimal fluctuation.

Text Inspector assigned a mean CEFR level of 5.55 (C1+), indicating that most texts were classified at an advanced proficiency level. Both the median and mode were recorded at 5.50 (C1+), suggesting a strong concentration of classifications at this level. The SD of 0.50 reflects some variability, but overall, the tool maintained a consistent classification within the C1+ range.

The Text Analyzer (RTG) assigned a mean CEFR level of 4.25 (B2), suggesting that most texts were classified at the upper-intermediate level. Both the median and mode were recorded at 4.00 (B2), reinforcing the consistency of this classification. The SD of 0.52 indicates moderate variability, meaning that some texts were rated slightly higher or lower than B2.

Lastly, the Oxford Text Checker provided only word percentage data rather than assigning a specific CEFR level. The results showed that most words in the analyzed texts were classified at the A1 level, indicating a basic vocabulary range. This suggests that the tool tends to classify texts at a lower proficiency level based primarily on word frequency.

The key results of this study indicate substantial inconsistency among various automated CEFR text classification tools, highlighting significant challenges in the field of text difficulty and assessment and its applications in educational settings. The result of the study is in line with Natova [15] regarding the use of a quantitative framework in mapping CEFR level of a text using quantitative measures. In Natova's [15] study, three quantitative tools were proposed relating to the Lexile Framework for Reading and the Lexile Analyzer, the Flesh Reading Ease index and the lexical profiler Text Inspector. Compared to Natova's [15] study, however, additional web-based tools designed specifically to map text difficulty aligning to a specific CEFR level are investigated such as the AI Cathoven and Pearson GSE. These inconsistencies raise critical questions about the reliability and accuracy of these automated tools, which are essential for teachers, material developers, and language assessment professionals.

Table 4 Automated CEFR analyzer for B2 texts

Title of text	Total word count	ChatGPT 3.5	Cathoven AI Version 2.0 (general level)	GSE Text Analyzer (Pearson)	Text Inspector (overall lexical scale)	Text Analyzer (RTG)	Oxford Text Checker (Oxford Learner Dictionary) Wordlist Oxford 5000
Las Vegas	214	3 (B1)	3 (B1)	4 (B2)	5 (C1)	4 (B2)	A1 (69%) (With B2 at 4%)
Liechtenstein	152	3 (B1)	3 (B1)	4 (B2)	5.5 (C1+)	4 (B2)	A1 (70%) (With B2 at 5%)
Trip to New York	222	3 (B1)	3 (B1)	3.5 (B1+)	4.5 (B2+)	3 (B1)	A1 (85%) (With B2 at 1%)
Santa Claus around the World	180	3 (B1)	4 (B2)	4.5 (B2+)	5 (C1)	4 (B2)	A1 (70%) (With B2 at 2%)
Boston	260	4 B2	3 (B1)	4.5 (B2)	5.5 (C1+)	4 (B2)	A1 (68%) (With B2 at 5%)
Cruise Ships: Floating Marvels of Marine Life	233	3 (B1)	3 (B1)	4.5 (B2+)	5.5 (C1+)	6 (C2)	A1 (67%) (With B2 at 10%)
Benjamin Franklin and the Kite	202	4 (B2)	4 (B2)	4 (B2)	6 (C2)	5 (C1)	A1 (59%) (With B2 at 6%)
Hot Air Balloons: A Fun History	264	4 (B2)	4 (B2)	4 (B2)	6.5 (C1+)	5 (C1)	A1 (73%) (With B2 at 7%)
The longest and most scenic Train Ride	199	4 (B2)	4 (B2)	4 (B2)	5 (C1)	4 (B2)	A1 (72%) (With B2 at 2%)
Animal	323	4 (B2)	3 (B1)	4 (B2)	6 (C2)	5 (C1)	A1 (78%) (With B2 at 5%)
Mean (CEFR level)	-	3.50 (B1+)	3.40 (B1)	4.05 (B2)	5.55 (C1+)	4.25 (B2)	-
Median (CEFR level)	-	3.50 (B1+)	3.00 (B1)	4.00 (B2)	5.50 (C1+)	4.00 (B2)	-
Mode (CEFR level)	-	3.00 (B1)	3.00 (B1)	4.00 (B2)	5.50 (C1+)	4.00 (B2)	-
SD	-	0.53	0.52	0.15	0.50	0.52	-

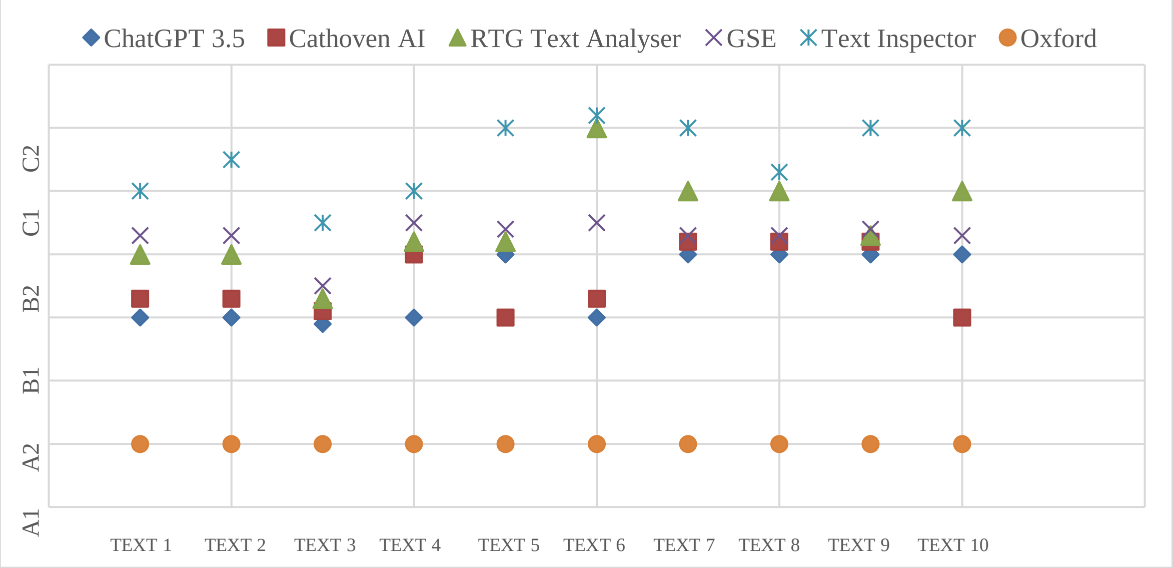

Figure 2. Automated CEFR analyzer mapping for B2 text

The variability in CEFR levels assigned by different tools for the same texts underscores the complexity of automated text classification. For instance, the ChatGPT 3.5 model and Cathoven AI often provided different CEFR levels compared to traditional tools like the GSE Text Analyzer and Text Inspector. This variability can be attributed to the distinct algorithms and linguistic features each tool emphasizes. ChatGPT, for example, utilizes advanced NLP techniques, whereas GSE and Text Inspector rely on more traditional readability metrics and lexical profiling frameworks. The mean CEFR levels assigned by the tools show a range from B1 to C1+ for B1 texts and from B1 to C2 for B2 texts. Such broad ranges indicate a lack of consensus among the tools, which could lead to challenges in standardizing text difficulty levels across different educational contexts for teachers, material developers, and even policymakers. Alignment validity is a concern for CEFR [23] and can be a challenge in providing coherent results. For instance, it is still questionable whether different readers at the same B2 level can perform all the tasks described in A1 and A2. An ongoing discussion of whether to take linguistic quality and quantity as a measure of language ability is still in question. There is no empirical evidence to support that language learners at a B2 level can do all the A2 tasks to reach B2. Hence, more research is needed to explore different dimensions of CEFR that are still unknown. For instance, it could be seen from the findings that the topic or content of a passage may influence the overall outcome of CEFR level by automated instruments.

It is noteworthy to mention that across these ten B1 texts, 9 out of 10 stories are about ‘location’, with one about an event (Valentine’s Day*)*. Examining the automated instruments used, it appears that most automated instruments map one level more advanced for the CEFR. For instance, Chat GPT 3.5, an NLP tool with a machine learning algorithm, generated a text about Valentine’s Day to be at the level of B2 whereas the majority of other tools map the passage to be at a B1 level. Similarly, the result from the B2 texts shows that when a passage is about a ‘location’, Chat GPT 3.5 tends to map them to be at the B1 level, with historical text passages (Benjamin Franklin and the Kite and Hot Air Balloons: A Fun History) and with historical events (The longest and most scenic Train Ride) to be at the B2 level. Comparing the result between ChatGPT 3.5 and the lexical profile framework by Text Inspector from the B1 passages, Text Inspector maps an overall CEFR of a story about valentine’s day to be at C2 level where the majority passages relating to ‘location’ are at the C1 and C1+ level. Similar to the passages from B2 texts, the lexical profile framework by Text Inspector maps the overall CEFR level of content about ‘location’ to be at B1+ to C1. However, when the passages are about ‘animal’ or ‘history’, the result changes to C2 (Animal and Benjamin Franklin and the Kite). Here, it is possible that content and words used within different text genres may influence the overall CEFR level, similar to the context of previous studies [24], [25].

Not only does the traditional lexical profile framework from Text Inspector show an interesting result that may be influenced by what the story is about in generating the overall CEFR but also the Cathoven AI version 2.0. From B1 texts, Cathoven AI generated 9 out of 10 texts about a location to be at B1-B2 level with only one to the C1 level. Similar to the previous automated instruments, Cathoven AI generated an overall CEFR level for valentine’s day to be C1. Similar to the group of B1 texts, Cathoven AI appears to rate most B2 texts as B1 with only a few at B2, especially when the topic is related to historical text types. This shows that the text about valentine's day may contain words, phrases, or grammatical structures that may be more complex or infrequent. Compared to all automated instruments, Text Inspector maps a text about valentine's day to be at the highest level of CEFR at C2. To examine further, the scorecard from Text Inspector was investigated which shows that the overall CEFR determined by the lexical sophistication from the EVP and BNC is at C2+. This indicates that although a corpus-based approach may provide users with fast and consistent results in generating text difficulty based on the CEFR guidelines, the accuracy of the generated CEFR indicator for a particular text needs to be carefully interpreted. Previous studies have explored different aspects of CEFR linguistic features using a corpus-based approach, but most relates to identifiable features, not conceptual association [26]–[28].

The GSE Text Analyzer by Pearson provides a very consistent result (SD=0.15). It can be seen in Figure 1 that the overall CEFR result generated by the GSE text analyzer is very consistent although the intended level of B1 may not match. Since the GSE text analyzer generated the valentine's day text to be at B2, another text was used to compare (The Grand Canyon) which mapped to be at the same B2 level by the GSE Text Analyzer (the CEFR level of The Grand Canyon is different when compared by other text analyzer programs). Alignment of scores of the GSE scale to the CEFR is shown in Table 5. The GSE scale for valentine's day is at 64-68 whereas The Grand Canyon is at 62-66. The GSE range for B12 is at 59-75, therefore, the result generated by the GSE text analyzer is not surprising. As intended by the Pearson research team, GSE could provide a fine-grain granular descriptive system to define graded learning, teaching, and assessment objectives. This implies that while CEFR can indicate the difficulty level as an overview, other scales could be used as an additional measurement for a more accurate representation of text difficulty. Figure 2 shows that the GSE text analyzer can accurately map the level of all texts to the B2 level as the text material producers intended to produce. More research needs to be done on the leveling of text difficulty in the GSE text analyzer and CEFR, especially on the alignment of specific point ranges and CEFR levels.

Table 5. Alignment of scores to the CEFR

CEFR	GSE
Below A1	10-21
A1	22-29
A2	30-35
A2+	36-42
B1	43-50
B1+	51-58
B2	59-66
B2+	67-75
C1	76-84
C2	85-90

The study findings could be a starting point to investigate further the coherence of automated instruments when analyzing CEFR Texts. Users may also delve into linguistic features of text, not only the traditional linguistic features like vocabulary and grammar points used but also consider the role of concepts embedded in a text. For instance, semantic association as one of the linguistic features of text may be a factor in influencing text difficulty level. Although a text identified to be at the same level of B2 may be challenging to read due to differences in content (e.g., history, narrative, and science fiction), the pattern of abstract concepts embedded in the content presented in B2 reading materials is often overlooked. If this is the case, additional qualitative metrics may need to be proposed and included in addition to the traditional quantitative measures and AI tools available today [29]–[33].

5. CONCLUSION

This study highlights the significant inconsistencies among various automated CEFR text classification tools, including ChatGPT 3.5, Cathoven AI, GSE Text Analyzer, Text Inspector, and Oxford Text Checker. Despite advancements in NLP and machine learning, these tools demonstrate varied results for the same texts, underscoring the complexity and challenges of automated text classification in educational settings. Traditional readability metrics and lexical profiling frameworks often provide consistent but higher-level classifications compared to advanced NLP-based tools. The broad range of CEFR levels assigned for the same texts raises concerns about reliability and standardization, which are critical for educators, material developers, and language assessment professionals. The content and genre of texts significantly influence CEFR classification, indicating a need for further research into qualitative aspects of text classification, including semantic associations and thematic content. To enhance the reliability and applicability of these tools, future research should expand sample sizes, include diverse texts, and integrate qualitative metrics with traditional quantitative measures. Continuous advancements and validation efforts are essential to improve the accuracy and reliability of automated CEFR text classification tools for effective language teaching and assessment. As communicative language teaching becomes more prominent across the globe, the CEFR guideline provides what language learners need to be able to do with language at different levels, not how many complex sentences they can produce. Therefore, the use of language differs since cultural backgrounds and genres are always involved and could be reflected through written text.

As for the broader community, one significant finding that the study provides is that the field of computer science and algorithmic developers will need to fine-tune or calibrate algorithms in different automated tools with a better understanding of what is involved in the classification of texts based on the CEFR guideline. Factors relating to cultural norms of different communities and genres will need to be taken into account. Future tool development should account for the influence of content and genre on CEFR alignment to improve accuracy. Researchers and practitioners should also integrate both quantitative and qualitative methods to provide a more comprehensive assessment of text difficulty. Addressing these issues will help create more reliable tools that better support language education and research. Lastly, one of the major implications of this study is that it could initially be used to confirm the need for a more qualitative approach when doing text difficulty classification. Policymakers in language education may need to be aware of the consistency among today’s automated CEFR analyzers. Before making a decision relating to text difficulty for target learners, policymakers may have to ensure that the generated results from automated tools are consistent with the qualitative CEFR text identification approach.

FUNDING INFORMATION

The authors declare no funding for this research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Piyathat Siripol	✓	✓		✓	✓	✓		✓	✓	✓	✓		✓	
Seongha Rhee		✓			✓	✓			✓	✓		✓		
Suthathip Thirakunkovit		✓				✓				✓		✓		
Aphiwit Liang-Itsara	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author [AL-I], upon reasonable request.

REFERENCES

[1] R. Gopal, M. Maniam, N. A. Madzlan, S. S. B. Shukor, and K. Neelamegam, “Readability formulas: An analysis into reading index of prose forms,” *Studies in English Language and Education*, vol. 8, no. 3, pp. 972–985, 2021, doi: 10.24815/siele.v8i3.20373.

[2] T. Tanprasert and D. Kauchak, “Flesch-Kincaid is Not a Text Simplification Evaluation Metric,” in *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021*, pp. 1–14, 2021, doi: 10.18653/v1/2021.gem-1.1.

[3] E. Matricciani, “Readability Indices Do Not Say It All on a Text Readability,” *Analytics*, vol. 2, no. 2, pp. 296–314, 2023, doi: 10.3390/analytics2020016.

[4] S. A. Crossley, S. Skalicky, and M. Dascalu, “Moving beyond classic readability formulas: new methods and new models,” *Journal of Research in Reading*, vol. 42, no. 3–4, pp. 541–561, 2019, doi: 10.1111/1467-9817.12283.

[5] Y. T. Sung, W. C. Lin, S. B. Dyson, K. E. Chang, and Y. C. Chen, “Leveling L2 Texts Through Readability: Combining Multilevel Linguistic Features with the CEFR,” *Modern Language Journal*, vol. 99, no. 2, pp. 371–391, 2015, doi: 10.1111/modl.12213.

[6] B. Deygers, C. H. Carlsen, N. Saville, and K. van Gorp, “The Use of the CEFR in Higher Education: A Brief Introduction to This Special Issue,” *Language Assessment Quarterly*, vol. 15, no. 1, pp. 1–2, 2018, doi: 10.1080/15434303.2017.1421957.

[7] T. Phaisannan, K. Charttrakul, and A. Damnet, “The CEFR-TBL in Fostering Thai Pre-service Teachers’ English Speaking Ability Using the Peer Interview Task,” *Advances in Language and Literary Studies*, vol. 10, no. 5, p. 10, 2019, doi: 10.7575/aiac.alls.v.10n.5p.10.

[8] A. Khan, A. R. David, A. H. Ahmad, A. Ali, and S. C. Lah, “Initial Insights into CEFR Adoption at a Language Faculty of a Public University in Malaysia,” *PASSA Journal*, vol. 67, pp. 330–360, 2023, doi: 10.58837/chula.pasaa.67.1.11.

[9] B. Waluyo, A. Zahabi, and L. Ruangsung, “Language Assessment at a Thai University: A CEFR-Based Test of English Proficiency Development,” *rEFLections*, vol. 31, no. 1, pp. 25–47, 2024, doi: 10.61508/refl.v31i1.270418.

[10] S. Tangsakul and K. Poonpon, “Aligning Academic Reading Tests to the Common European Framework of Reference for Languages (CEFR),” *rEFLections*, vol. 31, no. 2, pp. 614–638, 2024, doi: 10.61508/refl.v31i2.275057.

[11] A. H. M. Khair and P. M. M. Shah, “ESL Teachers’ Perceptions on the Implementation of CEFR in Malaysian Primary Schools: Issues and Challenges,” *Journal of Advances in Education Research*, vol. 6, no. 1, 2021, doi: 10.22606/jaer.2021.61005.

[12] H. Ma, J. Wang, and L. He, “Linguistic Features Distinguishing Students’ Writing Ability Aligned with CEFR Levels,” *Applied Linguistics*, vol. 45, no. 4, pp. 637–657, 2024, doi: 10.1093/applin/amad054.

[13] N. S. Lagutina, K. V. Lagutina, A. M. Brederman, and N. N. Kasatkina, “Text classification by CEFR levels using machine learning methods and BERT language model,” *Modeling and Analysis of Information Systems*, vol. 30, no. 3, pp. 202–213, 2023, doi: 10.18255/1818-1015-2023-3-202-213.

[14] A. Escobar-Acevedo, J. Guerrero-García, and R. Guzmán-Cabrera, “A Model Text Recommendation System for Engaging English Language Learners: Facilitating Selections on CEFR,” *Rupkatha Journal on Interdisciplinary Studies in Humanities*, vol. 14, no. 3, 2022, doi: 10.21659/rupkatha.v14n3.17.

[15] I. Natova, “Estimating CEFR reading comprehension text complexity,” *Language Learning Journal*, vol. 49, no. 6, pp. 699–710, 2021, doi: 10.1080/09571736.2019.1665088.

[16] V. J. Schmalz and A. Brutti, “Automatic assessment of English CEFR levels using BERT embeddings,” *CEUR Workshop Proceedings*, vol. 3033, 2021, doi: 10.4000/books.aaccademia.10828.

[17] E. Limbong and B. Wadham, “Exploring Pre-Service English Teachers’ Digital Competence in Creating Interactive Instructional Materials,” *Studies in English Language and Education*, vol. 11, no. 2, pp. 748–767, 2024, doi: 10.24815/siele.v11i2.35103.

[18] S. Mohamed, “Text classification and gradation in Arabic textbooks,” *The Language Learning Journal*, vol. 52, no. 6, pp. 629–649, Nov. 2024, doi: 10.1080/09571736.2023.2213695.

[19] A. Mekuria, E. W. Bushisho, and H. Wubshet, “The effects of reading strategy training on students’ reading strategy use and critical reading ability in EFL reading classes,” *Cogent Education*, vol. 11, no. 1, 2024, doi: 10.1080/2331186X.2024.2310444.

[20] S. Kim and S. H. Park, “Young Korean EFL Learners’ Perception of Role-Playing Scripts: ChatGPT vs. Textbooks,” *Korean Journal of English Language and Linguistics*, vol. 23, pp. 1136–1153, 2023, doi: 10.15738/kjell.23.202312.1136.

[21] N. I. Idris, N. L. Ali, and F. Jaafar, “Developing Wordlist for Primary Science Textbooks in Dual Language Programme,” *International Journal of Academic Research in Business and Social Sciences*, vol. 14, no. 1, 2024, doi: 10.6007/ijarbss/v14-i1/19937.

[22] M. Á. Z. Tejada, C. N. Gallardo, M. C. M. Ferradá, and M. I. C. López, “2L English Texts and Cohesion in upper CEFR Levels: A Corpus-Based Approach,” *Procedia - Social and Behavioral Sciences*, vol. 212, pp. 192–197, 2015, doi: 10.1016/j.sbspro.2015.11.319.

[23] C. Harsch and J. Hartig, “What Are We Aligning Tests to When We Report Test Alignment to the CEFR?” *Language Assessment Quarterly*, vol. 12, no. 4, pp. 333–362, 2015, doi: 10.1080/15434303.2015.1092545.

[24] L. A. Ficayuma, “Genre Analysis and Textual Boundaries to Cultivate Literacy and to Formulate Language English Policy: A Corpus Study of Pisa, TOEFL, English Book and Curriculum,” in *International Conference on English Language Teaching (4th ICONELT 2024),* 2014, pp. 1–15, doi: https://doi.org/10.15642/iconelt.2024.4.1-15.

[25] N. Yazici, İ. A. Gürsoy, and G. D. D. Aydemir, “Aspects of Narrative Competence in Language Levels and Narrative Texts in the CEFR1,” *Kastamonu Eğitim Dergisi*, vol. 29, no. 3, pp. 559–573, Sep. 2021, doi: 10.24106/kefdergi.774173.

[26] L. F. Huang, Y. L. Lin, and T. Gráf, “Development of the use of discourse markers across different fluency levels of CEFR A learner corpus analysis,” *Pragmatics*, vol. 33, no. 1, pp. 49–77, 2023, doi: 10.1075/prag.21016.hua.

[27] L. C. Chen, K. H. Chang, S. C. Yang, and S. C. Chen, “A Corpus-Based Word Classification Method for Detecting Difficulty Level of English Proficiency Tests,” *Applied Sciences (Switzerland),* vol. 13, no. 3, 2023, doi: 10.3390/app13031699.

[28] X. Hu, “Predicting CEFR Levels in L2 Oral Speech, Based on Lexical and Syntactic Complexity,” *Asia Pacific Journal of Corpus Research,* vol. 2, no. 1, pp. 35–45, 2021.

[29] J. Kim, S. Leijnen, and L. Beinborn, “Considering Human Interaction and Variability in Automatic Text Simplification,” in *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, 2024, pp. 52–60, doi: 10.18653/v1/2024.tsar-1.6.

[30] T. Gurmani, I. A. Abbasi, and Z. A. Jatoi, “Readability and Syntactic Complexity in Undergraduate Essays on Technology,” *Journal of Arts and Linguistics Studies*, vol. 3, no. 1, pp. 219–242, Feb. 2025, doi: 10.71281/jals.v3i1.219.

[31] A. C. Uyar and D. Büyükahıska, “Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT,” *International Journal of Assessment Tools in Education*, vol. 12, no. 1, pp. 20–32, Feb. 2025, doi: 10.21449/ijate.1517994.

[32] K. Kongsuwannakul, “Making a case for a change to using CEFR-oriented placement test scores: A reflexive ethnographic decision making,” *Heliyon*, vol. 6, no. 1, 2020, doi: 10.1016/j.heliyon.2020.e03324.

[33] K. Savski, “CEFR as Language Policy: Opportunities and Challenges for Local Agency in a Global Era,” *The English Teacher*, vol. 50, no. 2, pp. 60–70, 2021, doi: 10.52696/aide2513.

BIOGRAPHIES OF AUTHORS

Piyathat Siripol is a Ph.D. candidate in Applied Linguistics, Mahidol University, Thailand. He earned a BA in English from Khon Kaen University and an MA in Applied Linguistics for English Language Teaching from King Mongkut's University of Technology Thonburi. He is currently a lecturer in the international undergraduate program in TESOL at the Faculty of Education, Siam Technology College, Thailand. His research interests include second language writing assessment, semantics, and discourse analysis. He can be contacted at email: piyathat.sir@student.mahidol.ac.th.

Seongha Rhee is a professor of Linguistics at Faculty of Liberal Arts, Mahidol University, Thailand and Professor Emeritus of Hankuk University of Foreign Studies, Korea. He received his Ph.D. in linguistics from the University of Texas, Austin in 1996. He published World Lexicon of Grammaticalization (2019, CUP, co-author); book chapters in Grammaticalization Scenarios (2020, Mouton), The Cambridge Handbook of Korean Linguistics (2021, CUP), The Oxford Handbook of Grammaticalization (2011, OUP), and research articles in Nature (co-author), Journal of Pragmatics, Language Sciences, and Lingua, among others. His primary research interest is to identify cognitive mechanisms of grammaticalization from crosslinguistic, typological perspectives. He can be contacted at email: srhee@hufs.ac.kr.

Suthathip Thirakunkovit is an assistant professor in the International Graduate Program in Applied Linguistics at the Faculty of Liberal Arts, Mahidol University, Thailand. Her research interests include language test development, test validation, corpus linguistics, and second language writing. She can be contacted at email: suthathip.thi@mahidol.edu.

Aphiwit Liang-Itsara is an assistant professor in the International Graduate Program in Applied Linguistics, Faculty of Liberal Arts, Mahidol University. He earned his Ph.D. in Instruction and Learning, with specializations in English Education and Applied Linguistics, from the University of Pittsburgh, USA. His research interests include Teaching English as a foreign language (TEFL), the psychology of language teaching and learning, and functional linguistics. He can be contacted at email: aphiwit.lia@mahidol.edu.