Linguistic Complexity Analysis of Readability-Controlled Simplified Texts

Su-Youn Yoon[1], Kexin Bian[2], Mamoru Komachi[2], Masanori Hayashi[1]

[1]EduLab, Inc. [2]Hitotsubashi University

{su-youn.yoon, masanori.hayashi}@edulab-inc.com

{dm240020@g, komachi@r}.hit-u.ac.jp

Abstract

To develop automated systems that generate reading materials for language learners by simplifying existing English texts, it is essential to achieve semantic consistency with the source text while ensuring the generated output maintains a linguistic complexity appropriate for the learner's proficiency level. To address this, many researchers utilize similarity score with expert-simplified texts during evaluation. In this study, we investigated the linguistic complexity of the texts used as the "gold standard" in a common evaluation dataset for such systems. Our findings indicate that these gold-standard texts were more complex than those originally written for students of a similar level, highlighting the importance of developing metrics to carefully monitor linguistic complexity across various linguistic dimensions.

1 Introduction

Creating reading passages that maintain appropriate difficulty levels while incorporating authentic, real-world content remains a challenging task, even for experienced language teachers and item writers. The advancement of large language models (LLMs) has transformed various fields, including education, and their ability to generate human-like, natural text has prompted the development of automated reading material generation for both native and non-native learners. One research direction is readability-controlled text simplification (or proficiency-controlled text generation), which involves selecting existing English texts and transforming them into levels suitable for students. A recent example of this trend is the shared task on Text Simplification, Accessibility, and Readability at the 2025 EMNLP workshop (hereafter, TSAR workshop).

To assess the quality of system-generated texts, a valid and accurate evaluation process is essential. Research in this area relies heavily on automated metrics. Specifically, difficulty is often measured using automated CEFR (Common European Framework of Reference for Languages) level predictors, while content appropriateness is evaluated based on similarity to manually simplified reference texts or the original source texts. While these methods are valuable and cost-effective, they have inherent limitations. Most notably, evaluation results depend heavily on human-generated reference texts, and detailed linguistic characteristics are often overlooked because the CEFR score provides only a single, aggregate metric.

Simplified texts are derived from original documents and adapted to a target level while preserving the core content. In this process, lexical and grammatical expressions exceeding the target level are replaced with simpler alternatives. While successful simplification should ideally result in vocabulary and grammar within the target threshold, their distribution may differ from texts originally authored for that specific CEFR level. Consequently, these simplified texts may remain more challenging for non-native learners to comprehend.

This study explores the linguistic complexity of expert-simplified texts from TSAR workshop. By comparing them with reading passage originally written for the same CEFR level, we will investigate whether their linguistic profiles are comparable or not.

2 Relevant studies

Automated text simplification (TS) systems aim to improve the text readability by reducing the linguistic complexity while preserving the original content [1]. Due to the efficiency of the system evaluation, researchers in this field utilized automated metrics based on the expert-simplified texts. One such metric is D-SARI, which assesses how appropriately a system manages deletions, additions, and retentions of of the original content relative to human-authored references[2].

However, recent studies such as [3] raised a concern about using generic evaluation metrics. He pointed out that types of the reading difficulties vary by the target audience and it is important to develop TS systems to address audiences' own difficulties and goals. To evaluate such diverse systems, the evaluation metrics need to be adapted to assess success or failure of these specific goals.

For TS systems for L2 learners, one of the important metric is its adherence to specific linguistic levels, such as CEFR. Furthermore, for the pedagogical purpose, the goal of TS is not merely to reduce the linguistic complexity and make a text "easy," but to ensure it is pedagogically appropriate. Thus, high-quality simplification systems must balance the semantic distance from the source text with the specific goal of providing the "right level" of linguistic challenge—neither too overwhelming nor too simple.

To address this goal, [4] proposed three evaluation metrics: linguistic fluency, semantic consistency, and CEFR alignment. While semantic consistency was evaluated by the content match with the source text, the CEFR level was evaluated based on the predicted labels via regression models of linguistic features. Similarly, the recent TSAR workshop proposed CEFR alignment, semantic similarity to source texts, and similarity to expert-simplified texts as evaluation metrics.

This study evaluates the TSAR workshop references through this lens of goal-oriented simplification, examining how well these texts serve as benchmarks for readability-controlled tasks.

3 Data

3.1 Simplified texts from TSAR workshop

We used expert-simplified texts from TSAR workshop which was used as a gold standard for the evaluation of automated simplification systems. All source materials were obtained from the British Council's LearnEnglish website, having been originally authored by professional educators and categorized by CEFR levels. One hundred texts from the B2 and C1 levels were selected and simplified to B1 and A2 target levels by experienced EFL teachers. Before simplifying texts, they underwent a training session to familiarize themselves with the guidelines, ensuring that the simplified versions reduced lexical and syntactic complexity while preserving the original content. The detailed information is provided in [5].

Table 1 Descriptive analysis of word counts across grade levels

	mean	std	min	max
cambridge-A2	126.6	36.6	76	198
cambridge-B1	236.5	93.8	134	428
TSAR-A2	72.2	17.9	37	117
TSAR-B1	80.6	19.2	42	122

3.2 Cambridge reading passages

The original texts are from the Cambridge English Readability Dataset [6], which comprises reading passages extracted from five Cambridge English Examinations designed for L2 learners. We selected a total of 100 samples from this dataset, with 50 samples each for the A2 and B1 levels.

The descriptive analysis of text length for both datasets is summarized in Table 1.

The length of the Cambridge reading passage dataset was generally longer than the simplified texts. It was approximately 1.8 times longer for A2 and 2.9 times longer for B1, respectively.

4 Experiment

4.1 linguistic features

For the linguistic analysis, we used CEFR-based Vocabulary Level Analyzer, Version 3 [7], designed to estimate the CEFR-J level of reading texts. CVLA3 provides 8 linguistic features designed to assess the lexical and syntactic complexities and text readabilities. In this study, we used the following five representative features[1]:

- **AvrDiff**: The mean CEFR level of content words, calculated by mapping levels A1 through C2 to a numerical scale of 1 to 6.
- **BperA**: Ratio of Level B words relative to Level A words.

[1] The detailed description of features are provided in [8] and [7]

Table 2 Means and standard deviations (in parenthesis) of linguistic features for CEFR A2 Level texts

Metric	TSAR	Cambridge
AvrDiff	1.40 (0.17)	1.24 (0.11)
BperA	0.12 (0.09)	0.04 (0.04)
LenNP	2.75 (0.89)	2.73 (0.74)
VperSent	2.46 (0.62)	2.21 (0.50)
ARI	5.99 (1.88)	4.85 (2.27)

- **LenNP**: The average length of noun phrases within the text.
- **VperSent**: The average number of verbs per sentence.
- **ARI (Automated Readability Index)**: A readability score calculated using the formula established by Senter and Smith (1967) [9].

AvrDiff and BperA are designed to assess lexical complexity. In particular, AvrDiff and BperA focus on the distribution of content words. LenNP and VperSent are intended to evaluate grammatical complexity. LenNP measures the average length of noun phrases based on the assumption that longer noun phrases increase overall sentence difficulty. Finally, the ARI is designed to assess holistic text difficulty and is sensitive to both sentence and word length. For all measures, a higher value implies a greater level of lexical, syntactic, holistic text complexity.

4.2 Automated CEFR Level Labeling

We assigned a CEFR label to each text using an automated CEFR level prediction system. This system, provided as part of the TSAR-2025 shared task evaluation metrics, utilizes an ensemble of three fine-tuned ModernBERT models. We downloaded the model from the official repository (https://github.com/tsar-workshop/tsar-2025-shared-task), and specific implementation details are provided in [5].

5 Results

5.1 Linguistic Complexity Analysis

First, we conducted a descriptive analysis of linguistic features. Table 2 compares the TSAR texts and Cambridge reading passages at the A2 level.

On average, the TSAR texts exhibited higher mean values across all features compared to the Cambridge passages. Based on the t-test results, this trend was statistically significant for all features except VperSent. Higher values indicate greater textual complexity, suggesting that the simplified texts exhibit more complex linguistic characteristics than the Cambridge passages authored for the same CEFR level.

Table 3 Descriptive analysis of linguistic features for CEFR B1 Level texts

Metric	TSAR	Cambridge
AvrDiff	1.60 (0.20)	1.54 (0.13)
BperA	0.20 (0.12)	0.15 (0.07)
LenNP	3.48 (1.67)	3.99 (1.68)
VperSent	3.07 (0.77)	3.02 (0.65)
ARI	8.90 (1.90)	8.48 (1.97)

Table 3 presents the comparison between simplified texts and Cambridge passages at the B1 level. A similar trend was observed; however, statistically significant differences were limited to the lexical complexity features: AvrDiff and BperA. As with the A2 level, the simplified texts had higher means, implying that they are lexically more complex than the Cambridge passages.

In Section 3, we reported a substantial difference in length between the two text types, with the TSAR texts being significantly shorter than the Cambridge passages. Since differences in text length can impact linguistic features—specifically by lowering values in the longer Cambridge passages due to word repetition—we performed a correlation test between the linguistic features and the word count of each text. For the lexical complexity features (AvrDiff and BperA), the correlations were not statistically significant. For the remaining features (LenNLP, VperSent, and ARI), significant correlations were observed, but the relationships were weak; the highest correlation coefficient was under 0.2, indicating a low risk of length-based bias. These tests support the conclusion that the differences in linguistic complexity are driven by the text type rather than text length, resulting in simplified texts that are inherently more complex.

5.2 Automated Label Prediction Analysis

Finally, we analyzed the labels generated by the automated CEFR level prediction system. Since the workshop organizers did not provide performance metrics, the ac-

Table 4 CEFR Target Distribution

	Predicted CEFR		
target CEFR	A2	B1	B2
A2	64	36	0
B1	26	61	13

curacy of these automated labels remains unknown.[2)] We proceeded under the assumption that if the TSAR dataset is comparable to the Cambridge dataset, the level of agreement between their target and predicted CEFR levels should also be comparable.

Large differences were observed between the two datasets. For the Cambridge passages, the agreement between the target levels and automated labels was 100%; specifically, all A2 passages were labeled as A2, and all B1 passages were labeled as B1. In contrast, the agreement for the TSAR texts was 62.5%. The perfect agreement for the Cambridge passages should be interpreted with caution. Because these passages have been publicly available since 2016, they may have been included in the system's training data, leading to inflated accuracy. While we intend to verify this with the TSAR workshop organizers, it is noteworthy that the agreement for the TSAR texts is substantially lower.

For a more detailed analysis, Table 4 provides a cross-tabulation of the target CEFR levels and automated CEFR labels for the TSAR texts. The majority of the TSAR texts were labeled with their corresponding target CEFR levels: 64% for target level A2 and 61% for target level B1. In addition, all prediction errors fell into adjacent CEFR levels.

When the target level was A2, there was a tendency for the system to predict a higher level (36%), suggesting that the human-authored texts may be more complex than their intended level. This finding is consistent with our linguistic feature analysis. However, for target level B1, the system exhibited trends of both "undershooting" (predicting A2) and "overshooting" (predicting B2), and the undershooting rate was twice that of overshooting. Such conflicting classification trends increase within-group variance, explaining why few linguistic features show statistically significant differences and highlighting the need for further investigation of the B1 level.

6 Conclusion

The linguistic analysis of TSAR texts demonstrates that simplified versions often exhibit greater complexity than original texts written for the same CEFR level. Despite the involvement of trained experts, automated labeling reveals frequent "overshooting" and "undershooting," suggesting that aligning texts with specific CEFR standards is a challenge even for professionals.

This raises concerns regarding the practicality of creating such reference texts and the subsequent impact on system evaluations. Although the evaluation procedure incorporates automated CEFR labels to ensure that the created texts meet the appropriate linguistic complexity, this remains a single summary metric. As such, it may be insufficient for ensuring quality across various linguistic dimensions such as grammar, vocabulary, coherence, and topics.

Recent studies investigating LLM-based simplification have reported various challenges regarding CEFR-level adjustment. [11] noted that ChatGPT-generated simplifications lack lexical sophistication, while [12] found that texts targeting lower CEFR levels were too simplistic, while texts targeting higher levels were overly complex compared to textbook standards.

While both studies utilized relatively simple prompting approaches—consisting of a single instruction with or without few-shot examples—their findings emphasize the inherent difficulty of CEFR level adjustment. This highlights the importance of developing evaluation metrics that carefully monitor linguistic complexity to ensure texts are appropriate yet sufficiently challenging for language learners. Traditionally, numerous valid linguistic measures for assessing grammar, vocabulary, and topics have been available and open-sourced. Actively incorporating these measures into text simplification for the generation of educational materials represents a valuable approach for future research.

[2)] For reference, we report the performance of CEFR level prediction systems developed for English texts. For instance, UniversalCEFR—a project focused on creating a large-scale, multilingual, and multidimensional dataset of CEFR-annotated texts—developed automated prediction systems. Their best-performing model, based on ModernBERT, achieved a 76% weighted F_1 score on English texts; a detailed description is provided in [10]. Additionally, [4] developed an automated CEFR level prediction model based on regression and linguistic features, reporting an R^2 of 0.8.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number 25K03178.

References

[1] Suha S Al-Thanyyan and Aqil M Azmi. Automated text simplification: a survey. **ACM Computing Surveys (CSUR)**, Vol. 54, No. 2, pp. 1–36, 2021.

[2] Renliang Sun, Hanqi Jin, and Xiaojun Wan. Document-level text simplification: Dataset, criteria and baseline. **arXiv preprint arXiv:2110.05071**, 2021.

[3] Sian Gooding. On the ethical considerations of text simplification. **arXiv preprint arXiv:2204.09565**, 2022.

[4] Ali Malik, Stephen Mayhew, Christopher Piech, and Klinton Bicknell. From Tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation. In **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 15670–15693, 2024.

[5] Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In **Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)**, pp. 116–130, 2025.

[6] Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. Text readability assessment for second language learners. In **Proceedings of the 11th workshop on innovative use of NLP for building educational applications**, pp. 12–22, 2016.

[7] Satoru Uchida and Masashi Negishi. Estimating the CEFR-J level of English reading passages: Development and accuracy of CVLA3. 英語コーパス研究, Vol. 32, pp. 165–174, 2025.

[8] Satoru Uchida and Masashi Negishi. Assigning CEFR-J levels to English texts based on textual features. In **Proceedings of Asia Pacific Corpus Linguistics Conference**, Vol. 4, pp. 463–467, 2018.

[9] Edgar A Smith and RJ Senter. **Automated readability index**, Vol. 66. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, 1967.

[10] Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R Jablonkai, et al. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. In **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 9714–9766, 2025.

[11] Fengkai Liu, Xiaofei Lu, Tan Jin, Mengchao Kang, and Haomin Zhang. Does ChatGPT simplify texts like expert teachers? linguistic features of simplified texts. **Reading and Writing**, pp. 1–21, 2025.

[12] Satoru Uchida. Generative AI and CEFR levels: Evaluating the accuracy of text generation with ChatGPT-4o through textual features. **Vocabulary Learning and Instruction**, Vol. 14, No. 1, pp. 2078–2078, 2025.