Findings of the TSAR 2025 Shared Task on Readability-Controlled Text Simplification

Fernando Alva-Manchego[1], **Regina Stodden**[2], **Joseph Marvin Imperial**[3,4], **Abdullah Barayan**[1,5], **Kai North**[6], **Harish Tayyar Madabushi**[3]
[1]Cardiff University, [2]Bielefeld University, [3]University of Bath,
[4]National University Philippines, [5]King Abdulaziz University, [6]Cambium Assessment
CORRESPONDENCE: alvamanchegof@cardiff.ac.uk

Abstract

This paper presents the findings of the first Shared Task on Readability-Controlled Text Simplification at TSAR 2025. The task required systems to simplify English texts to specific target readability levels of the Common European Framework of Reference for Languages (CEFR). We received 48 submissions from 20 participating teams, with approaches predominantly based on large language models (LLMs), which included iterative refinement, multi-agent setups, and LLM-as-a-judge pipelines. For this shared task, we developed a new dataset of pedagogical texts and evaluated submissions using a weighted combination of semantic similarity and CEFR-level accuracy. The results of the participating teams demonstrate that while LLMs can perform substantially well on this task, dependable and controlled simplification often requires complex, multi-iterative processes. Our findings also suggest that the capabilities of current systems are beginning to saturate existing automatic evaluation metrics, underscoring the need for re-evaluation and practicality.

1 Introduction

Text simplification consists of automatically rewriting a text to make it easier to read and understand while preserving meaning, supporting applications in education, accessibility, and second-language learning (Alva-Manchego et al., 2020). Many previous shared tasks have focused on lexical simplification (Specia et al., 2012; Saggion et al., 2022; Shardlow et al., 2024) or on complexity prediction (Paetzold and Specia, 2016; Yimam et al., 2018; Shardlow et al., 2021). However, these tasks typically do not require control over output readability level, which is essential if simplification is to be adaptive to learner needs.

In this work, we introduce a new shared task for readability-controlled text simplification in English, in which systems must simplify a source text to a specified CEFR level (e.g., A2 or B1). In this way, the target complexity explicitly aligns with educational and pedagogical goals. Our task builds on, but also departs from, earlier shared tasks in lexical simplification, in that participating systems must simplify short passages under a CEFR constraint, rather than single words or phrases based on a target audience.

To support this shared task, we curated a **new CEFR-based reference dataset of 100 paragraph-level English texts** drawn from pedagogical reading materials for language learners. Each source text was manually simplified by experienced English-language teachers to two lower target levels, resulting in a total of 200 reference simplifications. In addition, we trained a **CEFR evaluator model** to estimate the readability level of system outputs automatically. The model was trained on CEFR-annotated texts and fine-tuned to classify English texts into CEFR levels with high reliability. Both resources are publicly released to support further research.

The shared task challenged participants to generate simplified versions of the same source texts at specified CEFR targets, requiring systems to demonstrate both readability control and semantic fidelity. Submissions were assessed using the CEFR evaluator model to measure compliance with the target level and MeaningBERT (Beauchemin et al., 2023) to assess source and reference-wise meaning preservation. The resulting metric scores were combined into a single ranking using AUTORANK (Kocmi et al., 2025), an aggregation method that normalizes metric scales and mitigates the effect of outliers.

2 Related Work

The task of readability-controlled text simplification (RCTS) aims to generate simplified text aligned with specific difficulty levels, often us-

Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 116–130
November 4-9, 2025

ing frameworks such as the Common Core Standards or the CEFR for reference (Xu et al., 2015; Uchida et al., 2018; Scarton et al., 2018). The main challenge with RCTS is that it requires finer-grained generation control across multiple versions of the exact text, tailored to different audiences. Early approaches treated RCTS as a supervised sequence-to-sequence problem requiring a parallel corpus annotated with readability (Scarton and Specia, 2018), which then followed by more advanced techniques including the addition of low-level control tokens (Nishihara et al., 2019; Agrawal and Carpuat, 2023), lexical-based constrained decoding for better handling of complex words (Zetsu et al., 2022), and reinforcement learning to reward readability-aligned texts (Yanamoto et al., 2022; Ribeiro et al., 2023; Malik et al., 2024).

Despite the substantial progress, current methods remain heavily dependent on gold-standard parallel data. To address this, more recent works have explored techniques that take advantage of instruction-tuned LLMs' few-shot learning capabilities for shifting the readability levels of texts without the need for additional training data (Kew et al., 2023; Imperial and Tayyar Madabushi, 2023; Imperial et al., 2024; Farajidizaji et al., 2024; Malik et al., 2024; Barayan et al., 2025). However, achieving precise readability control while maintaining the quality of generated simplifications remains challenging, motivating further research into fine-grained level alignment.

3 CEFR Evaluator Model

We detail the steps we followed to train the CEFR-based evaluator model we will use to evaluate the CEFR alignment of system submissions with gold-standard reference simplifications.

3.1 Data

We used the English subset of the UNIVERSAL-CEFR (Imperial et al., 2025) dataset, which contains gold-standard CEFR-annotated texts at multiple granularities (sentence-, paragraph-, and document-level). We constructed three separate training sets with variations of granularities and language coverage:

TRAIN_DOC_EN This split contains 650 English documents from CAMBRIDGEEXAMS (Xia et al., 2016) and ELG-CEFR-EN (Breuker, 2022).[1]

TRAIN_DOC_SENT_EN This split contains 13,476 English sentence- and document-level texts, combining TRAIN_DOC_EN with additional data from CEFR-SP (Arase et al., 2022) and README++ (EN) (Naous et al., 2024).[2]

REFERENCE_ALLLANG This is the largest split with 56,963 multilingual instances at the sentence, paragraph, and document levels, integrating data from multiple languages together with TRAIN_DOC_EN.[3] See Table 7 in the Appendix for the list of datasets and languages.

For our validation and test sets, we use the corresponding test splits of the English document-level subsets CAMBRIDGEEXAMS and ELG-CEFR-EN. Stratified sampling was applied to maintain a proportional representation across CEFR levels, with 15% of the data allocated to validation and 15% to testing. The final distribution of instances across splits and CEFR levels is shown in Table 6 (Appendix).

3.2 Base Model Architecture

We fine-tuned MODERNBERT-BASE (Warner et al., 2024), a 395M-parameter LLM on each of the three training sets previously discussed. Using this training regime, we produce the following: (1) a document-level English evaluator model using the TRAIN_DOC_EN split, (2) a combination of sentence and document multi-level English evaluator model using the TRAIN_DOC_SENT_EN split, and (3) a multilingual sentence, paragraph, and document evaluator model using the REFERENCE_ALLLANG split. All evaluator model variations were trained for 10 epochs, with the best checkpoint selected based on the highest weighted F1 score on the validation set. Additional training details can be found in Table 4 of the Appendix.

In addition to training MODERNBERT-BASE, we explored two ensemble-based strategies for resolving the final CEFR prediction:

- **Majority Vote:** Labels were assigned based on agreement among at least two models. In cases without a majority, the median CEFR level was chosen.
- **Confidence-Based:** Predictions were taken from the model with the highest confidence score for the given instance.

[1] AbdullahBarayan/ModernBERT-base-doc_en-Cefr

[2] AbdullahBarayan/ModernBERT-base-doc_sent_en-Cefr

[3] AbdullahBarayan/ModernBERT-base-reference_AllLang2-Cefr2

Model Setup	A1	A2	B1	B2	C1	C2	Avg	AdjAcc	RMSE
TRAIN_DOC_EN	0.80	0.90	0.84	0.84	0.74	0.83	0.83	0.97	0.50
TRAIN_DOC_SENT_EN	0.00	0.83	0.90	0.94	0.85	0.86	0.86	0.99	0.38
REFERENCE_ALLLANG	0.50	0.86	0.89	0.97	0.88	0.88	0.89	1.00	0.32
MAJORITY VOTE	0.50	0.87	0.92	0.95	0.85	0.86	0.89	0.99	0.35
CONFIDENCE-BASED	0.00	0.89	0.94	0.94	0.87	0.89	0.89	0.99	0.34

Table 1: Performance of various training data and model prediction setups integrated with ModernBERT-Base on the **test set**. We selected the CONFIDENCE-BASED setup for our final CEFR evaluator model due to its optimal performance (low RMSE and high averages).

3.3 Results

Table 1 reports the performance of the three fine-tuned models and the two ensemble strategies on the test set (Table 8 in the Appendix reports results in the validation set). Results are presented in terms of class-wise F1 scores, weighted average F1, adjusted accuracy, and RMSE. The results demonstrate that all three fine-tuned models achieve strong performance, with the confidence-based ensemble providing the most consistent accuracy and lowest error across both validation and test sets. We therefore adopt this ensemble to assess CEFR compliance of simplified outputs.

4 Shared Task Dataset

Our primary shared task dataset was designed to support readability-controlled simplification in English aligned with the Common European Framework of Reference for Languages (CEFR). The dataset is also aimed towards evaluating systems that can simplify texts to a target readability level while preserving meaning.

4.1 Data Source

All texts were extracted from The British Council's LearnEnglish website[4], a major UK-based open-access platform offering pedagogical content for learners of English. We were granted formal permission to use the materials for research and distribution as part of this shared task. The acquired material includes graded reading passages, each associated with a specific CEFR level. The content covers a range of everyday topics and was authored and reviewed by professional educators.

From the available materials, we selected a subset of texts originally labeled as C1 and B2 to serve as source texts for simplification. These upper-intermediate and advanced texts provide sufficient lexical and syntactic complexity to enable meaningful simplification toward lower CEFR levels.

[4] https://learnenglish.britishcouncil.org/

4.2 Data Annotation

The annotation process involved producing simplified versions of the original C1 and B2 source texts at lower CEFR levels. Each selected text was simplified to B1 and A2 target levels, resulting in two simplified versions per source. The dataset was divided into two parts: (1) **trial data** (20 instances), simplified by one annotator and released for system development; and (2) **test data** (80 instances), simplified by two annotators and used for official evaluation. All annotators were teachers of English as a foreign language and familiar with the CEFR framework, although with varied experience levels.

Before the main annotation phase, all annotators completed a qualification task to ensure a consistent understanding of CEFR levels and simplification principles. During annotation, each annotator received: (1) the original C1 or B2 text; (2) the target CEFR level (A2 or B1); and (3) a set of annotation guidelines describing the expected linguistic characteristics of the simplified output. These guidelines specified that simplifications should preserve meaning, reduce syntactic and lexical complexity in line with the target level, and maintain grammaticality and fluency. The full qualification task and annotation guidelines are included in the dataset release.

4.3 Quality Control

All simplifications were reviewed by the organisers for formatting consistency and completeness. No additional post-editing or filtering was applied to preserve the natural stylistic variation introduced by each annotator.

Inter-annotator agreement was not computed, as the annotators worked on disjoint subsets of the data. Instead, we assessed reliability by comparing the target CEFR levels assigned during annotation with the levels predicted by our automatic CEFR evaluator model (Sec. 3). The average RMSE was 0.6, suggesting moderate agreement between hu-

man and model estimates. While lower values would indicate stronger alignment, this level of divergence is expected given the subjectivity of CEFR judgments and the coarse step size between adjacent levels. Detailed metrics per annotator are provided in the Appendix.

5 Evaluation Setup

We describe the evaluation pipeline, combining normalization procedures, weighting decisions, and ranking methodology, used to assess the performance of system submissions using the shared task data and a trained CEFR evaluator model. Following the AUTORANK framework proposed in the WMT 2025 General Machine Translation Shared Task (Kocmi et al., 2025), we aggregate multiple evaluation metrics into a single overall ranking to increase robustness against outliers and improve interpretability. We present two forms of AUTORANK rankings: one for all submitted runs and another for the best run per team.

5.1 Metrics

We evaluated each system submission based on three variables: CEFR level compliance, meaning preservation, and gold-standard reference similarity. We describe the metrics chosen to measure each variable below:

1. **CEFR Level Compliance**. We use the Root Mean Squared Error (RMSE) from the trained CEFR model evaluator to assess CEFR level compliance. Lower RMSE values indicate better control of a submitted system with respect to the target CEFR readability level.

2. **Meaning Similarity**. We use MeaningBERT (Beauchemin et al., 2023) to measure the semantic similarity between the source text and the submitted system's output.

3. **Reference Similarity**. Similar to Meaning Similarity, we also use MeaningBERT (Beauchemin et al., 2023) to measure the semantic similarity between the expert-written simplifications and a submitted system's output.

We considered other computed metrics, including adjacent accuracy, weighted F1, and BERTScore, but they were ultimately not used in the official ranking for several reasons. Adjacent accuracy is less informative than RMSE because it does not account for the degree of mismatch. On the other hand, RMSE penalizes predictions proportionally to their distance from the target level. Weighted F1 reflects categorical performance but does not capture the severity of misclassification. Lastly, while BERTScore (Zhang et al., 2020) is a popular general-purpose similarity metric, it was not trained for simplification and often overestimates similarity when there is lexical overlap without true semantic preservation. MeaningBERT, on the other hand, was trained on human annotations for preserving meaning during simplification and is a more task-appropriate choice.

5.2 Submission Filtering

We observed that some teams submitted model simplifications with fewer runs than the expected total of 200. Since missing outputs would bias the evaluation, all runs with fewer than 200 outputs were discarded before scoring and ranking.

5.3 Normalization of Metrics

Metrics operate on different scales and distributions. RMSE included very low outliers, including values close to 0.0, representing perfect or near-perfect CEFR compliance. MeaningBERT values were tightly clustered. If we combined these raw values directly, it is evident that RMSE would dominate because of its larger relative variance. To address this, following (Kocmi et al., 2025), we applied *median–interpercentile scaling* to each metric. This normalization method reduces the influence of outliers while making scores comparable across metrics. Unlike min–max scaling, which is highly sensitive to outliers, this approach ensures that mid-ranked systems remain fairly distinguished. For RMSE, since lower values are better, we invert the scaled scores so that for all metrics, higher is always better.

5.4 Weighting

To reflect the balance required to optimize readability control and meaning preservation, we assign equal global weights of 50% to these two variables. While both semantic similarities in the source text and reference simplifications are essential, the latter generally correlates more strongly with expert judgments of simplification quality. For this reason, we weight the reference-based score twice as much as the source text-based score. The final weights are 0.500 for RMSE, 0.167 for meaning similarity via MeaningBERT, and 0.333 for reference similarity via MeaningBERT.

5.5 Aggregation and AUTORANK Mapping

After normalization and weighting, we computed the system-level scores as weighted averages across the three metrics. These averages were linearly scaled to the range of $[1, N]$, where N is the number of valid runs. Following the WMT25 AUTORANK convention, we apply a final linear mapping where the best-performing system was assigned AUTORANK = 1, the worst-performing system was assigned AUTORANK = N, and the intermediate middle-ranking systems were spaced proportionally between these endpoints.

6 Participants and System Descriptions

6.1 Overview

We received overall 48 submissions by 20 participating teams. Each team was allowed to submit outputs of up to three systems or runs. The most dominant strategy of the submissions was prompting (28 submissions), including evaluations with a CEFR labeling system (14 submissions) or an LLM as a judge (6 submissions). Other strategies ranged from rule-based systems (4 submissions), agentic approaches (7), fine-tuning of LLMs (4), training of neuronal networks (3), and other approaches (2).

6.2 System Summaries

Archaeology (Roscan and Nisioi, 2025) submitted three submissions. For two of them, they prompt an LLM (Claude-Sonnet-4 vs. Llama-3.1-8B-Instruct) to generate simplifications iteratively. They added feedback on the enforced CEFR level to the prompt at each iteration until the level is reached or 5 runs have been completed. They select the best candidates of the model with Minimum Bayes Risk. Additionally, they fine-tuned a lightweight Llama model on synthetic data with CEFR levels A2 and B1 and repeated the process previously described, but achieved lower scores with this approach.

BU-IntelPA proposed two multi-agent systems with zero-shot simplification using either GPT-OSS-20B or Mistral-NeMo-12B. Unfortunately, we cannot provide more information as no system description paper has been submitted.

Cappuccino submitted system outputs based on zero-shot prompting using LLMs. Their prompts are focusing on simplification, meaning preservation, short sentences, controlled vocabulary, and a consistent style. Unfortunately, we cannot provide more information as this team has not submitted a system description paper.

EasyJon (Barbu et al., 2025) proposed a system that selects the best simplification out of seven LLMs' outputs. They prompt seven different LLMs (Qwen3-235B-A22B, Llama-3.3-70B-Instruct, DeepSeek-R1, Gemma-3-27B, GPT-OSS-120B, Claude-Sonnet-4, Mistral-Medium-3.1) with either short, descriptive, and descriptive prompts with examples and use an LLM-as-a-judge approach to select the best candidates.

EhiMeNLP (Miyata et al., 2025) is ranked first overall with one of their three different systems. All three systems follow the same two-fold strategy of generating simplified candidates by prompting LLMs with four prompt types (fine-grained simplification, controlling CEFR level, US grade levels, and edit operations), and evaluating the candidates based on CEFR labelling and meaning preservation relative to the source text. In their best performing submission, they ensemble several LLMs (GPT-5, GPT-4.1, o3, GPT-OSS-20B, Qwen3-32B, Llama-3.3-70B-Instruct) and all prompting strategies. In the other approaches, they use GPT-5 and combine only two of the four prompting strategies.

GPLSI submitted three runs, including one that is based on a fine-tuned Flan-T5 model, whereas the others are based on prompting Llama-3.2-3B-Instruct. Unfortunately, we cannot provide more information as this team has not submitted a system description paper.

GRIPF (Alfter and Gombert, 2025) proposed three different systems variations. The simplifications of their first system are the outputs of a discussion between two LLMs (i.e., GPT-5 and Claude-Opus-4.1,) which generate, criticize, and revise each others simplified outputs. The submitted candidate is either selected by the two models or, if they do not agree, by a third LLM (Llama-3.2-3B-Instruct), which judges the best output. In their second approach, the authors also provide specific vocabulary for each CEFR level to the LLMs. The third approach uses GPT-4o to generate the simplified candidates. Afterwards, another LLM provides feedback based on the CEFR level of the generated texts and potentially rewrites the text, provided the level matches (or after three runs).

HIT-YOU (Shimada et al., 2025) explored readability-controlled simplification with a prompt-

ing and LLM-as-a-judge approach. For two of their three approaches, they generated multiple candidates with 4 LLMs (i.e., GPT-5, Gemini-2.5-Flash, Claude-4-Sonnet, and o1) and three prompting techniques, and used an LLM-as-a-judge (either Gemini or GPT-5). Their prompts are either zero-shot, one-shot, or include a round-trip translation. The third approach contains a self-refinement loop in which a CEFR labeling system provides continuous feedback until the generated text matches the expected level or the maximum number of iterations is reached.

HOPE (Maharjan and Shrestha, 2025) proposed two rule-based approaches and an additional multi-stage pipeline. Their first rule-based approach focuses only on lexical simplification, whereas the second approach combines lexical and syntactical simplification via word substitution and sentence splitting. In the multi-stage pipeline, they combined lexical and syntactical preprocessing with zero-shot prediction of a T5 model.

HULAT-UC3M (Sanchez-Gomez et al., 2025) submitted two systems to the shared task. Both approaches are based on prompting LLMs (Ettin Suite and LLama-3). Their reinforced prompts contain either detailed descriptions of CEFR levels or only brief descriptions.

ITU (Dinç et al., 2025) explored prompting GPT-4o in a three-step manner. First, they ask the model to simplify with respect to syntactic simplification. In the second round, they ask the model to consider lexical simplification during generation based on syntactically simplified text. Finally, they ask the model also to include elaborations in the simplification. Each prompt includes some rules for simplification as well as examples for it.

Know-AI (Wu et al., 2025) proposed two different submissions. Both submissions can be summarized as an iterative generation of simplifications using GPT-4o until the target CEFR level is reached. In the first submission, the CEFR level is estimated based on an alignment between the CEFR levels and the Flesch-Kincaid Grading Level. For the other approach, they evaluate the readability with the CEFR leveling system provided by the shared task organizers.

MMU_NLP tackled readability-controlled simplification via neural networks. They use existing parallel simplification corpora and enrich them first with CEFR levels. Afterwards, they generate sentence embeddings for the additional data using SONAR and train a simple feed-forward neural network on them and the target level. This model is then used to generate simplifications of the sentences of the test set. In the different approaches of MMU_NLP, a) separate models are trained per CEFR level (run 1), b) the models are trained on all data at once (run 2), or c) are trained at once, but using 1-hot vectors (run 3). Unfortunately, this team has not submitted a system description paper, which would provide more insight into their approach.

OneNRC (Vajjala, 2025) proposed two agentic approaches using Gemini-2.5-Flash and Gemma-3-12B. For both submissions, they use zero-shot prompting and two tools to evaluate the generated simplifications. The first tool is a CEFR labeling system, and the second tool measures the preservation of meaning in the original text.

OUNLP (Huynh and Cao, 2025) submitted the output of two systems, which were built on top of each other. The first model is rule-based, whereas the second model uses the output of the first as part of an input prompt for an LLM, i.e., GPT-4o. The prompt is enriched with instructions for improving readability, e.g., synonym replacement, clause trimming, sentence splitting, and word limit restrictions.

SQUREL (Sokova et al., 2025) used a fine-tuning approach for two of their three submissions. In the first approach, they fine-tune Llama-3.2-1B Instruct with two reward functions via the CEFR level of the generated text and sentence similarity between the generated text and the source texts. In the second approach, they focus on simplifications with a larger gap between the source and target CEFR levels. Therefore, they use a more relaxed CEFR reward function that reduces penalties for larger gaps between levels. In their third approach, they focus on lexical simplification. For the substitution, they iteratively select words that could be simplified based on the WordNet lexicon. Afterwards, they use an LLM for integrating the words well into the text.

STARLING (Przybyła, 2025) proposed three submissions to the shared task. All three submissions are prompting-based approaches based on the BLESS benchmark using Gemma-3-27B. Based on multiple generated candidates, the best

one is selected using the CEFR classifier provided by the shared task organizers. This team compared whether selecting the best candidates works better when providing 5 good candidates, 10 good candidates, or 5 random candidates.

taskGen (Oviedo et al., 2025) submitted three approaches using prompting strategies with the same LLM, i.e., Llama-3.1-8B. Each of their prompt lists the relevant CEFR descriptors and examples of appropriate vocabulary, morphological, and grammatical structures. In comparison, their first submission contains no cleaning, whereas the second includes some cleaning, and the third contains candidate selection based on CEFR levels and meaning preservation.

Uniandes (Russi et al., 2025) proposed a few-shot learning and agent-based approach using different LLMs for each of the three submissions. For all submissions, they use an LLM as a judge, which provides feedback for the next iteration in the simplification loop. In the first run, Gemini-2.5-Pro is prompted, Gemini-2.5-Flash in the second, and GPT-OSS-120B and Gemini-2.5-Pro in the third.

UoL-UPF (Hayakawa et al., 2025) addressed the task of readability-controlled simplification via prompting and candidate selection out of different model and prompt combinations. For candidate selection, they use Minimum Bayes Risk and the CEFR labeler provided by the shared task organizers. One of their approaches focuses on simplifying on the paragraph level, another on the sentence level, and the third one combines the sentence and paragraph levels.

7 Results

We report two separate rankings to provide a comprehensive view of system performance. Table 2 presents every valid submission, allowing us to compare the relative performance of all system variants submitted by each team. In contrast, Table 3 shows only the highest-scoring run per team to highlight each team's most performant approach. Analyzing model variations across the submitted systems, we identified that 34 distinct LLMs were used for the shared task, of which 23 were open-source and the remaining were commercial.

Comparing approaches that used only one LLM, system submissions that used commercial ones like GPT-5 or Gemini 2.5 performed substantially better in terms of average performance (EhiMeNLP, HIT-YOU, or Uniandes) than with system submissions that used open-weight models like Llama or Gemma (taskGen, Archaeology, and STARLING). The best model, which uses only open-weight models, is taskGen, with an AUTORANK score of 7.48 using Llama-3.1-8B. On the other hand, system submissions that used multiple LLMs achieved even better results, including the top 5 submissions, where one is from EhiMeNLP and two from UoL-UPF and HIT-YOU. All these top-scoring system submissions used at least four LLM variants, except for EasyJon, which has also used a collection of seven LLMs, but was only ranked 8.88. Thus, it is clear that the selection of LLMs and the prompting technique largely affect the performance. Likewise, we also observe a clear pattern from the system submissions where the use of commercial GPT-based models leads to a substantially stronger performance.

Based on the system submissions, prompting strategies achieved, on average, the best AUTORANK placement of 7.70, followed by agentic approaches with 8.48 and rule-based approaches with 13.00. Comparing the prompting strategies, the approaches with an LLM-as-a-judge performed slightly better, with an AUTORANK placement of 6.33, than the approaches using a CEFR labeling system for evaluation, with 6.70. Due to the lack of training data, some participants used external resources, such as related simplification corpora, lexicons, and more informative descriptions of CEFR levels to enhance their prompts. The teams also used reference-less metrics, such as readability-based formulas like Flesch-Kincaid and semantic similarity via MeaningBERT, to pre-evaluate their systems and feed the results back into their models in an iterative process.

8 Discussion

Are LLMs the Only Way Forward for Text Simplification? The use of LLMs as a core resource across the majority of system submissions reflects a decisive shift away from traditional rule-based simplification methods to generative models. We observe a similar pattern with the proliferation of LLMs achieving state-of-the-art performance across general readability control benchmarks (Kew et al., 2023; Imperial et al., 2025). This transition has significant implications for evaluation, shifting from measuring output quality to assessing real-world impact (Reiter, 2025). As a

Team	Model	RMSE	MeaningBERT (Src)	MeaningBERT (Ref)	Avg	AUTORANK
EhiMeNLP ★	run1	0.000	0.902	0.845	0.636	1.000
UoL-UPF ★	uol-upf_submission3	0.000	0.856	0.857	0.603	1.410
UoL-UPF ★	uol-upf_submission1	0.000	0.849	0.856	0.590	1.580
HIT-YOU	run2_gpt_ensemble_4	0.158	0.852	0.835	0.429	3.610
HIT-YOU	run1_gemini_ensemble_4	0.187	0.863	0.833	0.424	3.670
EhiMeNLP	run3	0.234	0.847	0.840	0.390	4.100
EhiMeNLP	run2	0.200	0.838	0.816	0.322	4.970
HIT-YOU	run3_self_refine	0.245	0.822	0.820	0.282	5.460
Uniandes	run_1	0.212	0.817	0.814	0.275	5.560
Uniandes	run_2	0.200	0.825	0.803	0.260	5.740
Archaeology	claude_sonnet_4	0.122	0.779	0.804	0.238	6.010
Uniandes	run_3	0.510	0.847	0.813	0.138	7.280
taskGen	submit_3	0.628	0.856	0.826	0.122	7.480
ounlp	test_data_output_for_First_Program	0.755	0.855	0.849	0.121	7.500
Archaeology	llama_3.1_8b	0.265	0.782	0.789	0.109	7.640
BU-intelPA	run1	0.628	0.831	0.830	0.099	7.780
Cappuccino	Cappuccino_TSAR2025_Submission	0.718	0.826	0.843	0.077	8.050
GRIPF	tsar2025_ezscalar_lexical_gripf	0.689	0.857	0.820	0.070	8.130
GRIPF	tsar2025_ezscalar_nonlexical_gripf	0.721	0.856	0.824	0.060	8.270
Know-AI	run2	0.700	0.821	0.835	0.053	8.350
Know-AI	run1	0.659	0.801	0.832	0.036	8.560
EasyJon	run_1	0.822	0.838	0.836	0.011	8.880
SQUREL	SQUREL_Run3	1.153	0.979	0.819	-0.022	9.300
HULAT-UC3M	run2_llama3-8b_reinforced-prompt	0.608	0.793	0.806	-0.028	9.370
oneNRC	onenrc_google25flash_withtoolcall	0.534	0.772	0.800	-0.033	9.440
STARLING	starling_1_g5-best	0.621	0.811	0.791	-0.053	9.690
ITUNLP	itunlp	0.632	0.797	0.797	-0.063	9.820
oneNRC	onenrc_gemma312b_react_notool	0.579	0.761	0.803	-0.069	9.880
SQUREL	SQUREL_Run1	0.718	0.821	0.797	-0.076	9.980
taskGen	submit_1	0.592	0.791	0.786	-0.084	10.070
HULAT-UC3M	run1_llama3-8b_reinforced-prompt	0.682	0.790	0.791	-0.122	10.560
GRIPF	tsar2025_saga_gripf	0.831	0.827	0.796	-0.140	10.780
SQUREL	SQUREL_Run2	0.632	0.779	0.778	-0.153	10.950
STARLING	starling_2_g10-best	0.678	0.795	0.777	-0.160	11.040
Archaeology	llama_3.2_1b	0.212	0.706	0.731	-0.165	11.100
taskGen	submit_2	0.561	0.752	0.773	-0.169	11.150
STARLING	starling_3_g5-random	0.812	0.816	0.785	-0.180	11.280
HOPE	HOPE_run1	1.428	0.945	0.815	-0.255	12.230
GPLSI	run1_llama_knowledge	0.998	0.865	0.772	-0.258	12.270
GPLSI	run2_llama_zs	0.640	0.772	0.750	-0.258	12.270
ounlp	test_data_output_for_Second_Program	0.714	0.865	0.701	-0.313	12.960
HOPE	HOPE_EXPERT_run1	1.402	0.919	0.797	-0.337	13.260
MMU_NLP	mmu_tsar25_test_system2	1.005	0.845	0.754	-0.350	13.430
BU-intelPA	run2	0.612	0.715	0.739	-0.368	13.650
MMU_NLP	mmu_tsar25_test_system3	1.010	0.832	0.752	-0.381	13.830
MMU_NLP	mmu_tsar25_test_system_1	0.822	0.735	0.676	-0.664	17.390
HOPE	HOPE_SOTA_run1	1.600	0.841	0.730	-0.795	19.030
GPLSI	run3_flan_knowledge	0.883	0.221	0.182	-3.093	48.000

Table 2: Final ranked results for all submitted runs using AUTORANK with custom weighting.

Team	Model	RMSE	MeaningBERT (Src)	MeaningBERT (Ref)	Avg	AUTORANK
EhiMeNLP ★	run1	0.000	0.902	0.845	0.636	1.000
UoL-UPF ★	uol-upf_submission3	0.000	0.856	0.857	0.603	1.410
HIT-YOU ★	run2_gpt_ensemble_4	0.158	0.852	0.835	0.429	3.610
Uniandes	run_1	0.212	0.817	0.814	0.275	5.560
Archaeology	claude_sonnet_4	0.122	0.779	0.804	0.238	6.010
taskGen	submit_3	0.628	0.856	0.826	0.122	7.480
ounlp	test_data_output_for_First_Program	0.755	0.855	0.849	0.121	7.500
BU-intelPA	run1	0.628	0.831	0.830	0.099	7.780
Cappuccino	Cappuccino_TSAR2025_Submission	0.718	0.826	0.843	0.077	8.050
GRIPF	tsar2025_ezscalar_lexical_gripf	0.689	0.857	0.820	0.070	8.130
Know-AI	run2	0.700	0.821	0.835	0.053	8.350
EasyJon	run_1	0.822	0.838	0.836	0.011	8.880
SQUREL	SQUREL_Run3	1.153	0.979	0.819	-0.022	9.300
HULAT-UC3M	run2_llama3-8b_reinforced-prompt	0.608	0.793	0.806	-0.028	9.370
oneNRC	onenrc_google25flash_withtoolcall	0.534	0.772	0.800	-0.033	9.440
STARLING	starling_1_g5-best	0.621	0.811	0.791	-0.053	9.690
ITUNLP	itunlp	0.632	0.797	0.797	-0.063	9.820
HOPE	HOPE_run1	1.428	0.945	0.815	-0.255	12.230
GPLSI	run1_llama_knowledge	0.998	0.865	0.772	-0.258	12.270
MMU_NLP	mmu_tsar25_test_system2	1.005	0.845	0.754	-0.350	13.430

Table 3: Best run per team using AUTORANK with custom weighting.

result, the central question is no longer *"Can this model generate simplified texts?"* But rather, *"Are this model's outputs of sufficient quality to be used with CEFR-based learners?"* Hence, this shared task supports further research on pedagogical validations (e.g., expert-in-the-loop) to ensure that LLM-generated simplifications are aligned with CEFR-based learning objectives rather than producing superficial, simpler texts.

Where Do We Position Synthetic Data? While the shared task was ultimately successful in producing a new parallel CEFR-based reference dataset for validation and testing, it is worth noting the substantial effort and financial support required for resource development. In line with this, we ask the question *"Is it time to use LLM-generated CEFR data to complement expert-produced data?"* While adjacent NLP tasks like grammatical correction and essay scoring have benefited from the performance advantages of synthetic data, its practicality for readability and text simplification applications has only ever been explored recently (Stahlberg and Kumar, 2021; Klöser et al., 2024; Latouche et al., 2024; Qwaider et al., 2025). For text simplification that is anchored on a real-world language proficiency framework like CEFR, there are both opportunities and risks. On the one hand, LLMs are capable of generating fluent text guided by CEFR specifications, as evident in the top submissions of this shared task. This could potentially be valuable for low-resource languages across various domains, genres, and text types where expert-annotated resources are scarce. However, using synthetic data without careful validation checks risks producing noisy approximations of CEFR levels, which may reinforce undesirable simplification patterns. As such, our stance on this is that we should establish a community-accepted framework to integrate synthetic data, involving steps such as filtering and expert validation stages. For the next iteration of this shared task, we propose a direction exploring how the use of LLM-generated synthetic data can perform well on test and validation data generated by experts.

Cost-Performance Tradeoffs Our leaderboard results show that top-ranked systems achieved strong performance by leveraging ensembles of multiple LLMs. However, these submissions come with substantial computational and financial costs. For example, the rank-one system `EhiMeNLP` combined six LLMs, including GPT-5, GPT-4.1, o3, GPT-OSS-20B, Qwen3-32B, and Llama-3.3-70B-Instruct to generate up to 120 candidate simplifications per input. While such methods clearly demonstrate the potential of ensemble-based techniques for producing precise readability-controlled simplifications, their heavy resource demands may limit their adoption in educational and resource-constrained contexts where computational budget to run more than one commercial model is unavailable. Future work should therefore explore approaches that strike a balance between achieving decent performance and computational efficiency. Likewise, combining *resource-aware evaluation procedures* with performance-based metrics may encourage participants to propose innovative and computationally viable methods that are effective in real-world CEFR-based text simplification settings.

9 Conclusion

The TSAR 2025 Shared Task introduced a benchmark for readability-controlled text simplification explicitly aligned with CEFR levels. Two main resources were developed as contributions to the community: a) a CEFR-aligned dataset of pedagogical paragraph-level English texts simplified to A2 and B1 levels, and b) a CEFR evaluator model fine-tuned to estimate text difficulty along the CEFR scale.

The results from participating teams indicate that although LLMs achieve strong performance on this task, achieving dependable and fine-grained control over simplification often relies on complex, iterative generation strategies. Moreover, our analysis suggests that current systems are approaching the limits of existing automatic evaluation metrics, underscoring the need to adapt these metrics for greater robustness and practical relevance.

Limitations

We acknowledge several limitations in the conduct of this shared task, which we believe can serve as a springboard for future iterations.

Dataset and Annotation Coverage. For our primary shared task dataset, we use a single-source English-only dataset from The British Council. This is the result of our prioritized search for a new gold-standard CEFR-based parallel dataset that has not been published before, as a contribution to the community from the Shared Task. The

dataset we acquired was provided in paragraph-level format due to the unavailability of resources in longer forms. One can argue for transforming the paragraph-level data into sentence-level data to account for variance; however, we did not pursue this option since sentence-level CEFR data are already available, such as CEFR-SP (Arase et al., 2022) and ReadMe (Naous et al., 2024). Likewise, due to budget constraints, we were only able to provide one expert-written simplification for each instance of the trial and test set. We acknowledge that expanding this, for example, by asking more language experts to produce separate annotations will allow better convergence of evaluation scores.

Automatic Metrics for System Evaluations. Our main evaluation pipeline, which determined the system rankings, primarily relied on automatic metrics including a combination of weighted RMSE and ModernBERT scores, which compared system outputs to reference and target expert-written simplifications. While this setup is convenient from a shared task perspective, an additional round of expert validation from text simplifications of each system submission would be valuable in assessing linguistic and pedagogical appropriateness that automatic metrics do not capture. However, we were unable to conduct this due to time and funding constraints.

Ethics Statement

The dataset used for this shared task was acquired with the British Council's formal permission. The collection of manual simplifications received a favourable opinion from the Ethics Committee of the School of Computer Science and Informatics at Cardiff University. All artifacts, including the dataset with expert reference simplifications and evaluation scripts, will be released to the research community to support future work. All participating teams are credited for their submissions. System description papers were included when available, ensuring transparency and proper attribution of their methods.

We acknowledge that the LLMs used by participants and in our evaluation tools may contain inherent biases reflecting their training data. This work is an analysis of system performance and does not constitute an endorsement of these models for direct pedagogical applications without further expert-in-the-loop validation. Our aim is to benchmark the current state of the art to encourage the responsible development of text simplification technology.

Acknowledgements

We thank all participants who expressed interest in joining the shared task and worked towards submitting their proposed systems. We also thank the annotators for their valuable contributions to create the reference text simplification dataset. Regina Stodden was supported by the European Regional Development Fund within the project: LLM4KMU - Optimierter Einsatz von Open Source Large Language Modellen in KMU. Joseph Imperial is supported by the National University Philippines and the UKRI Centre for Doctoral Training in Accountable, Responsible, and Transparent AI [EP/S023437/1] of the University of Bath.

References

Sweta Agrawal and Marine Carpuat. 2023. Controlling pre-trained language models for grade-specific text simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819, Singapore. Association for Computational Linguistics.

David Alfter and Sebastian Gombert. 2025. GRIPF at TSAR 2025 Shared Task: Towards controlled CEFR level simplification with the help of inter-model interactions. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.

Paul-Gerhard Barbu, Adrianna Lipska-Dieck, and Lena Lindner. 2025. EasyJon at TSAR 2025 Shared Task: Evaluation of Automated Text Simplification with

LLM-as-a-Judge. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. Meaningbert: assessing meaning preservation between sentences. *Frontiers in Artificial Intelligence*, 6.

Mark Breuker. 2022. CEFR labelling and assessment services. In *European Language Grid: A Language Technology Platform for Multilingual Europe*, pages 277–282. Springer International Publishing Cham.

Kutay Arda Dinç, Fatih Bektaş, and Gülşen Eryiğit. 2025. ITU NLP at TSAR 2025 Shared Task: A Three-Stage Prompting Approach for CEFR-Oriented Text Simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339, Torino, Italia. ELRA and ICCL.

Akio Hayakawa, Nouran Khallaf, and Horacio Saggion. 2025. UoL-UPF at TSAR 2025 Shared Task: A Generate-and-Select Approach for Readability-Controlled Text Simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Cuong Huynh and Jie Cao. 2025. OUNLP at TSAR 2025 Shared Task: AI-Generated Multi-Round Sentence Simplifier. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. UniversalCEFR: Enabling Open Multilingual Research on Language Proficiency Assessment. *arXiv preprint arXiv:2506.01419*.

Joseph Marvin Imperial, Gail Forey, and Harish Tayyar Madabushi. 2024. Standardize: Aligning language models with expert-defined standards for content generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1573–1594, Miami, Florida, USA. Association for Computational Linguistics.

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

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

Lars Klöser, Mika Beele, Jan-Niklas Schagen, and Bodo Kraft. 2024. German text simplification: Finetuning large language models with semi-synthetic data. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 63–72, St. Julian's, Malta. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, and 9 others. 2025. Preliminary Ranking of WMT25 General Machine Translation Systems. *Preprint*, arXiv:2508.14909.

Gaetan Lopez Latouche, Marc-André Carbonneau, and Benjamin Swanson. 2024. Zero-shot cross-lingual transfer for synthetic data generation in grammatical error detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3002–3016, Miami, Florida, USA. Association for Computational Linguistics.

Sujal Maharjan and Astha Shrestha. 2025. HOPE at TSAR 2025 Shared Task: Balancing Control and Complexity in Readability-Controlled Text Simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

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

Rina Miyata, Koki Horiguchi, Risa Kondo, Yuki Fujiwara, and Tomoyuki Kajiwara. 2025. EhiMeNLP at TSAR 2025 Shared Task: Candidate Generation via Iterative Simplification and Reranking by Readability and Semantic Similarity. In *Proceedings of*

the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025), Suzhou, China. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable text simplification with lexical constraint loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266, Florence, Italy. Association for Computational Linguistics.

Juan Cruz Oviedo, Elisabet Comelles Pujadas, Laura Alonso Alemany, and Jordi Atserias Batalla. 2025. taskGen at TSAR 2025 Shared Task: Exploring prompt strategies with linguistic knowledge. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Piotr Przybyła. 2025. STARLING at TSAR 2025 Shared Task: Leveraging Alternative Generations for Readability Level Adjustment in Text Simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Chatrine Qwaider, Bashar Alhafni, Kirill Chirkunov, Nizar Habash, and Ted Briscoe. 2025. Enhancing Arabic automated essay scoring with synthetic data and error injection. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 549–563, Vienna, Austria. Association for Computational Linguistics.

Ehud Reiter. 2025. We Should Evaluate Real-World Impact. *Computational Linguistics*, pages 1–13.

Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687, Singapore. Association for Computational Linguistics.

Rares-Alexandru Roscan and Sergiu Nisioi. 2025. Archaeology at TSAR 2025 Shared Task: Teaching Small Models to do CEFR Simplifications. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Felipe Arias Russi, Kevin Cohen Solano, and Ruben Manrique. 2025. Uniandes at TSAR 2025 Shared Task: Multi-Agent CEFR Text Simplification with Automated Quality Assessment and Iterative Refinement. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Jesus M. Sanchez-Gomez, Lourdes Moreno, Paloma Martínez, and Marco Antonio Sanchez-Escudero. 2025. HULAT-UC3M at TSAR 2025 Shared Task on Readability-Controlled Text Simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. Text simplification from professionally produced corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, and 3 others. 2024. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Se-*

mantic Evaluation (SemEval-2021), pages 1–16, Online. Association for Computational Linguistics.

Mao Shimada, Kexin Bian, Zhidong Ling, and Mamoru Komachi. 2025. HIT-YOU at TSAR 2025 Shared Task: Leveraging Similarity-Based Few-Shot Prompting, Round-Trip Translation, and Self-Refinement for Readability-Controlled Text Simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Daria Sokova, Anastasiia Bezobrazova, and Constantin Orasan. 2025. SQUREL at TSAR 2025 Shared Task: CEFR-Controlled Text Simplification with Prompting and Reinforcement Fine-Tuning. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. SemEval-2012 task 1: English lexical simplification. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355, Montréal, Canada. Association for Computational Linguistics.

Nicolas Spring, Annette Rios, and Sarah Ebling. 2021. Exploring German multi-level text simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1339–1349, Held Online. INCOMA Ltd.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online. Association for Computational Linguistics.

Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463, Toronto, Canada. Association for Computational Linguistics.

Satoru Uchida, Shohei Takada, and Yuki Arase. 2018. CEFR-based lexical simplification dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sowmya Vajjala. 2025. OneNRC at TSAR2025 Shared Task: Small Models for Readability Controlled Text Simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Laura Vásquez-Rodríguez, Pedro-Manuel Cuenca-Jiménez, Sergio Morales-Esquivel, and Fernando Alva-Manchego. 2022. A benchmark for neural readability assessment of texts in Spanish. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 188–198, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Yiheng Wu, Anisia Katinskaia, Jue Hou, and Roman Yangarber. 2025. Know-AI at TSAR 2025 Shared Task: Difficulty-aware Text Simplification System. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. 2022. Controllable text simplification with deep reinforcement learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 398–404, Online only. Association for Computational Linguistics.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Tatsuya Zetsu, Tomoyuki Kajiwara, and Yuki Arase. 2022. Lexically constrained decoding with edit operation prediction for controllable text simplification.

In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 147–153, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

A Appendix

A.1 Data Distribution

Table 6 shows the distribution of instances across CEFR levels in the training, validation, and test splits.

A.2 Hyperparameter Values

Table 4 reports the hyperparameters used to fine-tune the pre-trained MODERNBERT model.

Hyperparameter	Value
Learning rate	3.6×10^{-5}
Train batch size	3
Evaluation batch size	3
Random seed	42
Gradient accumulation steps	16
Total effective batch size	48
Optimizer	`adamw_torch_fused`
Betas	$(0.9, 0.999)$
Epsilon	10^{-8}
Learning-rate scheduler	linear
Warm-up ratio	0.1

Table 4: Hyperparameter values used to fine-tune the pre-trained MODERNBERT model.

A.3 Annotators' Reliability

Table 5 shows different metrics between the target CEFR level assigned during annotation with the levels predicted by our automatic CEFR evaluator model in the shared task dataset.

Annotator	Data (N)	ρ	Acc	AdjAcc	RMSE
1	Trial (20)	0.68	0.63	1.00	0.61
2	Test (20)	0.37	0.65	1.00	0.59
3	Test (80)	0.43	0.61	1.00	0.62

Table 5: Reported reliability and accuracy scores of the annotators with respect to the data splits they produced the reference text simplifications.

A.4 Classifier Performance

Table 8 presents the performance of the CEFR classifier in the validation set.

Split	A1	A2	B1	B2	C1	C2	Total
TRAIN_DOC_EN	17	122	152	148	115	96	650
TRAIN_DOC_SENT_EN	323	2,066	4,080	4,374	2,236	397	13,476
REFERENCE_ALLLANG	2,318	19,838	22,270	7,257	3,802	1,478	56,963
VALIDATION	4	26	33	31	24	21	139
TEST	3	26	33	32	25	21	140

Table 6: Distribution of instances across CEFR levels in training, validation, and test splits.

Source Name	Language
`cefr-sp` (Arase et al., 2022)	en
`apa-lha` (Spring et al., 2021)	de
`deplain-apa-doc` (Stodden et al., 2023)	de
`deplain-apa-sent` (Stodden et al., 2023)	de
`deplain-web-doc` (Stodden et al., 2023)	de
`elg-cefr-de` (Breuker, 2022)	de
`elg-cefr-nl` (Breuker, 2022)	nl
`hablacultura` (Vásquez-Rodríguez et al., 2022)	es
`kwiziq` (Vásquez-Rodríguez et al., 2022)	es
`kwiziq` (Imperial et al., 2025)	fr
`learn_welsh_cy` (Imperial et al., 2025)	cy
`readme` (Naous et al., 2024)	en, ar, fr, hi, ru

Table 7: List of datasets and languages included in the REFERENCE_ALLLANG split

Model Setup	A1	A2	B1	B2	C1	C2	Avg	AdjAcc	RMSE
TRAIN_DOC_EN	0.57	0.85	0.77	0.81	0.78	0.91	0.81	0.97	0.50
TRAIN_DOC_SENT_EN	0.40	0.87	0.82	0.81	0.84	0.95	0.84	0.99	0.42
REFERENCE_ALLLANG	0.00	0.84	0.79	0.78	0.81	0.95	0.80	1.00	0.43
MAJORITY VOTE	0.40	0.87	0.84	0.82	0.84	0.95	0.84	0.99	0.42
CONFIDENCE-BASED	0.40	0.88	0.87	0.84	0.86	0.98	0.87	0.99	0.39

Table 8: Performance of various training data and model prediction setups integrated with ModernBERT-Base on the **validation set**. We selected the CONFIDENCE-BASED setup for our final CEFR evaluator model due to its optimal performance (low RMSE and high averages).