Know-AI at TSAR 2025 Shared Task: Text Simplification with Difficulty Critics

Yiheng Wu[‡] **Jue Hou**[†‡] **Anisia Katinskaia**[‡] **Roman Yangarber**[‡]

[†]Department of Computer Science
[‡]Department of Digital Humanities
University of Helsinki, Finland
first.last@helsinki.fi

Abstract

Text simplification is an active research topic with applications in many domains. In a simplification pipeline, assessment of text difficulty plays a crucial role as a quality control mechanism: it acts as a "critic," and guides models to generate text at the difficulty level that is required by the user. In this paper, we present a LLM-based Text Simplification pipeline using several difficulty critics. We evaluate the pipeline using the TSAR shared task dataset and discuss the challenges in building models for assessment of text difficulty and simplification, including the construction of corpora for training difficulty models.

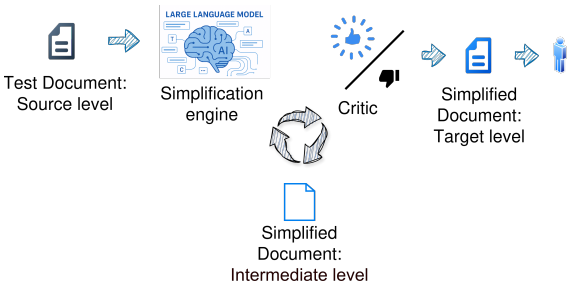

Figure 1: Overview of simplification pipeline.

1 Introduction

Text simplification is a widely studied task in natural language processing (NLP), with applications in accessibility, education, and communication. It is important in many applications where the users—e.g., non-native speakers—struggle to understand complex or standard language. The goal is to reduce the linguistic complexity of a text, while maintaining the original text's core meaning and coherence. Increasingly, official legislation in Europe (Inclusion Europe) requires government organizations, NGOs and other public agencies to provide information to clients in clear and accessible form, including for readers who may be unable to understand *standard* language. We are motivated especially by applications of simplification in second-language (L2) education, where *personalized* learning is supported by adapting text to the learner's proficiency level (Katinskaia and Yangarber, 2018; Hou et al., 2019).

Our simplification pipeline,[1] shown in Figure 1, uses a critic consisting of two parts: (a) *difficulty*—it evaluates the difficulty level of a text simplified by a large language model (LLM), and (b) *semantic similarity*—it checks how well the simplified text preserves the semantics/meaning of the original text. This framework was introduced in (Katinskaia et al., 2025), in L2 education. In this paper, we adapt the framework for simplification in English. The pipeline iteratively attempts to generate a "simplified" version of an input text. If the generated text is above the target level of difficulty, then feedback—including the generated text and its currently assessed level—is sent back to the LLM to revise the output. The pipeline makes several attempts at simplification to reach the target difficulty level. We experiment with several critics in the pipeline, including an open-source transformer-based model that classifies text by difficulty level, and a regression model that we train using English-language texts labeled with difficulty levels.

The paper is organized as follows: Section 2 gives a brief overview of related work. Section 3 describes the shared task and the evaluation methods. Section 4 presents the architecture of our simplification pipeline. Section 4.1 describes the experiments with controlling the behavior of a LLM via the difficulty critic. Section 5 presents results and analysis. Section 6 concludes the paper and discusses directions for future work.

[1] simplification.py

Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 266–272

2 Related Work

Prior approaches to text simplification relied on assessment of text difficulty to identify sentences requiring simplification. For example, Gasperin et al. (2009) trained a model to detect linguistically complex sentences; Aluísio et al. (2010) developed readability assessment tools to support simplifying texts for low-literacy readers. Readability metrics have also been incorporated directly into rule-based simplifiers: Woodsend and Lapata (2011) integrate the Flesch-Kincaid grade formula (Flesch, 1948) into optimization-based simplification.

More recent approaches to simplification leverage readability predictors as feedback within generation loops. Alkaldi and Inkpen (2023) use a readability classifier in a reinforcement learning framework to iteratively simplify text until it reaches the desired difficulty. Large-scale neural systems have combined readability prediction with controllable generation techniques to produce text at the target difficulty level (Agrawal and Carpuat, 2023).

3 Task Description

The Shared Task on Readability-Controlled Text Simplification (Alva-Manchego et al., 2025) involves simplifying English-language paragraphs written at upper-intermediate or advanced levels. Participants are required to produce simplified versions at a target readability, specified as a CEFR level: Common European Framework of Reference for Languages (Council of Europe, 2001).

Our experiments are based on the *test* dataset provided by the TSAR shared task. The test set consists of English paragraphs at level B2 or higher, each associated with a target level (A1, A2, or B1). No training data, and no reference simplifications are provided. The evaluation involves measuring multiple aspects of the simplified texts:

- Compliance with target CEFR level is determined using a CEFR-level classifier, which checks whether the generated paragraph meets the specified target proficiency level.
- Meaning preservation is assessed via semantic similarity between the *original* source paragraph and the simplification, ensuring that the essential meaning is retained.
- Similarity to a *reference* simplification is computed, to quantify how closely the system output matches the provided references.

These metrics are calculated using the official evaluation scripts released by the shared task organizers with the test dataset. The semantic similarity in the evaluation scripts uses `meaningbert` (Beauchemin et al., 2023). `meaningbert` is a BERT-based semantic similarity model that measures how well meaning is preserved between two texts, particularly for tasks such as text simplification and paraphrase assessment.

4 System Overview

We next describe how we use the critic model to guide in LLM-based text simplification pipeline (see Figure 1).

The pipeline begins by determining the difficulty of a source text, either with a difficulty model or manual annotation. The text, together with the target CEFR level and a prompt, is passed to a LLM, which produces a candidate output. The critic model evaluates the candidate's difficulty; if it matches the target level, the process ends. Otherwise, the LLM is re-prompted with the previous output and the discrepancy from the target. This loop continues for up to N iterations—a predefined maximum, to balance between cost and quality. The system then outputs either a satisfactory simplification, or an error if the target is not reached.

4.1 Methodology

In the context of the shared task, we experiment with two difficulty assessment models in the pipeline critic:

- **Statistical model**: we use the Flesch–Kincaid Reading Ease score (Flesch, 1948; Kincaid et al., 1975), implemented in the Spacy library.[2] This model assigns a numeric readability value based on word and sentence length, with higher scores indicating simpler text. To relate these scores to CEFR levels, we apply an approximate mapping[3] shown in Table 1. This enables us to interpret Flesch–Kincaid scores within a CEFR framework and use them as difficulty estimates in the simplification pipeline.
- **Transformer-based model**: we use the model `AllLang2-Cefr2`,[4] which classifies its input into the 6 CEFR levels: A1–C2. This model is also used in the official evaluation in the Shared Task. We use its prediction on the

[2] spacy.io/universe/project/spacy_readability
[3] Flesch-Kincaid readability analysis and CEFR map
[4] ModernBERT-base-reference_AllLang2-Cefr2

LLM-generated text to determine whether to stop the iterative simplification process.

To control the complexity of the generated texts, we use LLM prompts, based on the target CEFR level. Each prompt has three main components, parameterized by the target CEFR level:

- **Role Instruction:** The LLM is instructed to act as an *expert in teaching English*, to adapt English texts for learners to the specified CEFR level (`level_target`).
- **Output Format:** The LLM must produce a JSON object containing the key `"SIMPLIFICATION"`, to ensure that the result is structured and machine-readable.
- **Adaptation Guidelines:** The LLM is instructed to adapt the input `text` according to the target CEFR: the prompt contains a description of what the reader can/cannot understand easily (based on the definitions of the CEFR levels). The simplified text should *preserve the meaning* of the original text while matching the target proficiency level.

Detailed prompt templates for all CEFR levels are provided in Appendix A.

To perform the simplification, the pipeline uses GPT-4o (OpenAI, 2024) with the prompts described above. For all test documents, we monitored the simplification process by recording the CEFR level at each iteration and computing the cosine similarity between each intermediate output and the original text.

To measure semantic similarity in the critic—to check how well the simplification preserves meaning—we applied a semantic similarity model `all-mpnet-base-v2` (Reimers and Gurevych, 2020), and used a threshold of 0.7 (determined heuristically), retaining only those simplifications that have semantic similarity to the original above this value. Each document was simplified for up to $N = 5$ iterations; the process terminates earlier if the critic judges the text's difficulty to be at or below the target CEFR level, and its similarity with the original is above the threshold.

5 Results and Analysis

In this section, we examine the results of simplification with different critic models. Beyond exact-match accuracy, we assess how well the predicted difficulty levels match the intended simplification direction. The *Direction Consistency* metric measures whether predictions respect the target level ordering for each input.

Flesch-Kincaid	CEFR
90–100	A1
80–89	A2
70–79	B1
60–69	B2
50–59	C1
0–49	C2

Table 1: Mapping from Flesch-Kincaid Reading Ease scores to CEFR levels.

Consistent Example			**Inconsistent Example**		
Target	**Pred**	**Cons.?**	**Target**	**Pred**	**Cons.?**
B1	B1	Yes	B1	B1	No
A2	A1		A2	B1	

Table 2: Examples of direction consistency: left = consistent (trend preserved), right = inconsistent (trend violated).

Using the Flesch–Kincaid Reading Ease model as critic in the simplification pipeline, exact match between predicted and target CEFR levels is 38%: only a minority of simplifications reach the target level. Direction consistency measure is 62%, indicating that while the model often misses the exact target, it usually adjusts difficulty in the correct relative direction. Thus, the model offers coarse guidance on text difficulty, but lacks the precision needed for strict level control.

Figure 2 shows that most errors are deviations of ± 1 level, with about 60% of the misclassified samples exceeding the target by one level. Exact matches and predictions one level below are less frequent, and deviations beyond ± 2 levels are rare. Figure 3 illustrates that when the target is A2, outputs often simplify to level B1, while for B1 target, many texts remain at the original difficulty. In rare cases, predictions even drop to A1. These patterns indicate that the model fails to align reliably with CEFR standards, limiting the accuracy of the critic and yielding only modest control over target difficulty in the simplification pipeline.

Using `AllLang2-Cefr2` as critic in the simplification pipeline, performance improves. Exact-

		meaningbert	
Critic model	*RMSE*	*Origin.*	*Refer.*
Flesch-Kincaid	0.659	0.801	0.832
`AllLang2-Cefr2`	0.700	**0.821**	**0.835**
Regression	**0.600**	0.772	0.815

Table 3: Performance of difficulty critics on *simplification*

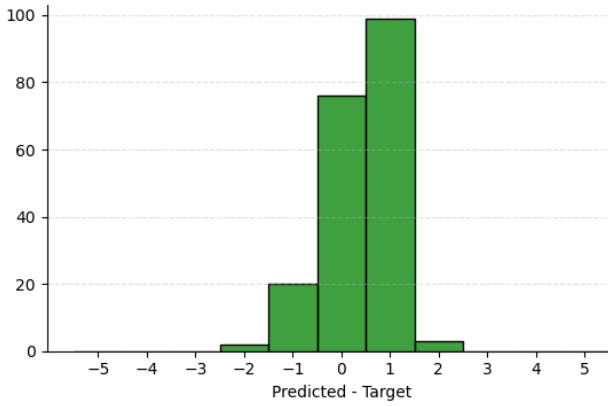

Figure 2: Distribution of the difference between estimated difficulty level of the simplified output and target difficulty level, using Flesch-Kincaid model as critic. X-axis is difference between the estimated level of the output (simplified) text and target level. Y-axis is the *number* of instances in test set.

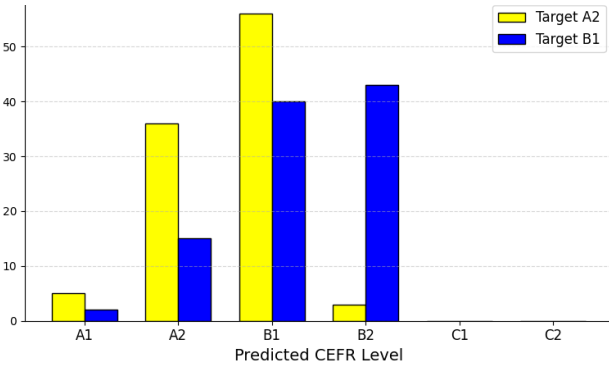

Figure 3: Distribution of estimated difficulty of simplified output texts for different target levels (A2,B1), using Flesch-Kincaid critic. X-axis represents the estimated CEFR levels (A1–C2) of output text; Y-axis indicates the *percentage* of samples at each estimated level. Different target levels are color coded.

match accuracy rises to 57%, well above that of the Spacy critic, while Direction Consistency is 63.5%. As shown in Figure 4, deviations never exceed ± 1 level, with exact matches most common and overshooting by one level less often. Figure 5 shows that when the target is A2, most samples are correctly simplfied to A2, with the rest at B1. For target B1, about 60% reach B1, some remain at B2, and the rest overshoot to A2.

The official ranking in the Shared Task is based on (a weighted average of) the three measures reported in Table 3, namely, on correct difficulty and on semantic similarity; *origin* indicates similarity of the simplified text to the original, according to `meaningbert`; *refer* indicates similarity of the simplified text to reference simplification.

6 Discussion and Future Work

The effectiveness of our proposed pipeline depends on the choice of difficulty assessment model used in the critic, since it guides the simplification pro-

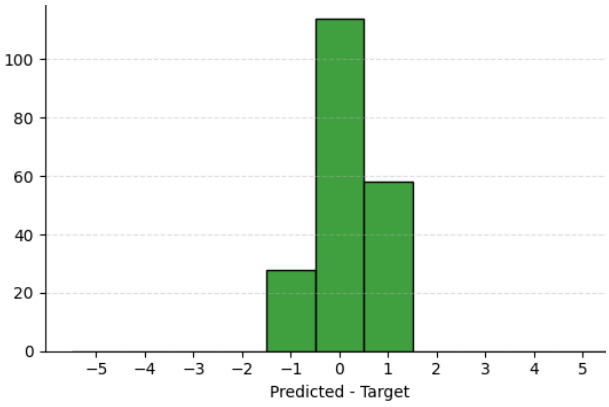

Figure 4: Distribution of the difference between estimated difficulty level of the simplified output and target difficulty level, using `AllLang2-Cefr2` as critic.

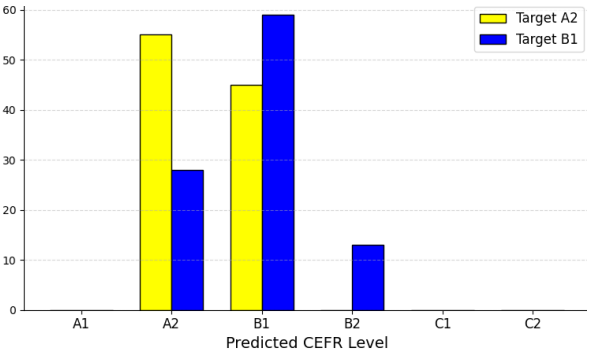

Figure 5: Distribution of estimated difficulty of simplified output texts for different target levels, using `AllLang2-Cefr2` as critic.

cess. In addition to the models above, we experimented with training our own difficulty assessment model. Although this approach did not appear in our submissions for the Shared Task, it shows much promise for future work. This section summarizes the lessons learned from this attempt.

First, since no training data were provided for the Shared Task, we construct a training, development and test set—Test Set 1—by taking an existing corpus[5] described in (Katinskaia et al., 2025) and translating it from Finnish into English, using the OPUS machine translation (MT) toolkit[6] (Tiedemann et al., 2023). It is crucial to note that we found that the OPUS models are particularly strong at preserving the CEFR levels of the original source text in the MT output text.[7] We also use the reference set provided by TSAR as a second test

[5]Test Set 1 contains intermediate CEFR levels: A2-B1, B1-B2, etc. For comparability, we applied a special adjustment for `AllLang2-Cefr2`, in which each intermediate level was "mapped" down to the lower adjacent level.

[6]We use the sla-eng MT model (Slavic-to-English).

[7]This property of the OPUS-MT models—that they preserve the CEFR level well from the input to the output text—was confirmed through manual inspection by experts in Simple Language. These findings need to be confirmed more rigorously in quantitative terms, in future work.

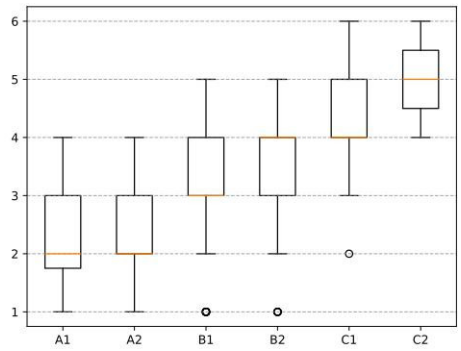

Figure 6: Difficulty estimation using `AllLang2-Cefr2` in Test Set 1. Red line means the median of current CEFR level

set—Test Set 2.

Second, following the methodology of Katinskaia et al. (2025), we train a regression model to predict difficulty. We were unable to gather a sufficient amount of training data and tune our regression model in time for the actual TSAR competition; therefore, as a fallback, we used `AllLang2-Cefr2` rather than the regression model as a critic in our submission for the Shared Task.

We next check how well difficulty prediction works—on its own, apart from the simplification task. For Test Set 1, the difficulty prediction results are in Figures 6 and 7. The regression model shows a clear advantage over the `AllLang2-Cefr2` model, exhibiting a clear step-wise pattern that aligns well with CEFR levels. It consistently outperforms the baseline across all evaluation metrics. The evaluation metrics for difficulty prediction are shown in the top part of Table 4.

For Test Set 2, the evaluation metrics for difficulty prediction are in the bottom of Table 4. The R^2 values are negative for both models, indicating a limited overall fit to the data. The difficulty prediction results for Set 2 are in Figures 8 and 9.

Several factors may compromise the performance of our regression model. First, the dataset

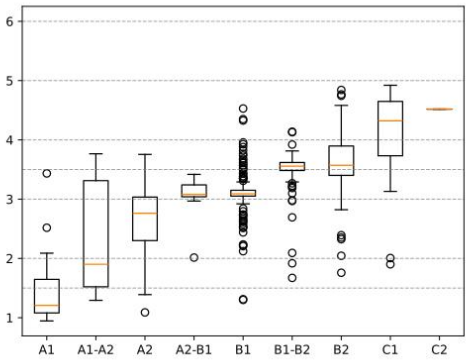

Figure 7: Difficulty estimation error distribution of regression model in Test Set 1.

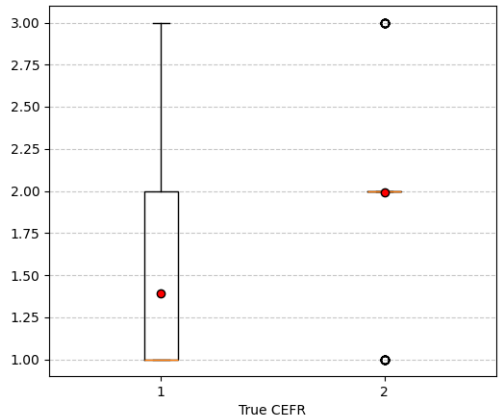

Figure 8: Difficulty estimation using `AllLang2-Cefr2` in Test Set 2. Red dot shows the mean score of this CEFR level

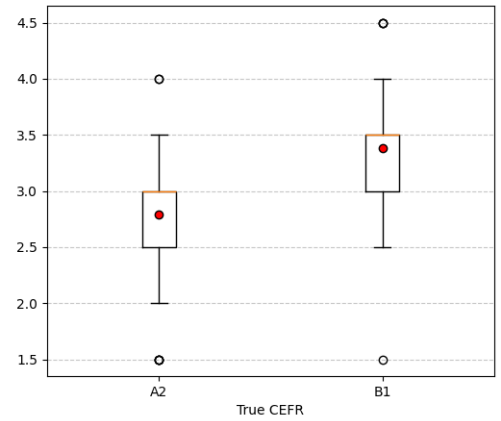

Figure 9: Difficulty estimation using regression model in Test Set 2.

Test	Model	MSE	RMSE	MAE	R^2
1	`AllLang2-Cefr2`	1.12	1.06	0.82	-0.46
	Regression	**0.32**	**0.57**	**0.34**	**0.56**
2	`AllLang2-Cefr2`	0.43	0.66	0.39	-0.72
	Regression	0.64	0.80	0.65	-1.57

Table 4: Performance of difficulty estimation models on two test sets; top section Test Set 1, bottom Test Set 2.

is machine-translated, which may distort the true difficulty of the texts. Ideally, training data is manually annotated for difficulty. However, manual annotation is very complex and time-consuming. Second, the translated dataset is still small, restricting the model's ability to generalize across different linguistic phenomena.

In future work, we plan to extend the setup relying solely on GPT-4o for text simplification, to consider other models, including smaller models fine-tuned for the simplification task. We will investigate more advanced models to improve the assessment of difficulty, which is central for the simplification pipeline. Larger, more accurate, and more diverse training datasets should further improve performance and generalization.

7 Lay Summary

This study investigates text simplification, in the context of the Shared Task on Text Simplification, Accessibility, and Readability (TSAR).

We present a difficulty-aware simplification pipeline based on large language models (LLMs) and small models for simplification assessment. We use text data in English, of varying levels of difficulty, ranging from A1 to C1 on the CEFR scale. We evaluate performance according to several criteria, including error rates of difficulty assessment models in their assessment of the difficulty of texts in a held-out test set, and the success rates of the simplification pipeline, relative to reference texts provided by the organizers of the shared task.

The paper A. discusses the performance of a number of critic models for assessing difficulty of a text, and B. compares the performance of the simplification pipeline driven by the different critics.

Acknowledgements

This work was supported by Project *"Easy Language for accessible workplace communication,"* funded by BusinessFinland Agency for Technology and Innovation (Grant 4173/31/2024); Project *"Generative AI-Enhanced Knowledge Management in Business"* (GAIK), funded by the European Regional Development Fund (EAKR), and High-performance Computing Project *"Know-AI,"* funded by Academy of Finland (Grant 359285).

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.

Wejdan Alkaldi and Diana Inkpen. 2023. Text simplification to specific readability levels. *Mathematics*, 11(9):2063.

Sandra Aluísio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications*, Los Angeles, California.

Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the 4th Workshop on Text Simplification, Accessibility, and Readability*, Suzhou, China.

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

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Caroline Gasperin, Lucia Specia, Tiago F. Pereira, and Sandra M. Aluísio. 2009. Learning when to simplify sentences for natural text simplification. In *Proceedings of the Encontro Nacional de Inteligência Artificial (ENIA)*, Bento Gonçalves, Brazil.

Jue Hou, Maximilian W Koppatz, José Maria Hoya Quecedo, Nataliya Stoyanova, Mikhail Kopotev, and Roman Yangarber. 2019. Modeling language learning using specialized Elo ratings. In *BEA: 14^{th} Workshop on Innovative Use of NLP for Building Educational Applications, ACL: 56th annual meeting of Association for Computational Linguistics*.

Inclusion Europe. Information for all: European standards for making information easy to read and understand. https://www.inclusion-europe.eu/easy-to-read-standards-guidelines/. Accessed: 2025-10-12.

Anisia Katinskaia, Anh-Duc Vu, Jue Hou, Ulla Vanhatalo, Yiheng Wu, and Roman Yangarber. 2025. Estimation of text difficulty in the context of language learning. In *BEA: 20th Workshop on Innovative Use of NLP for Building Educational Applications*, Vienna, Austria.

Anisia Katinskaia and Roman Yangarber. 2018. Digital cultural heritage and revitalization of endangered Finno-Ugric languages. In *Proceedings of the 3rd Conference on Digital Humanities in the Nordic Countries*, Helsinki, Finland.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report No. RBR875.

OpenAI. 2024. Gpt-4o system card. https://arxiv.org/abs/2410.21276.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raúl Vázquez, and Sami Virpioja. 2023. Democratizing neural machine translation with opus-mt. *Language Resources and Evaluation*, 58:713–755.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland.

A Prompts for CEFR-based Text Simplification

CEFR	Prompt Description
A1	You must always output a JSON object with an `"SIMPLIFICATION"` key. You are an expert in English language and language teaching. You will be given a text in English. Your task is to read it first and then to provide an adaptation into CEFR level A1. Do not significantly change the meaning of the input text. A1 is the simplest, beginner level with short sentences and easy grammar. Imagine that you are teaching a complete beginner, your adaptation should fit their proficiency level. **This is the text to simplify:** `{text}`
A2	You must always output a JSON object with an `"SIMPLIFICATION"` key. You are an expert in English language and language teaching. You will be given a text in English. Your task is to provide an adaptation into CEFR level A2. Do not significantly change the meaning of the input text. A2 is just above the beginner level and should contain simple grammar and vocabulary. Imagine that you are teaching someone who just started learning the language. **This is the text to simplify:** `{text}`
B1	You must always output a JSON object with an `"SIMPLIFICATION"` key. You are an expert in English language and language teaching. You will be given a text in English. Your task is to provide an adaptation into CEFR level B1. Do not significantly change the meaning of the input text. B1 is an intermediate level. Learners can understand the main points of clear standard input and produce connected text on familiar topics. Adapt the text accordingly. **This is the text to simplify:** `{text}`
B2	You must always output a JSON object with an `"SIMPLIFICATION"` key. You are an expert in English language and language teaching. You will be given a text in English. Your task is to provide an adaptation into CEFR level B2. Do not significantly change the meaning of the input text. B2 corresponds to an upper-intermediate level, allowing complex text understanding and fluent communication. The adapted text should be more advanced than B1 or A2. **This is the text to simplify:** `{text}`
C1	You must always output a JSON object with an `"SIMPLIFICATION"` key. You are an expert in English language and language teaching. You will be given a text in English. Your task is to provide an adaptation into CEFR level C1. Do not significantly change the meaning of the input text. C1 corresponds to an advanced level, capable of producing fluent, well-structured, detailed text with complex grammar and cohesive devices. The adapted text can therefore be more sophisticated and lexically rich. **This is the text to simplify:** `{text}`

Table 5: Prompts used for CEFR-based text simplification to target CEFR levels A1–C1. Each prompt instructs the model to adapt the input text to the linguistic characteristics of the target CEFR level.