

Right at My Level: A Unified Multilingual Framework for Proficiency-Aware Text Simplification

Jinhong Jeong¹ Junghun Park² Youngjae Yu²

¹Yonsei University ²Seoul National University

jjhsnail0822@yonsei.ac.kr

Abstract

Text simplification supports second language (L2) learning by providing comprehensible input, consistent with the Input Hypothesis. However, constructing personalized parallel corpora is costly, while existing large language model (LLM)-based readability control methods rely on pre-labeled sentence corpora and primarily target English. We propose **Re-RIGHT**, a unified reinforcement learning framework for adaptive multilingual text simplification without parallel corpus supervision. We first show that prompting-based lexical simplification at target proficiency levels (CEFR, JLPT, TOPIK, and HSK) performs poorly at easier levels and for non-English languages, even with state-of-the-art LLMs such as GPT-5.2 and Gemini 2.5. To address this, we collect 43K vocabulary-level data across four languages (English, Japanese, Korean, and Chinese) and train a compact 4B policy model using Re-RIGHT, which integrates three reward modules: vocabulary coverage, semantic preservation, and coherence. Compared to the stronger LLM baselines, Re-RIGHT achieves higher lexical coverage at target proficiency levels while maintaining original meaning and fluency.

1 Introduction

Second language (L2) acquisition is most effective when language learners are exposed to sufficient “comprehensible input” at levels between their current proficiency (i) and the marginally higher level ($i + 1$), according to the Input Hypothesis (Krashen, 1981). Linguistic research has further reported that L2 learners need to know 95–98% of the words in a given text to achieve fluent reading comprehension (Hu and Nation, 2000; Schmitt et al., 2011). Accordingly, adequate vocabulary coverage plays a crucial role in providing level-appropriate language input. In real educational environments, however, continuously supplying level-specific reading materials requires considerable effort from profession-

als, making it impractical in terms of cost and time.

Text simplification addresses this problem by controlling lexical complexity to match a target reader’s vocabulary knowledge while preserving original meaning (Al-Thanyyan and Azmi, 2021). Despite substantial advances in large language models (LLMs) and their strong zero-shot capabilities (Wei et al., 2021), they still fall short of reliably generating text that is precisely tailored to a specific proficiency levels (Barayan et al., 2025).

To overcome these limits, recent studies have explored reinforcement learning (RL)-based approaches to improve controllability in text simplification (Zhang and Lapata, 2017; Malik et al., 2024; Li et al., 2025b). However, these methods require pre-labeled, level-specific sentences and are predominantly restricted to English, thereby offering limited applicability in multilingual environments.

In this context, we introduce **Re-RIGHT** (Reinforcement learning for Readability Improvement via Generation from Hard Texts), an adaptive and unified multilingual text simplification framework for vocabulary learning that does not rely on predefined sentence corpora. Our framework supports four languages (English, Japanese, Korean, and Chinese) from different language families and writing systems, yet sharing similar practical learning needs (Yang, 2003).

To systematically control lexical proficiency, we construct a vocabulary level dataset based on the standardized proficiency scales (CEFR, JLPT, TOPIK, and HSK 3.0), and quantitatively define a vocabulary coverage score as the proportion of level-appropriate vocabulary within the content words of a given text. Using this criterion, we first examine the ability of state-of-the-art LLMs like GPT-5.2 and Gemini 2.5 to generate texts that match target proficiency levels. Our findings reveal that LLMs struggle particularly with easy proficiency levels, where this performance gap becomes even larger in non-English languages. We address

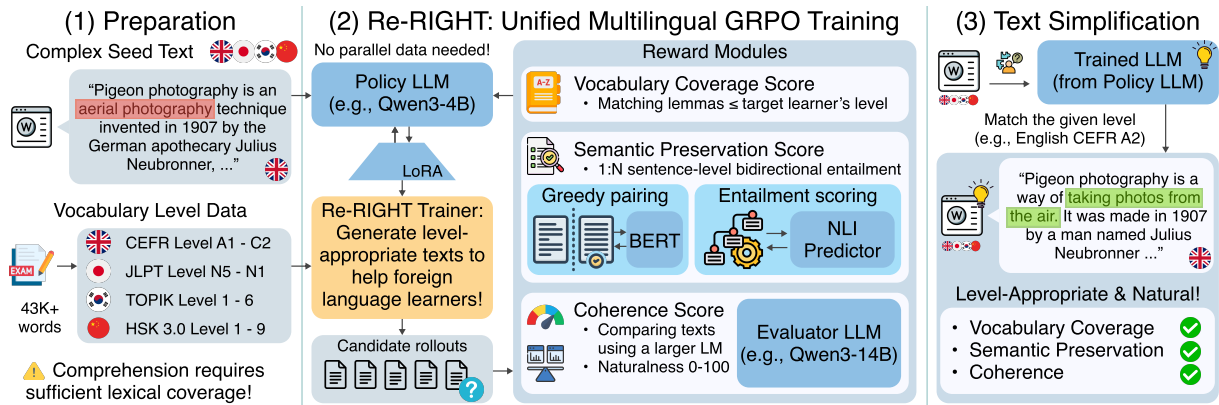


Figure 1: We propose **Re-RIGHT**, a unified multilingual GRPO training framework for text simplification at learners’ proficiency levels (§6) without relying on parallel training corpora, overcoming the limitation of the prompting-based approach (§4) that fails to meet the target vocabulary coverage. (1) At the preparation phase, we collect “Featured Articles” in Wikipedia as a training seed dataset, while constructing 43,786 vocabulary level data from official proficiency standards across four languages. (2) We then train a 4B policy model with three reward modules: vocabulary coverage, semantic preservation, and coherence. (3) The trained policy model enables adaptive multilingual text simplification with higher vocabulary coverage than GPT-5.2 and Gemini 2.5, generating level-appropriate, semantically preserved, and coherent outputs.

this challenge by training a small, unified policy model using Re-RIGHT. Our main contributions are summarized as follows:

1. We propose Re-RIGHT, an adaptive, multilingual text simplification framework that supports language learning through level-appropriate vocabulary without relying on pre-labeled sentences, as illustrated in Figure 1.
2. We provide a comprehensive multilingual benchmark for LLM-based text simplification using 43K+ vocabulary level data, demonstrating that even state-of-the-art LLMs struggle to reliably generate texts at easy proficiency levels and in non-English settings.
3. We introduce three novel reward modules in Re-RIGHT, integrating (1) vocabulary coverage grounded in standardized proficiency scales, (2) semantic preservation based on a 1:N sentence-level bidirectional entailment score, and (3) coherence reward through comparative prompting.

Experimental results demonstrate that a 4B policy model with Re-RIGHT significantly increases the proportion of level-appropriate vocabulary while preserving original meaning and fluency. Our study addresses the fundamental limitation of current LLMs in generating level-appropriate texts, providing a generalizable foundation toward practical L2 learning support across multiple languages.

2 Related Work

2.1 Controlled Vocabulary in Applied Linguistics

The idea of controlling vocabulary for language learning can be traced back to the “Basic English” developed by [Ogden \(1930\)](#), who suggested 850 core words for simplified English. Since then, contemporary applied linguists have consistently shown that achieving sufficient (mainly 95–98%) lexical coverage is essential for effective comprehension of L2 input ([Laufer, 1989](#); [Hu and Nation, 2000](#); [Schmitt et al., 2011](#); [Durbahn et al., 2024](#)), implying that learners’ lexical knowledge plays a pivotal role in providing comprehensible input based on the Input Hypothesis ([Krashen, 1981](#); [Webb, 2021](#)). Studies have also reported strong statistical correlations between L2 learners’ vocabulary knowledge and their reading comprehension performance ([Zhang, 2012](#); [Jeon and Yamashita, 2022](#); [Zhang and Zhang, 2022](#)). These findings have been utilized in the level classification of authoritative language proficiency standards such as CEFR and learner’s vocabulary lists ([Milton, 2010](#); [Brezina and Gablasova, 2013](#); [Capel, 2015](#)), clearly demonstrating the impact of appropriate vocabulary control in language education.

2.2 Text Simplification via Language Models

Researchers have widely investigated language model based text simplification ([North et al., 2025](#)),

constructing readability-controlled datasets at various difficulty levels for English (Kogan et al., 2025) and non-English languages (Naous et al., 2024; Imperial et al., 2025; Anschütz et al., 2025). Several studies have also introduced LLM-based benchmarks (Ryan et al., 2023; Maddela et al., 2023; Kew et al., 2023) or agentic simplification framework (Mo and Hu, 2024). Since applying simple prompt engineering approach has been viewed as insufficient for the task (Barayan et al., 2025), controlled decoding strategies (Kew and Ebling, 2022; Zetsu et al., 2022) and RL-based LLM training (Yanamoto et al., 2022; Malik et al., 2024) have gained prominence.

Specifically, Li et al. (2025b) adopts PPO algorithm (Schulman et al., 2017) to train text simplification models that require data labeled at each CEFR level instead of explicit parallel corpora, with the purpose of generating level-appropriate texts for English as a Second Language (ESL) learners. Our work expands this research by utilizing GRPO algorithm (Shao et al., 2024) with linguistically designed reward modules to eliminate the necessity for predetermined sentence complexity labels, thereby facilitating unified adaptability in multilingual readability control at specific proficiency levels. Furthermore, our framework ensures naturalness in simplified outputs by measuring coherence through prompting multilingual LLMs, unlike Siddharthan (2003), Vásquez-Rodríguez et al. (2023), and Vásquez-Rodríguez et al. (2024), which rely on algorithmic approaches or fine-tuned language models.

3 Datasets

3.1 Multilingual Vocabulary Level Data

We collect vocabulary lists for four languages (English, Japanese, Korean, and Chinese), according to each language’s authoritative proficiency levels. We used Common European Framework of Reference (CEFR) (Council of Europe, 2001) for English, Japanese Language Proficiency Test (JLPT) (The Japan Foundation and Japan Educational Exchanges and Services, 2012) for Japanese, Test of Proficiency in Korean (TOPIK) (National Institute for International Education, 2021) for Korean, and Hanyu Shuiping Kaoshi (HSK) 3.0 (Chinese Test International, 2018)¹ for Chinese as our

¹Levels of HSK7, HSK8, and HSK9 are integrated into the same HSK7-9 level in accordance with the official vocabulary list’s formulation.

vocabulary level criteria.

For vocabulary level data, English data are based on the English Vocabulary Profile (Capel, 2015), while Korean and Chinese data are derived from official vocabulary lists published by national institutions (National Institute of Korean Language, 2017; Ministry of Education of the People’s Republic of China, 2021). Since no official level-specific vocabulary list exists for Japanese, we utilize a widely adopted unofficial online resource.²

After collecting the data, words are lemmatized using the spaCy (Honnibal et al., 2020) library.³ In cases where multiple levels are assigned to the same lemma, the lowest level is used as the basis, resulting in a total of 43,786 lemmas. The summary of the proficiency level data is shown in Table 5 in the Appendix B.

3.2 Wikipedia’s Featured Article Data

In contrast to prior studies, our framework does not require level-annotated training data. Instead, we refine the “Featured Articles” in Wikipedia⁴ as a seed dataset, which contains high-quality and sufficiently complex texts for each language.

We extract a total of 8,057 articles from Wikipedia across four languages. Considering the training memory, we chunk each article into units of up to 512 policy model tokens, while preserving paragraph boundaries to maintain contextual coherence. We then generate multiple proficiency-level variations for each text and uniformly sample across languages. Consequently, this process results in 69,220 data chunks, of which we use 10% as the test set and the remainder as the training set.

3.3 Additional Test Set: Parallel Global Voices

In addition to the Wikipedia Featured Article test set,⁵ we examine the generalizability of our framework by constructing a test set from another domain, namely news data. We use Parallel Global Voices dataset (Prokopicidis et al., 2016), which is preprocessed in the same manner as in §3.2. We uniformly sample 300 level-duplicated news chunks for each language, building a test set with a total of 1,200 documents. Details on the dataset construction in §3 are provided in the Appendix B.

²<https://www.tanos.co.uk/>

³Except Japanese, where lemmatization for standalone words yields low performance, so source lemmas are used.

⁴<https://www.wikipedia.org/>

⁵This data is used as a main training and test set throughout this work unless otherwise specified.

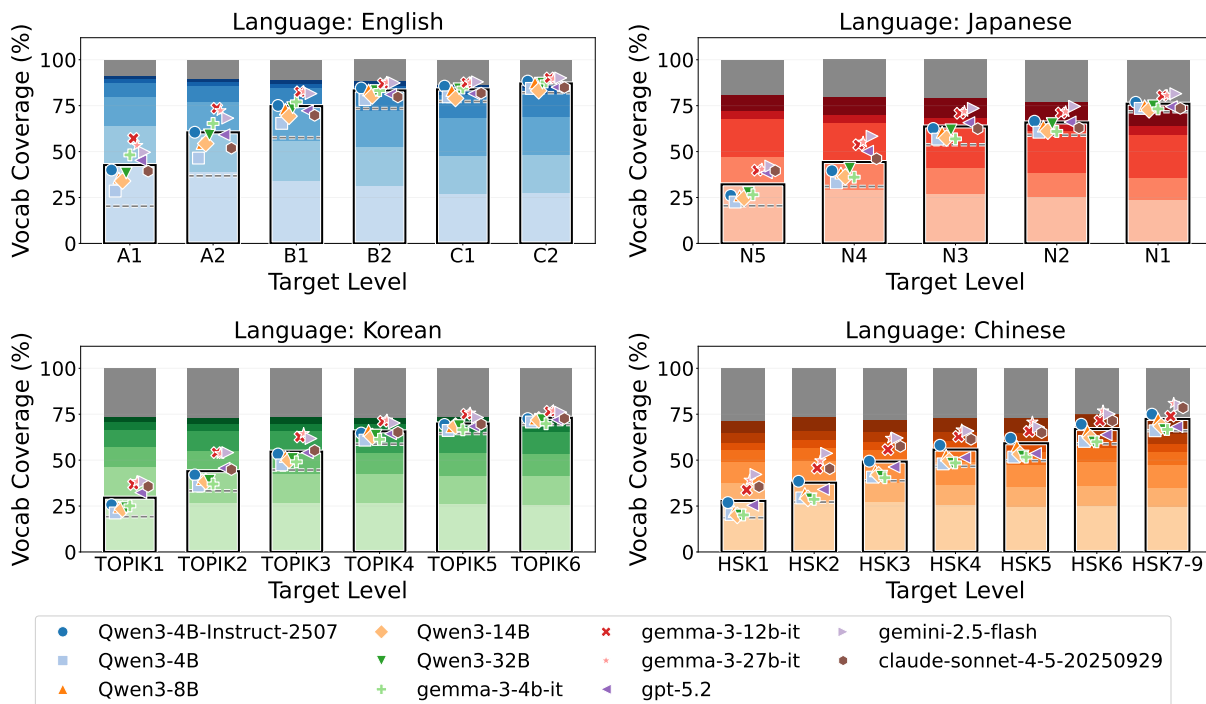


Figure 2: Vocabulary coverage scores (percentage of content words generated at or below the target level) across models and target level variants for each language. The results show that as the target vocabulary level becomes easier, almost all models, including GPT-5.2, Gemini 2.5, and Claude 4.5, fail to provide high vocabulary coverage via the prompting-based approach. The black line box means the average score across models, while the gray dotted line indicates the baseline score of the reference texts. The color depth in each bar plot also represents average vocabulary coverage of each proficiency level from model-generated results for that language (i.e., the lightest color stands for the easiest level), and the gray parts on the top indicate unknown words beyond all levels.

4 Can Current LLMs Generate Level-Appropriate Texts?

We first investigate the performance of existing state-of-the-art LLMs in text simplification at a given learner’s vocabulary level.

4.1 Methodology

We compare vocabulary coverage scores for model-generated candidate texts from various LLMs including GPT-5.2, Gemini-2.5-Flash, Claude Sonnet 4.5, Gemma-3, and Qwen3.⁶ For example, if the target level is CEFR B1, we measure what percentage of the total content words are at or below the B1 level (elaborated in §5.1). The models are zero-shot prompted with the language, target proficiency level, and an original text in the Wikipedia dataset from §3.2. The prompt is as shown in Table 1.

4.2 Results

The experimental models struggle to generate level-appropriate outputs. Figure 2 demonstrates the results for overall vocabulary coverage scores.

⁶In this paper, all models operate in non-reasoning mode.

Text Simplification Test

You are a careful rewrite assistant.
Rewrite the <TEXT> in {language} so that every word, except proper nouns or proper adjectives, is at or below the {level} vocabulary level.
Replace or simplify any other words above {level} level with easier alternatives while preserving the original meaning and coherence.
Do not skip, shorten, or omit any part of the text. Keep sentence count and structure.
Output only the fully converted text with no explanations, instructions, or extra words.

```
<TEXT>
{original_text}
```

Table 1: Text simplification prompt used in §4 and §6. We provide LLMs with the original text along with the target language and proficiency level, requesting text simplification that preserves the original meaning.

Performance Degradation at Easy Levels. For most languages and models, the vocabulary coverage score tends to decrease significantly as the target level becomes easier. This phenomenon implies that even state-of-the-art LLMs lack the ability to generate text that precisely aligns with the learner’s

vocabulary knowledge while keeping original information as required by the prompt in Table 1. Notably, even GPT-5.2 shows low performance at the English CEFR A1 level, only scoring 45.1% of vocabulary coverage.

Performance Discrepancies in Languages.

Non-English languages generally exhibit lower text simplification performance. At the easiest English proficiency level (A1), the models’ average score reaches 42.6%, which is 22.4 points higher than the original text (20.2%). However, at the easiest Korean proficiency level (TOPIK1), it achieves 29.8%, only 10.3 points higher than the baseline (19.5%). The same pattern holds for Japanese (11.8 point gap) and Chinese (9.3 point gap). This tendency is also observed at intermediate proficiency levels, indicating that language models struggle more with adjusting text difficulty for non-English languages.

5 Re-RIGHT: Reward Modules

To overcome the limitation in §4, we introduce Re-RIGHT, a reinforcement learning framework based on the GRPO algorithm (Shao et al., 2024) to improve LLMs’ abilities to perform level-appropriate text simplification. Re-RIGHT integrates three reward modules: vocabulary coverage, semantic preservation, and coherence.

5.1 Vocabulary Coverage

We design a vocabulary coverage score to calculate whether the vocabulary level of a given text is at or below the target level by matching it with the reference vocabulary level data from §3.1.

First, a candidate (simplified) input is lemmatized and processed to remove function words, stopwords, and proper nouns.⁷ Then, the vocabulary coverage score is calculated by matching remaining content lemmas in the input.⁸ Namely, we define the vocabulary coverage score as follows:

$$\text{SCORE}_{\text{vocab}} = \frac{|\{w_i \in M(C) \mid \ell(w_i) \leq \ell_t\}|}{|M(C)|}, \quad (1)$$

where $M(C)$ is a multiset of content lemmas in a candidate input C , and $\ell(w_i)$ indicates the proficiency level of a lemma w_i , given a target level

⁷Before the English vocabulary matching, the module conducts phrase-level matching in descending order of length to mitigate the word-level breakdown of phrasal expressions.

⁸Chinese words are often compound expressions from individual morphological characters. Therefore, if matching fails, they are once more decomposed into individual characters.

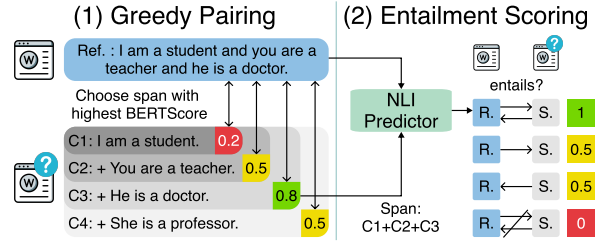


Figure 3: Entailment scoring process for the semantic preservation reward. We introduce a two-phase approach with (1) greedy pairing using BERTScore and (2) entailment scoring via an NLI predictor.

ℓ_t . Unknown words not in the reference vocabulary level data are considered as beyond the target level.

To assess improvements from the original text, we then define the final reward $r_{\text{vocab}} = \text{SCORE}_{\text{vocab_rollout}} - \text{SCORE}_{\text{vocab_original}}$.⁹

5.2 Semantic Preservation

We introduce 1 : N sentence-level bidirectional entailment score for assessing semantic preservation. Through this method, a complex reference sentence can be split into up to N simple corresponding sentences while preserving the meaning.

To achieve this, we adopt a two-phase approach as shown in Figure 3. First, for each reference sentence, we greedily combine 1 to n candidate sentences,¹⁰ measuring BERTScore (Zhang et al., 2020) between the reference and candidate sentence spanning pair. We then select the pair with the highest embedding similarity as an aligned result pair. Next, we use a small Natural Language Inference (NLI) model¹¹ to bidirectionally examine whether the aligned pairs of a reference sentence A_i and a candidate sentence spanning B_i are in an entailment relationship.

Then, $\text{score}_{\text{entailment}}$ is the average of all scores p_i of the aligned pairs, where p_i is defined as:

$$p_i = \begin{cases} 1.0 & \text{if } A_i \Rightarrow B_i \wedge B_i \Rightarrow A_i \\ 0.5 & \text{if } A_i \Rightarrow B_i \oplus B_i \Rightarrow A_i \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Finally, the semantic preservation reward

$$r_{\text{sem_pres}} = \text{SCORE}_{\text{entailment}}$$

⁹Since the reward is calculated using only content words, the vocabulary coverage score may not strictly adhere to the linguistically targeted vocabulary coverages. Nevertheless, the score remains useful due to the nature of the GRPO performing relative comparisons.

¹⁰In this paper, the maximum sentence spanning $n = 4$.

¹¹We use MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 model to determine the sentence entailment.

5.3 Coherence

We employ the LLM-as-a-judge (Li et al., 2025a) method to measure coherence score by comparatively prompting an evaluator model to assess the naturalness of the generated texts. The evaluator model is assumed to have more parameters than a policy model.

The evaluator model is given a pair of reference and candidate text, with a prompt that requests quality evaluation of the candidate text on a scale from 0 to 100, compared to the reference text. In the prompt, the module requires the evaluator to assign text quality to a predefined score range, heavily penalizing repetitive template phrasing and unnatural sentence patterns that may occur in case of reward hacking. After the evaluation, the coherence score is normalized to a float number between 0 and 1. For the detailed prompt, refer to the Appendix G.

During the training, the coherence score is passed through a quadratic transformation to impose a greater penalty on low-quality rollouts. In addition, to prevent the policy model from generating outputs copying the reference text, the coherence module slightly penalizes the score using the Jaccard similarity $J(\cdot)$. The coherence reward $r_{coherence}$ therefore defined as:

$$r_{coherence} = \max\left(1 - \left(\frac{1 - \text{score}_{coherence}}{1 - \alpha}\right)^2, 0\right) - \beta J(S_t(A), S_t(B)), \quad (3)$$

where α is a quality boundary constant,¹² β is a penalty coefficient,¹³ and $S_t(A)$ and $S_t(B)$ denote the sets of lemmas in a reference A and a candidate rollout B that exceed the target level t , respectively.

6 Re-RIGHT: Experiments

6.1 Baselines

We set baselines to compare relative performances of Re-RIGHT. Since the use of multilingual proficiency-labeled parallel corpora falls outside our scope, we exclude methods requiring such data.

Reference Text. As the most basic score, the original text is used directly as the result text.

Prompting Approach. We measure the evaluation scores for a base policy model (Qwen3-4B-Instruct-2507) and state-of-the-art models (GPT-

¹² $\alpha = 0.6$, which corresponds to the upper boundary of the prompt-defined “machine-generated” quality score.

¹³ $\beta = 0.05$, which empirically yields stable training.

5.2 and Gemini-2.5-Flash) performing zero-shot text simplification using the prompt in Table 1, as conducted in §4.

Constrained Decoding. We also measure scores using a constrained generation method, FUDGE (Yang and Klein, 2021), computing scores for top-100 candidate tokens each step. We use a FUDGE discriminator that predicts the probability of meeting the constraint at the end of the generation. Since a multilingual text level classifier is not available, we instead apply a rule-based approach: at each step, we compute the vocabulary coverage score of the current candidate token and last five tokens.¹⁴

6.2 Methodology

In the training stage, we adopt Qwen3-4B-Instruct-2507 as a policy model, and Qwen3-14B as an evaluator model. We also apply LoRA (Hu et al., 2021) for parameter-efficient training. It should be noted that a single policy model is trained for all languages and proficiency levels.

We use the Wikipedia dataset (§3.2) as a training seed data. For each prompt, eight candidate responses are sampled.¹⁵ Then, the GRPO objective is optimized using the three reward modules (vocabulary coverage, semantic preservation, and coherence). The final reward is computed as a weighted sum, enabling the level-appropriate simplification policy training that jointly satisfies the multifactorial goal of the experiment.

In the evaluation stage, we employ gemma-3-27b-it and Qwen3-32B as evaluator models and report the average coherence score of both models to avoid self-evaluation.

6.3 Results

Table 2 shows overall results, and Table 3 demonstrates simplified text examples in each language. Since easier proficiency levels lead to particularly low scores as observed in §4, we additionally focus on average scores at “easier” proficiency levels, defined as the lowest two levels for each language.¹⁶

Vocabulary Coverage Improvements. Re-RIGHT provides substantially higher vocabulary

¹⁴The last 5 tokens are considered since the spaCy pipeline and the simplification policy model use different tokenizers.

¹⁵Due to the computational resource constraints, prompts and outputs are limited to a maximum of 512 tokens.

¹⁶English: A1 and A2, Japanese: N5 and N4, Korean: TOPIK1 and TOPIK2, and Chinese: HSK1 and HSK2.

Lang.	Method	Vocabulary Coverage		Semantic Pres.		Coherence	
		Total (Std.)	Easy (Std.)	Total	Easy	Total	Easy
EN	Reference	58.3 (24.2)	28.9 (12.6)	–	–	–	–
	Base (Qwen3-4B)	72.6 (19.2)	50.4 (14.4)	75.3	72.3	86.9	83.7
	FUDGE (Qwen3-4B)	74.2 (18.4)	53.4 (14.4)	75.0	71.8	86.7	83.1
	GPT-5.2	71.0 (16.8)	52.4 (12.2)	76.1	70.4	84.6	75.8
	Gemini 2.5	77.7 (16.3)	59.1 (13.5)	72.0	69.2	82.8	78.7
	Re-RIGHT	81.6 (13.6)	66.9 (12.3)	80.8	78.1	82.9	77.0
JA	Reference	46.6 (20.3)	26.4 (9.1)	–	–	–	–
	Base (Qwen3-4B)	53.9 (20.4)	33.7 (10.8)	65.4	64.9	88.7	88.2
	FUDGE (Qwen3-4B)	54.3 (20.3)	34.2 (10.9)	64.5	65.0	88.7	88.2
	GPT-5.2	58.8 (15.7)	45.1 (10.9)	59.7	57.5	84.7	79.9
	Gemini 2.5	65.8 (16.5)	51.1 (12.3)	61.7	62.5	85.2	82.2
	Re-RIGHT	76.0 (15.3)	60.4 (10.7)	80.6	80.1	83.1	79.6
KO	Reference	48.8 (19.9)	26.3 (11.2)	–	–	–	–
	Base (Qwen3-4B)	55.2 (19.3)	33.9 (13.0)	75.7	74.3	89.1	89.0
	FUDGE (Qwen3-4B)	56.0 (19.2)	34.9 (13.2)	75.3	73.4	89.2	88.8
	GPT-5.2	56.3 (17.4)	39.0 (12.6)	65.6	64.1	88.5	85.4
	Gemini 2.5	62.4 (16.4)	46.0 (13.9)	69.3	69.3	87.0	85.0
	Re-RIGHT	70.4 (15.7)	52.9 (12.6)	87.1	86.7	84.4	81.3
ZH	Reference	44.8 (20.3)	23.9 (9.0)	–	–	–	–
	Base (Qwen3-4B)	55.2 (19.4)	33.3 (10.4)	62.0	61.2	90.2	89.7
	FUDGE (Qwen3-4B)	55.4 (19.4)	33.5 (10.5)	61.7	61.7	90.2	89.8
	GPT-5.2	50.3 (19.8)	31.0 (11.0)	57.7	57.6	88.4	86.0
	Gemini 2.5	64.4 (15.4)	48.6 (11.9)	66.7	68.1	85.3	81.2
	Re-RIGHT	80.2 (12.2)	66.1 (9.7)	76.6	77.1	83.7	81.0

Table 2: Evaluation results across four languages, reporting averaged vocabulary coverage, semantic preservation, and coherence scores for overall proficiency levels (Total) and easier-level subsets (Easy: average over the two lowest levels for each language). Re-RIGHT outperforms baselines in vocabulary coverage while preserving meaning and fluency, only using a 4B policy model. In particular, Re-RIGHT improves text simplification performance at easier levels, while existing baselines (including GPT-5.2 and Gemini 2.5) fail to achieve high vocabulary coverage.

coverage than the baselines, including at easier proficiency levels. For instance, the vocabulary coverage score of the 4B policy model with Re-RIGHT for English simplification achieves 81.6% across all levels and 66.9% at easier levels, surpassing Gemini-2.5-Flash (the highest-performing LLM in this work) by 3.9 points and 7.8 points, respectively. Simultaneously, the semantic preservation score improves by 8.8 points to 80.8, while the coherence score increases by 0.1 points to 82.9. Since lower vocabulary coverage inevitably leads to higher coherence in outputs that closely resemble the reference text, achieving high vocabulary coverage while minimizing the coherence reduction is meaningful.

Effectiveness in Non-English Languages. Re-RIGHT also significantly improves vocabulary coverage performance for Japanese, Korean, and Chinese. The average vocabulary coverage improvement over the reference texts for non-English languages is 28.8 points, compared to an 8.5-point increase for the FUDGE baseline and 17.5 points for Gemini-2.5-Flash.

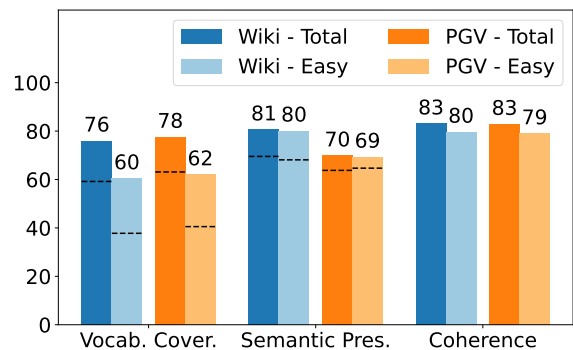


Figure 4: Average evaluation scores across languages on out-of-domain data (PGV), compared to the Wikipedia test data. We show that Re-RIGHT endows the model with general simplification abilities from complex seed texts of an arbitrary domain. The black dashed line represents the performance of the 4B base model.

6.4 Data and Model Generalizability

Out-of-Domain Test Set. We additionally evaluate our trained policy model on an out-of-domain news dataset from Parallel Global Voices (PGV; elaborated in §3.3) to demonstrate the general-

Lang.	Level	Text
EN	Reference	Pigeon photography is an aerial photography technique invented in 1907 by the German apothecary Julius Neubronner, who also used pigeons to deliver medications. ...
	CEFR A2	Pigeon photography is a way of taking photos from the air. It was made in 1907 by a man named Julius Neubronner from Germany. He also used pigeons to carry medicine. ...
	CEFR B2	Pigeon photography is a way of taking photos from the air that was first created in 1907 by a German man named Julius Neubronner. He also used pigeons to carry medicine. ...
JA	Reference	... ブリオンは「シューマンにとって、限界があり、厳密さを欠く文章表現よりも、... 緻密な言葉を提供した」と述べている。 (Translation: ... Brion states, “For Schumann, it offered precise language rather than the limited and imprecise nature of written expression ...”)
	JLPT N4	... ブリオンさんはこう言っていた。「シューマンにとって、文章というのは限界があって、とても正確でない場合が多い。... よく考えられた言葉をくれる。」
	JLPT N2	... ブリオンさんはこう言っている。「シューマンにとって、文章というのは限界があって、とても厳密でないものだった。... 正確な言葉を提供してくれた。」
KO	Reference	... 당시 자유주의자들은 독일인들의 열광적이고 광범위한 지지를 얻으며 세를 넓혀 나가고 있었다. 자유주의자들은 독일의 통일을 ... (Translation: ... At the time, the liberals were expanding their influence, gaining enthusiastic and widespread support from the German people. The liberals sought German unification ...)
	TOPIK2	... 그때 자유주의 사람들은 독일 사람들의 마음을 많이 얻고, 그 마음을 바탕으로 더 큰 힘을 얻어 갔다. 자유주의 사람들은 독일이 하나가 되는 것을 ...
	TOPIK4	... 그 시기 자유주의 그룹은 독일 사람들의 매우 강한 관심과 넓은 지지를 받으며 점점 더 큰 영향력을 얻어 갔다. 자유주의 그룹은 독일 전체를 하나의 나라로 만드는 ...
ZH	Reference	... 在这种背景下，库克在旅程途中也不时把新发现的地方宣告为英国领土，因此，如果说他的航海旅程完全不具“掠夺”性质，也不是准确的说法。 (Translation: ... Against this backdrop, Cook periodically declared newly discovered territories as British possessions during his voyages. Therefore, it would be inaccurate to claim that his expeditions were entirely devoid of a “predatory” nature.)
	HSK2	... 在这种情况下，库克在去的路上，也常常把新发现的地方说成是英国的地盘。所以，说他整个出海过程一点都没有“拿东西”的意思，其实也不对。
	HSK4	... 在这种情况下，库克在旅行过程中，也常常把新发现的地方说成是英国的地盘。所以，如果有人说库克的旅行一点都没有“想占便宜”的意思，那其实也不太对。

Table 3: Examples of proficiency-controlled simplification results across languages. Each block shows the reference and two simplified outputs aligned to different target levels. Without pre-labeled parallel corpora, Re-RIGHT enables the trained 4B policy model to generate level-appropriate sentences, preserving information and coherence.

izability of Re-RIGHT. Figure 4 shows that our framework trained on the Wikipedia dataset effectively operates on PGV test data, implying that sufficiently complex datasets from arbitrary domains can be used as training seed data.

Model Generalization. We also demonstrate that Re-RIGHT performs effectively on gemma-3-4b-it. Table 4 demonstrates that our framework can be easily integrated into other model families, maintaining the same level of high performance.

7 Conclusion

We present Re-RIGHT, an adaptive, unified, and multilingual text simplification framework tailored to second language learners’ proficiency levels, without relying on level-annotated parallel corpora. Our experiments show that even state-of-the-art LLMs have limited abilities to generate level-appropriate outputs, particularly at easy levels and in non-English languages. In contrast, Re-RIGHT

La.	Method	Voc.	Sem.	Coh.
EN	Base (Qw3 4B)	72.6	75.3	86.9
	Base (Ge3 4B)	74.5	69.0	84.7
	Re-RIGHT (Qw3 4B)	81.6	80.8	82.9
	Re-RIGHT (Ge3 4B)	81.5	79.9	80.8
JA	Base (Qw3 4B)	53.9	65.4	88.7
	Base (Ge3 4B)	50.2	61.6	90.0
	Re-RIGHT (Qw3 4B)	76.0	80.6	83.1
	Re-RIGHT (Ge3 4B)	72.5	76.3	79.5
KO	Base (Qw3 4B)	55.2	62.0	90.2
	Base (Ge3 4B)	51.9	69.7	90.1
	Re-RIGHT (Qw3 4B)	70.4	87.1	84.4
	Re-RIGHT (Ge3 4B)	72.5	84.5	82.0
ZH	Base (Qw3 4B)	55.2	62.0	90.2
	Base (Ge3 4B)	46.3	58.8	92.1
	Re-RIGHT (Qw3 4B)	80.2	76.6	83.7
	Re-RIGHT (Ge3 4B)	75.1	73.6	80.3

Table 4: Performances of Re-RIGHT on Gemma 3 4B, compared to the baselines with Qwen3 4B. We show that Re-RIGHT can be generalized to other models.

substantially improves lexical level controllability while preserving original meaning and fluency. We expect that Re-RIGHT can serve as a foundation for future research toward agentic, adaptive language tutoring systems.

Limitations

Scoring Criteria. Our vocabulary coverage module requires word level data for each language, where unknown words are treated as words exceeding all proficiency levels. Since it is impossible for researchers to collect all vocabulary used within a language, the proposed scores should not be regarded as absolute values. Nevertheless, the standardized language test criteria require learning a limited vocabulary list for each level, so our methodology can be considered meaningful for measuring vocabulary proficiency levels targeted at language learners.

Granularity of Complexity. Additionally, our current approach mainly considers lexical complexity. However, other variables such as syntactic complexity may hold importance in text simplification as well. In this research, we limit the scope of our study to lexical complexity, instead of measuring the hard-to-define syntactic complexity of sentences belonging to different language families.

Model Size Limits. Due to the limitation of training resources, we conduct experiments with relatively small 4B-size policy models and a 14B-size evaluator model. This may lead to the generation of slightly unnatural sentences. Moreover, the relatively small NLI predictor we used for the semantic preservation reward may not be able to perfectly determine multilingual entailment relationships. However, the demonstrated generalizability across model families implies that our framework can be readily scaled up to larger models that exhibit more natural multilingual performance.

Needs for Assessment in Real-World Educational Environments. Although our framework empirically generates outputs preserving original meaning and naturalness, the evaluations for semantic preservation and coherence are largely substantiated via automated scoring from language models. Therefore, determining the actual usefulness and effects of our framework for language learners in real-world educational settings may require long-term human evaluation experiments.

Ethical Considerations

Potential Pedagogical Misuse. Although level-appropriate language input can accelerate the second language acquisition process, instructors should not rely solely on automated simplification systems when providing educational materials to language learners. Given the complex nature of language learning, the use of LLM-based readability control frameworks should always be accompanied by appropriate human-in-the-loop review.

Learner Diversity Consideration. Language learners exhibit substantial variation in vocabulary knowledge when acquiring foreign languages. While our framework basically assumes learners' lexical proficiency to be aligned with standardized proficiency guidelines, actual vocabulary knowledge may differ depending on learning objectives, educational environments, cultural background, or individual aptitude. Such variation may lead to inappropriate or indiscriminate application of LLM-based frameworks.

Acknowledgments

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. RS-2024-00354218, No. RS-2024-00353125), and the Technology Innovation Program(RS-2025-25456760, Development of a humanoid robot specialized in chemical processes based on AI foundation model) funded By the Ministry of Trade, Industry and Resources(MOTIR, Korea). We express special thanks to KAIT GPU project. The ICT at Seoul National University provides research facilities for this study.

References

- Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. [Automated text simplification: A survey](#). *ACM Comput. Surv.*, 54(2).
- Miriam Anschutz, Thanh Mai Pham, Eslam Nasrallah, Maximilian Müller, Cristian-George Craciun, and Georg Groh. 2025. [German4all – a dataset and model for readability-controlled paraphrasing in german](#). *Preprint*, arXiv:2508.17973.

- Anthropic. 2025. [Introducing claude sonnet 4.5](#). Accessed: 2026-01-02.
- 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.
- Vaclav Brezina and Dana Gablasova. 2013. [Is there a core general vocabulary? introducing the new general service list](#). *Applied Linguistics*, 36(1):1–22.
- Annette Capel. 2015. The english vocabulary profile. *English profile in practice*, 5(1):9–27.
- Chinese Test International. 2018. Chinese test service website. <https://www.chinesetest.cn/>. Accessed: 2025-12-03.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.
- Marion Durbahn, Michael Rodgers, Marijana Macis, and Elke Peters. 2024. [Lexical coverage in l1 and l2 viewing comprehension](#). *Studies in Second Language Acquisition*, 46(4):1045–1068.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- Holger Hopp. 2018. The bilingual mental lexicon in l2 sentence processing. *Second language*, 17(1):5–27.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- M Hu and Paul Nation. 2000. [Unknown vocabulary density and reading comprehension](#).
- Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Munoz Sanchez, 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](#). *Preprint*, arXiv:2506.01419.
- Eun Hee Jeon and Junko Yamashita. 2022. L2 reading comprehension and its correlates. *Understanding L2 proficiency: Theoretical and meta-analytic investigations*, pages 29–86.
- 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.
- Tannon Kew and Sarah Ebling. 2022. [Target-level sentence simplification as controlled paraphrasing](#). In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 28–42, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.
- Jung-seob Kim. 2017. [A study on the application of international standard korean language curriculum](#).
- David Kogan, Max Schumacher, Sam Nguyen, Masanori Suzuki, Melissa Smith, Chloe Sophia Bellows, and Jared Bernstein. 2025. [Ace-cefr – a dataset for automated evaluation of the linguistic difficulty of conversational texts for llm applications](#). *Preprint*, arXiv:2506.14046.
- Stephen D Krashen. 1981. Second language acquisition and second language learning.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Batia Laufer. 1989. What percentage of text-lexis is essential for comprehension. *Special language: From humans thinking to thinking machines*, 316323.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025a. [From generation to judgment: Opportunities and challenges of LLM-as-a-judge](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

- Guanlin Li, Yuki Arase, and Noel Crespi. 2025b. **Aligning sentence simplification with ESL learner’s proficiency for language acquisition**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 492–507, Albuquerque, New Mexico. Association for Computational Linguistics.
- Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. **LENS: A learnable evaluation metric for text simplification**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. 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.
- James Milton. 2010. **The development of vocabulary breadth across the cefr levels**.
- Ministry of Education of the People’s Republic of China. 2021. Moe china notice (2021-03-29). http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202103/t20210329_523304.html. Accessed: 2025-12-03.
- Kaijie Mo and Renfen Hu. 2024. **ExpertEase: A multi-agent framework for grade-specific document simplification with large language models**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9080–9099, Miami, Florida, USA. 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.
- National Institute for International Education. 2021. Test of proficiency in korean (topik) official website. <https://www.topik.go.kr/>. Accessed: 2025-12-03.
- National Institute of Korean Language. 2017. 2017 study on the application of the internationally-used standard korean curriculum (phase 4). https://www.korean.go.kr/front/reportData/reportDataView.do?mn_id=207&report_seq=932. Accessed: 2025-12-03.
- Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2025. Deep learning approaches to lexical simplification: A survey. *Journal of Intelligent Information Systems*, 63(1):111–134.
- Charles Kay Ogden. 1930. **Basic english: A general introduction with rules and grammar**.
- OpenAI. 2025. **Introducing gpt-5.2**. Accessed: 2026-01-02.
- Prokopis Prokopidis, Vassilis Papavassiliou, and Stelios Piperidis. 2016. **Parallel global voices: a collection of multilingual corpora with citizen media stories**. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).
- Qwen Team. 2025. **Qwen3 technical report**. *Preprint*, arXiv:2505.09388.
- Ranks NL. 2025. **Japanese stopwords list**. <https://www.ranks.nl/stopwords/japanese>. Accessed: 2025-12-30.
- Michael J Ryan, Tarek Naous, and Wei Xu. 2023. **Revisiting non-English text simplification: A unified multilingual benchmark**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.
- Norbert Schmitt, Xiangying Jiang, and William Grabe. 2011. The percentage of words known in a text and reading comprehension. *The modern language journal*, 95(1):26–43.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. **Proximal policy optimization algorithms**. *Preprint*, arXiv:1707.06347.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. **Deepseekmath: Pushing the limits of mathematical reasoning in open language models**. *Preprint*, arXiv:2402.03300.
- Advait Siddharthan. 2003. **Preserving discourse structure when simplifying text**. In *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003) at EACL 2003*, Budapest, Hungary. Association for Computational Linguistics.
- The Japan Foundation and Japan Educational Exchanges and Services. 2012. Japanese-language proficiency test (jlpt) official website. <https://www.jlpt.jp/>. Accessed: 2025-12-03.
- Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2023. **Document-level text simplification with coherence evaluation**. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 85–101, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Galouédec. 2020. Trl: Transformer reinforcement learning. <https://github.com/huggingface/trl>.
- Laura Vásquez-Rodríguez, Nhung T. H. Nguyen, Piotr Przybyła, Matthew Shardlow, and Sophia Ananiadou. 2024. [Simple is not enough: Document-level text simplification using readability and coherence](#). *Preprint*, arXiv:2412.18655.
- Stuart Webb. 2021. Research investigating lexical coverage and lexical profiling: What we know, what we don't know, and what needs to be examined.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. [Optimizing statistical machine translation for text simplification](#). *Transactions of the Association for Computational Linguistics*, 4:401–415.
- Yanxiang Xu and Tiejian Luo. 2011. Measuring article quality in wikipedia: Lexical clue model. In *2011 3rd Symposium on Web Society*, pages 141–146. IEEE.
- 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.
- Jean Sook Ryu Yang. 2003. Motivational orientations and selected learner variables of east asian language learners in the united states. *Foreign Language Annals*, 36(1):44–56.
- Kevin Yang and Dan Klein. 2021. [FUDGE: Controlled text generation with future discriminators](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. 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.
- Dongbo Zhang. 2012. Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *The modern language journal*, 96(4):558–575.
- Songshan Zhang and Xian Zhang. 2022. The relationship between vocabulary knowledge and l2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, 26(4):696–725.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). *Preprint*, arXiv:1904.09675.
- Xingxing Zhang and Mirella Lapata. 2017. [Sentence simplification with deep reinforcement learning](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

A Experimental Settings

A.1 Training Hardware and Software

- Hardware: up to four NVIDIA RTX A6000 GPUs. For proprietary models, we use API calls.
- Software: TRL (von Werra et al., 2020) library for the reinforcement learning experiments. Evaluation (including evaluator LLM inferencing during the training process) is conducted using vLLM (Kwon et al., 2023) library.
- Reproducibility: When needed, the random seed is fixed to 42, and the temperature is set to 0.

A.2 Full List of Language Models Used

- Qwen3-4B-Instruct-2507 (Qwen Team, 2025)
- Qwen3-4B (Qwen Team, 2025)
- Qwen3-8B (Qwen Team, 2025)
- Qwen3-14B (Qwen Team, 2025)
- Qwen3-32B (Qwen Team, 2025)
- gemma-3-4b-it (Gemma Team et al., 2025)
- gemma-3-12b-it (Gemma Team et al., 2025)
- gemma-3-27b-it (Gemma Team et al., 2025)
- gpt-5.2 (OpenAI, 2025)
- gemini-2.5-flash (Comanici et al., 2025)
- claude-sonnet-4-5-20250929 (Anthropic, 2025)

A.3 Detailed Hyperparameters for Re-RIGHT Experiment

- LoRA configuration.
 - Rank: 16
 - LoRA alpha: 32
 - Target modules: All attention and MLP layers.
- Training configuration.
 - Learning rate: $3e-5$
 - Optimizer: adamw_8bit
 - Scheduler: linear
 - Number of sample generations: 8
 - KL penalty beta: 0.002
 - Batch size per device: 4
 - Gradient accumulation steps: 4
 - GPU hardware: three NVIDIA RTX A6000 GPUs for GRPO training and one NVIDIA RTX A6000 GPU for the evaluator LLM.
- Reward module weights.
 - Vocabulary coverage: 2.0
 - Semantic preservation: 1.0
 - Coherence: 1.0

The weight settings are empirically chosen. We observe that assigning a relatively high weight to vocabulary coverage yields stable training.

Due to the training resource constraint, models are trained for 0.25 epochs over the training dataset. Nevertheless, the reward curves have nearly plateaued.

A.4 Baseline Settings

- FUDGE: top-100 tokens are considered each generation step. At each time step, vocab coverage score of the candidate token and 5 previous tokens are calculated. The implementation uses the custom logit processor feature of the vLLM library.

B Dataset Details

B.1 Vocabulary Level Data

Table 5 shows the summary of proficiency level data across English, Japanese, Korean, and Chinese.

When processing Japanese vocabulary data, the Kanji, Hiragana, and Katakana representations of

a single word are treated as distinct lemmas sharing the same proficiency level. This encourages policy models to replace relatively difficult Kanji (Sino-Japanese) terms with easier Hiragana-based expressions.

B.2 Wikipedia Dataset

Wikipedia’s “Featured Articles” are the best documents of Wikipedia manually selected by editors, which can be regarded as a well-structured seed data that covers professional subjects with high-level vocabularies (Xu and Luo, 2011). As described in Table 6, we extract the Featured Articles from English (6,793 articles), Japanese (100 articles), Korean (140 articles), and Chinese (1,023 articles) Wikipedia, resulting in a total of 8,057 documents. As described in §3.2, we preprocess the data to be plain texts and chunk up to 512 tokens while maintaining paragraph information. The paragraph boundaries are determined by two new-line characters. We also remove paragraphs that have under 20 tokens or reference strings, such as ‘参考文献’ or ‘참고 문헌’. After preprocessing, we duplicate every chunk by the number of target proficiency levels for each language.¹⁷ We then uniformly sampled chunks from each language, making the final training and test dataset. In this way, our resulting dataset is not parallel, nor manually pre-annotated with proficiency levels, and can be substituted with any random dataset that contains sufficiently complex texts. Examples of each chunk are shown in Table 11.

B.3 Parallel Global Voices Dataset

We use Parallel Global Voices dataset (Prokopicidis et al., 2016) as an out-of-distribution test set. Multilingual corpora of English, Japanese, Korean, and Chinese are used for preprocessing, which is described in Table 6. The body of each document consists of several xml paragraphs. Since each paragraph of the data usually consists of only a few sentences, we concatenate all paragraphs with a blank in between. We truncate later paragraphs to maintain length within 512 tokens, retaining only 1 chunk per document. Paragraphs containing ‘crawlinfo’ tag (usually in another language), or tag ‘type’ and value ‘title’, ‘contributor’, ‘notes’, ‘tweet-info’, or ‘caption’ are removed. We only retain documents containing at least 300 tokens. Like

¹⁷Each data chunk is replicated across all target proficiency levels, with only the proficiency level and language condition prompts varied; the source text itself remains unchanged.

Lang.	Criterion	Level	# Lemma
EN	CEFR	A1, A2, B1, B2, C1, C2	9413
JA	JLPT	N5, N4, N3, N2, N1	14146
KO	TOPIK	Level 1, 2, 3, 4, 5, 6	9357
ZH	HSK 3.0	Level 1, 2, 3, 4, 5, 6, 7-9	10870
Total			43786

Table 5: Overall summary for our vocabulary level data. We collect proficiency level information of lemmas across four languages (English, Japanese, Korean, and Chinese) based on authoritative examinational standards. Vocabulary coverage module utilizes this data to calculate the proficiency level adequacy of an input.

Dataset	Preprocessing
Wikipedia	<ul style="list-style-type: none"> • Document Filtering & Chunking <ul style="list-style-type: none"> – Extract ‘Featured articles’ documents – Detect paragraphs by double newline characters – Remove reference paragraphs or paragraphs with < 20 tokens – Ensure each chunk is ≤ 512 tokens – Maintain paragraph boundaries • Sampling <ul style="list-style-type: none"> – Duplicate each chunk by number of target levels – Uniformly sample from each language
PGV	<ul style="list-style-type: none"> • Document Filtering & Chunking <ul style="list-style-type: none"> – Remove paragraphs with key ‘crawlinfo’ – Remove paragraphs with key ‘type’ and value ‘title’, ‘caption’, ‘contributor’, ‘notes’, or ‘tweet-info’ – Concatenate all paragraphs with blank in between – Remove documents with < 300 tokens – Truncate to retain ≤ 512 tokens only • Sampling <ul style="list-style-type: none"> – Duplicate each chunk by number of target levels – Uniformly sample from each language

Table 6: Checklist for dataset preprocessing.

the Wikipedia dataset, we duplicate every document by target proficiency levels for each language. We then uniformly sample 300 documents for each language, resulting in 1,200 documents. Examples of each chunk are shown in Table 12.

C Reward Module Details

C.1 Vocabulary Coverage

We remove stopwords and proper nouns from the total vocabulary count when calculating the vocabulary coverage score. During the scoring process, spaCy performs PoS tagging and removes lemmas that are labeled as “SYM”, “PUNCT”, “SPACE”, “X”, and “PROP”. Additionally, lemmas categorized as “ADP”, “AUX”, “PART”, “SCONJ”, “CCONJ”, “DET”, and “PRON” are deleted from the candidate text while processing agglutinative

languages, Japanese and Korean. The vocabulary coverage module also matches every lemma with pre-defined stopword lists in each language and excludes them from the text. For English and Chinese, we use pre-defined stopwords in the spaCy library. For Japanese, where no official stopword list exists, we crawl the stopword data from Ranks NL (2025). For Korean, we utilize an educational stopword list from National Institute of Korean Language (Kim, 2017).

C.2 Semantic Preservation

In our 1 : N sentence-level bidirectional entailment scoring for assessing semantic preservation, we first split all sentences in a reference text and a candidate text. As described in §5.2, we calculate the entailment score by a two-phase approach: greedy pairing and entailment scoring. Accordingly, the

module needs to align the sentence counts to effectively compare the reference and candidate text. When the candidate sentence count is larger than the reference sentence count, the module keeps proceeding with the two phases. If the candidate sentence count is equal to the reference sentence count, there is no need for calculating BERTScore for greedy pairing, so the module proceeds directly to the second phase. Otherwise, it means that the candidate text is so concise that the text could not preserve the original sentence structure, and the module produces no reward for that candidate.

C.3 Coherence

We observe that applying a linear penalty to the coherence reward consistently encourages the trained policy model to generate relatively unnatural outputs. To address this issue, we adopt quadratic transformation that attenuates coherence score reduction during training.

D Ablation Study

We further analyze the effect of each reward module. Effective text simplification requires achieving high vocabulary coverage while minimizing degradation in semantic preservation and coherence. We compare the average evaluation scores of languages and proficiency levels with the following settings:

- Vanilla model: Base policy model (Qwen3-4B-Instruct-2507) without training.
- Full framework: Vocabulary coverage, semantic preservation, and coherence reward modules.
- 1 : 1 sentence-level entailment: Vocabulary coverage, semantic preservation (maximum sentence spanning = 1), and coherence reward modules.
- Without semantic preservation: Vocabulary coverage and coherence reward modules only.
- Without coherence: Vocabulary coverage and semantic preservation modules only.

Ablation experiment results demonstrate the effectiveness of Re-RIGHT. Compared to the 1 : 1 sentence-level entailment setting (maximum sentence spanning = 1), the full framework (maximum sentence spanning = 4) achieves a 7.3-point improvement in semantic preservation, indicating that decomposing a complex reference sentence

Setting	Voc.	Sem.	Coh.
Vanilla	59.2	69.5	88.7
Full framework	76.0	80.6	83.1
1 : 1 entailment	74.0	73.3	83.7
No Sem.	84.0	52.7	83.1
No Coh.	88.6	79.4	54.6

Table 7: Evaluation results from the ablation study.

into multiple simpler sentences is beneficial for preserving meaning.

Although the setting without the semantic preservation reward yields higher vocabulary coverage than the full framework, its semantic preservation score drops substantially by 27.9 points. This suggests that the model tends to prioritize lexical simplification without preserving the original meaning of texts. Likewise, removing the coherence reward also produces high vocabulary coverage; however, the coherence score greatly decreases to 54.6, falling into the range defined as “frequently unnatural” by the coherence scoring prompt and leading to low-quality outputs.

E Further Discussion

E.1 Tradeoff between Vocabulary Coverage and Coherence

Due to the nature of comparative prompting with a reference text and a candidate simplification, coherence scores produced by LLM-based evaluators tend to decrease as the two texts become more dissimilar. This represents a tradeoff with our objective of lexical simplification, and it is inevitable that coherence scores decrease in experiments using relatively small (4B) models, especially in non-English language settings.

Importantly, our data and model generalization experiments (§6.4) demonstrate that our framework is not dependent on specific models, so researchers may be able to mitigate this issue by employing larger policy and evaluator models.

E.2 Choice of Evaluation Metrics

We acknowledge that SARI (Xu et al., 2016) and LENS (Maddela et al., 2023) are widely used metrics in the field of text simplification. However, both metrics require parallel reference sentence corpora. Since Re-RIGHT is trained in a fully unsupervised reinforcement learning setting, and no proficiency-specific parallel sentence datasets exist for Japanese, Korean, and Chinese to the best of our knowledge, constructing parallel level-aligned

corpora in four languages falls outside the scope of this work. Therefore, we instead mainly focus on actual vocabulary coverage scores grounded in language acquisition theories such as controlled vocabulary approaches.

E.3 Justifications for Excluding Syntactic Simplification

We acknowledge that syntactic simplification, in addition to lexical simplification, may also facilitate L2 comprehension. Nevertheless, we can offer both psycholinguistic and methodological justifications regarding our decision to exclude syntactic simplification from the scope of this work.

From a psycholinguistic perspective, our focus on lexical simplification can be supported by the Lexical Bottleneck Hypothesis (Hopp, 2018), which suggests that difficulties in L2 text processing often arise not from insufficient syntactic knowledge per se, but from limited lexical knowledge. In particular, limited lexical knowledge imposes significant cognitive load, preventing learners from effectively allocating resources to syntactic parsing. Conversely, increased lexical knowledge facilitates more efficient syntactic processing and promotes grammar acquisition from language input.

From a methodological perspective, this design choice enables our Re-RIGHT framework to be more effectively applied in multilingual settings. Due to substantial typological variation across languages from different language families, syntactic complexity is difficult to measure in a unified standard across languages. In contrast, lexical complexity can be measured in a more consistent and authoritative manner such as proficiency tests, making our framework more scalable and generalizable for multilingual text simplification.

Nevertheless, it is worth noting that our semantic preservation reward module (§5.2), which employs a 1:N sentence-level bidirectional entailment scoring mechanism, implicitly promotes syntactic simplification. Under our approach, a single sentence in the original text can be expressed as up to n sentences in the simplified output. This design naturally encourages the policy model to decompose long and syntactically complex sentences into multiple shorter and structurally simpler sentences.

F Detailed Results

F.1 Vocabulary Coverage of Prompting-Based Baselines

The following tables (Table 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, and 24) illustrate zero-shot vocabulary coverage scores for each experimental model, according to the criteria in Figure 2.

F.2 Re-RIGHT Generations

Table 29 and 30 are examples of Re-RIGHT generations by proficiency level in English and a non-English language. We use Qwen3-4B-Instruct-2507 as a trained policy model.

F.3 Re-RIGHT Generations Compared with a Prompting-Based Approach

We present a case where the prompting-based method fails to achieve high vocabulary coverage and Re-RIGHT successfully provides a sufficient number of level-appropriate words. In particular, base models exhibit low performance at easy proficiency levels. Table 8 demonstrates an example case for Japanese JLPT N5 (the easiest proficiency level).

Reference	... 2億円を超える落札額の産駒は以下の6頭いたが、獲得賞金が落札額を上回ったのはアドマイヤグルーヴのみであった。 ... (Translation: ... Six foals sold for over ¥200 million, but only Admire Groove’s career earnings exceeded its winning bid. ...)
Base model	... 2億円を超える落札額の馬は以下の6頭あるが、賞金が落札額を上回ったのはアドマイヤグルーヴだけだった。 ...
Re-RIGHT	... その中で、2億円以上で売られた子馬は全部で6頭あったが、実際にもらった賞金の方が売ったお金より多いのは、アドマイヤグルーヴという名前の馬だけだった。 ...

Table 8: An example comparison between generations from the base 4B model with the prompting method and the trained 4B model with Re-RIGHT.

In Table 8, the base model directly uses the term “落札額 (*winning bid*)”, which is a relatively complex Kango (Sino-Japanese word) for beginners. In contrast, Re-RIGHT efficiently converts the word to much simpler phrases, “売られた (*were sold*)” and “売ったお金 (*money from the sale*)”, while fitting the sentence context.

F.4 Case Study of Failure

Despite the powerful performance of Re-RIGHT, trained models occasionally produce undesirable outputs. We analyze several cases where our framework failed to perform as intended.

Reference	Judith Arlene Resnik (April 5, 1949 – January 28, 1986) was an American electrical engineer, software engineer, biomedical engineer, pilot and NASA astronaut who died in the Space Shuttle Challenger disaster.
CEFR A1	Judith Arlene Resnik (April 5, 1949 – January 28, 1986) was an American person who worked in electricity, software, medicine, flew planes , and also worked for NASA. She died in the Space Shuttle Challenger problem.

Table 9: An example failure case: grammatical error.

Reference	어린 프리드리히는 1840년대 이래 독일에서 맹위를 떨치던 자유주의 세력 중심의 혼탁한 정국을 겪었다. (<i>Translation: Young Friedrich experienced the turbulent political climate dominated by liberal forces that had been rampant in Germany since the 1840s.</i>)
TOPIK1	어린 프리드리히는 1840년대부터 독일에서 자유주의 세력이 강하게 나타나는 혼란스러운 정치 상황을 겪었다.
TOPIK6	어린 프리드리히는 1840년대부터 독일에서 자유주의 세력이 강하게 나타나는 혼란스러운 정치 상황을 겪었다.

Table 10: An example failure case: proficiency level misalignment (identical outputs).

Table 9 shows an example case where the trained policy model generates ungrammatical or unnatural outputs. Also, Table 10 represents a failure case to control proficiency levels, generating simpler but identical sentences for the easiest level (TOPIK1) and the most difficult level (TOPIK6) in Korean. These types of failures are more often observed in non-English generations, where smaller-parameter models can particularly be affected. Therefore, these cases may stem from the small size of the policy model (4B) and evaluator model we employ.

F.5 Re-RIGHT Result Statistics

Table 25, 26, 27, and 28 demonstrate the evaluation result of Re-RIGHT on Qwen3-4B-Instruct-2507 and gemma-3-4b-it with the Wikipedia and PGV test datasets.

G Prompt

Table 31 shows the exact prompt for coherence scoring via an evaluator LLM.

H Code and Data Release

We release the Re-RIGHT training code and available vocabulary level datasets online to contribute to the research community.¹⁸

I Use of AI Assistants

In this work, we use AI assistants for coding and checking grammatical errors.

¹⁸<https://github.com/jjhsnail0822/text-simplification>

Pigeon photography is an aerial photography technique invented in 1907 by the German apothecary Julius Neubronner, who also used pigeons to deliver medications. A homing pigeon was fitted with an aluminium breast harness to which a lightweight time-delayed miniature camera could be attached. Neubronner's German patent application was initially rejected, but was granted in December 1908 after he produced authenticated photographs taken by his pigeons. He publicized the technique at the 1909 Dresden International Photographic Exhibition, and sold some images as postcards at the Frankfurt International Aviation Exhibition and at the 1910 and 1911 Paris Air Shows.

Initially, the military potential of pigeon photography for aerial reconnaissance appeared interesting. Battlefield tests in World War I provided encouraging results, but the ancillary technology of mobile dovecotes for messenger pigeons had the greatest impact. Owing to the rapid development of aviation during the war, military interest in pigeon photography faded and Neubronner abandoned his experiments. The idea was briefly resurrected in the 1930s by a Swiss clockmaker, and reportedly also by the German and French militaries. Although war pigeons were deployed extensively during World War II, it is unclear to what extent, if any, birds were involved in aerial reconnaissance. The United States Central Intelligence Agency (CIA) later developed a battery-powered camera designed for espionage pigeon photography; details of its use remain classified.

(a) An example data chunk from English Wikipedia.

シューマンの読書好きは父親譲りで、主として文学と哲学を好んだ。シューマンは13歳のとき、当時興味を持った批評や詩、哲学的著作からの引用や自作の劇『精神』（未完）からの断章、両親の文章などを「スクランダー」というペンネームで『美しい黄金色の牧場の葉と花』としてまとめている。また、1825年から1828年の間に書いた自作の文集を「ムルデ河畔のロベルト」というペンネームで『雑録』としてまとめている。このころ、シューマンはゲーテの『ファウスト』をほとんど全部暗記し、友人たちからは「ファウスト」または「メフィスト」などと呼ばれていた。このほか、シューマンが手がけた文学作品として、コリオランを題材にした合唱付きの悲劇『ランデンドルファー兄弟』や喜劇『レオンハルトとマンテリエ』、ジャン・パウルから影響を受けた『6月の晩と7月の昼間』という小説があるが、いずれも未完である。シューマンが文学者をめざさず音楽の道を選んだことについて、ブリオンは「シューマンにとって、限界があり、厳密さを欠く文章表現よりも、音楽はずっと豊かで、多様で、陰影があり、緻密な言葉を提供した」と述べている。

(b) An example data chunk from Japanese Wikipedia.

어린 프리드리히는 1840년대 이래 독일에서 맹위를 떨치던 자유주의 세력 중심의 혼탁한 정국을 겪었다. 당시 자유주의자들은 독일인들의 열광적이고 광범위한 지지를 얻으며 세를 넓혀나가고 있었다. 자유주의자들은 독일의 통일을 희망하였고 입헌군주론자들은 새 헌법을 만들어 모든 인민들의 평등권 보장, 재산 보호, 그리고 기본 인권 보장을 구호로 내세웠다. 즉, 자유주의자들은 인민들의 입장을 대변하고 그들의 뜻에 따라 정책을 수립하는 정부를 원하였다. 프리드리히가 17살이 된 1848년, 민족주의자들과 자유주의자들은 독일 전 지역과 서유럽에 걸쳐 대규모의 시위를 주도하였다. 자유주의와 민족주의 세력은 집회와 결사의 자유, 언론의 자유 등의 자유권의 보장과 독일 의회의 수립, 그리고 헌법의 제정을 요구했다. 비록 독일에서의 혁명은 뚜렷한 족적을 남기지는 않았지만, 프리드리히가 어릴 때 목도한 이 자유주의는 훗날 그의 일생에 걸쳐 큰 영향력을 발휘하게 된다.

(c) An example data chunk from Korean Wikipedia.

在18世紀中期，西方文明對其他文明的影響是西方思想家的辯論焦點之一，不少學者認為西方文明優化了其他文明，但也有學者認為西方文明的入侵腐化了其他文明。庫克的三次航海探索正值這個辯論的高潮，因此他的航海發現或多或少讓西方思想家對地球另一邊鮮為西方所知的文化有稍進一步的了解。不過，庫克對這個命題並不特別關心，從他的周記所見，也不見得出他對「高貴野蠻人」（Noble Savage）這種在當時盛行的看法有特別的興趣。在19世紀，波蘭裔英國小說家約瑟夫·康拉德曾經對歷代航海家和探險家的動機作出比較，他認為庫克以前的航海家和探險家主要以「掠奪」（acquisitive）為動機，而庫克則主要以「科學」（scientific）為動機，因此兩者本質上具有顯著的分別。但有其他學者認為，庫克三次航海旅程的費用要由英國政府動用公帑承擔，這意味出海的計劃和目的受到納稅人監察，在這種背景下，庫克在旅程途中也不時把新發現的地方宣告為英國領土，因此，如果說他的航海旅程完全不具「掠奪」性質，也不是準確的說法。

(d) An example data chunk from Chinese Wikipedia (traditional Chinese character version).

Table 11: Examples of non-parallel “Featured Article” data chunks, excerpted from different articles.

In the months leading to the election of President John Evans Atta Mills, many Ghanaians, including those abroad, feared that a New Democratic Congress (NDC) win would morph into another reign by the party's founder and former military ruler, President Jerry John Rawlings. The transition was smooth, and the relationship between Presidents Rawlings and Mills, seemed cordial. JJ Rawlings, an unofficial blog site created to highlight the work and thoughts of the former president posted a positive report on May 18, 2009 about President Rawlings' trip to South Africa on behalf of President Mills. It reported: President Mills thanked President Rawlings for representing Ghana at the inauguration and assured him that there would be follow up measures to take advantage of the opportunities for mutually beneficial relations between Ghana and South Africa. But lately, tensions have been rising between President Mills and other members of the NDC party including President Rawlings. Ato Kwamena Dadzie recently wrote in his post, "A President Under Siege": President Mills is under siege. No doubt about that. He has become such an easy target that an increasing number of people are baying to take a hit at him. What is most intriguing is that the heaviest hitters are people within his own party. The opposition can take a vacation. Just last week, it was the majority leader in parliament, Alban Bagbin. He said the president, who likes to portray himself as a humble sheep, has surrounded himself with aides and confidantes who behave like foxes, hyenas and lions. These people, according to Mr. Bagbin, tend to intimidate and harass anyone who tries to offer some useful counsel to the president.

(a) An example English data chunk.

国際女性デーが世界中で祝われた数日前、女性として初めて宇宙に行ったワレンチナ・テレシコワが、75歳の誕生日を迎えた。宇宙開発競争の発端は、アメリカとソ連がロケット技術開発でしのぎを削った第二次世界大戦後に遡るが、1957年にソ連が人工衛星スプートニクを上げたことで本格的な競争が始まった。1961年には、ポストーク（ロシア語で東の意味）1号に搭乗したユーリイ・ガガーリンが人類として初めて宇宙に行った。数年後、ポストーク6号が打上られ、テレシコワが宇宙飛行した初の女性となった。Engineering Pathway Blogでは、去年、テレシコワの宇宙飛行を記念し、彼女の飛行時間の記録に注目した。大気圏に再突入した時点で、テレシコワの飛行時間は、それまでのアメリカ人宇宙飛行士全員の飛行時間の合計を上回っていたのだ。歴史上の今日：1963年6月16日、ソ連のポストーク6号に搭乗したワレンチナ・テレシコワが初めて宇宙に行った女性となった。3日間の宇宙飛行を終えたテレシコワの飛行時間は、当時のアメリカ人宇宙飛行士全員の飛行時間の合計を上回った。[...]

(b) An example Japanese data chunk.

해변가의 파라솔 그늘 아래 여유롭게 누워있거나, 관광객이 되어 손에 가이드북을 들고 생소한 도시의 거리를 돌아다니거나, 아니면 집에서 휴식을 취하며 자유시간을 즐기기. 이게 휴일을 보내는 일반적인 방법이 아닐까 싶다. 그러나 많은 일본인들은 직장에서 최대한 휴가를 내지 않으려 이러한 휴일을 매년 날려 버리곤 한다. 익스페디아 (Expedia) 일본지부가 선진국 11개국을 대상으로 한 '휴가 부족 (Vacation Deprivation) 상황'에 관한 설문조사'에 따르면 일본 노동자들은 1년에 평균 7.9일의 휴가를 보낸다. 일본에서는 평균적으로 15일의 유급 휴가가 주어지는데 이는 미국에 이어 두 번째로 낮은 수다. 사용하지 않은 휴가 일수로는 일본이 가장 높은 것으로 나타났다. 이는 일본인들이 경제적 위기에 중압감과 불안감을 느끼고 있고, 또한 휴가를 사용함으로써 "동료에게 업무 부담을 늘리는 폐"를 끼치지 않고자 하는 직장 내 분위기 때문인 것으로 보여진다.

(c) An example Korean data chunk.

今日世界看似平坦，从亚洲、非洲、欧洲到美洲，世界人民都感受到全球经济衰退的负面后果，本文试图记录在庞大金融危机之下，世界各地一般民众所见所受的社会冲击。这场风暴最显著的迹象为美国华尔街重挫与房地产市场崩盘，虽然我们不该小看这些经济灾害，但更要关注全球民众每日目睹与身陷的灾情。例如经济警讯已迫使许多韩国民众更改或放弃旅游计划，汶莱向来是韩国人的观光胜地，如今旅游业亦受波及；由于旅客人数下滑，埃及观光业人员已抱怨薪水未如期入帐。美国消费力萎缩压低孟加拉成长出口需求，影响当地雇员众多的成衣出口业，美国与英国为孟加拉产品主要出口国。俄罗斯金融危机反映在政府削减医疗支出；日本衰退情况包括缩短百货公司营业时间、车辆销量降低，以及临时屋宅、公园及网咖里出现愈来愈多失业游民。经济急冻在乌克兰非常具体，去年12月由于政府未准时缴纳水费，使首都基辅多数地区一星期无热水可用，Evie of Kiva Stories from the Field博客描述基辅居民受冻的困境：在零下低温与大陆型寒冬夹击下，一星期没有热水着实痛苦，人们甚至无法洗碗盘，因为水全都在管线里冻结，尽管热水恢复供应三天后，暖气装置功能尚未正常，住家仍然寒冷，民众也受到风寒感冒。香港向来以全球金融中心见长，很意外得知汇丰控股股价重挫至1995年来的最低点，更让一位电视评论员在提及此事时，因不可置信而落泪。

(d) An example Chinese data chunk.

Table 12: Examples of Parallel Global Voices data chunks, excerpted from different articles.

Lang.	Level	Original (Avg. \pm Std.)
EN	A1	20.2 \pm 7.5
	A2	37.5 \pm 10.5
	B1	58.0 \pm 10.4
	B2	73.5 \pm 9.0
	C1	77.2 \pm 8.5
	C2	81.9 \pm 7.6
JA	N5	20.6 \pm 6.3
	N4	31.3 \pm 8.2
	N3	53.5 \pm 10.2
	N2	58.8 \pm 9.5
	N1	71.6 \pm 10.2
KO	TOPIK1	19.5 \pm 8.6
	TOPIK2	33.4 \pm 9.1
	TOPIK3	44.4 \pm 9.9
	TOPIK4	58.0 \pm 10.4
	TOPIK5	63.9 \pm 10.3
	TOPIK6	68.9 \pm 10.6
ZH	HSK1	19.1 \pm 6.0
	HSK2	28.2 \pm 9.1
	HSK3	40.0 \pm 11.7
	HSK4	47.4 \pm 12.2
	HSK5	51.2 \pm 14.8
	HSK6	59.3 \pm 15.5
	HSK7-9	67.0 \pm 19.3

Table 13: Vocabulary coverage score baseline statistics for Reference Texts in Wikipedia Featured Article Dataset.

Lang.	Level	Simplified (Avg. \pm Std.)
EN	A1	29.0 \pm 10.9
	A2	47.4 \pm 12.0
	B1	65.5 \pm 11.6
	B2	78.7 \pm 10.1
	C1	79.8 \pm 8.9
	C2	84.4 \pm 7.8
JA	N5	22.9 \pm 7.1
	N4	34.0 \pm 9.7
	N3	56.7 \pm 10.7
	N2	61.4 \pm 10.3
	N1	73.6 \pm 11.3
KO	TOPIK1	21.7 \pm 9.5
	TOPIK2	36.2 \pm 9.5
	TOPIK3	47.4 \pm 10.2
	TOPIK4	60.7 \pm 10.4
	TOPIK5	65.9 \pm 10.4
	TOPIK6	70.5 \pm 10.6
ZH	HSK1	21.1 \pm 7.2
	HSK2	30.7 \pm 10.1
	HSK3	42.3 \pm 13.1
	HSK4	49.3 \pm 13.2
	HSK5	53.3 \pm 16.0
	HSK6	61.0 \pm 15.9
	HSK7-9	67.4 \pm 20.1

Table 15: Vocabulary coverage score statistics for Qwen3-4B.

Lang.	Level	Simplified (Avg. \pm Std.)
EN	A1	39.8 \pm 9.7
	A2	60.9 \pm 9.8
	B1	75.0 \pm 10.0
	B2	84.8 \pm 8.2
	C1	85.5 \pm 7.2
	C2	88.3 \pm 6.6
JA	N5	26.4 \pm 7.6
	N4	39.8 \pm 9.3
	N3	62.4 \pm 10.2
	N2	66.2 \pm 9.9
	N1	76.9 \pm 9.9
KO	TOPIK1	25.9 \pm 9.8
	TOPIK2	42.0 \pm 10.6
	TOPIK3	53.0 \pm 10.8
	TOPIK4	64.5 \pm 10.5
	TOPIK5	69.0 \pm 10.5
	TOPIK6	72.4 \pm 10.9
ZH	HSK1	27.4 \pm 7.6
	HSK2	38.9 \pm 9.6
	HSK3	50.8 \pm 11.8
	HSK4	58.7 \pm 10.3
	HSK5	62.7 \pm 12.4
	HSK6	70.2 \pm 12.2
	HSK7-9	76.3 \pm 14.1

Table 14: Vocabulary coverage score statistics for Qwen3-4B-Instruct-2507.

Lang.	Level	Simplified (Avg. \pm Std.)
EN	A1	35.3 \pm 11.1
	A2	54.9 \pm 11.3
	B1	72.9 \pm 11.3
	B2	83.6 \pm 9.1
	C1	83.2 \pm 8.3
	C2	87.2 \pm 7.4
JA	N5	26.1 \pm 8.3
	N4	38.6 \pm 10.4
	N3	60.5 \pm 10.8
	N2	63.8 \pm 9.9
	N1	75.1 \pm 10.0
KO	TOPIK1	24.6 \pm 9.7
	TOPIK2	39.4 \pm 10.0
	TOPIK3	51.9 \pm 10.2
	TOPIK4	63.2 \pm 11.6
	TOPIK5	68.4 \pm 10.9
	TOPIK6	72.2 \pm 10.7
ZH	HSK1	22.6 \pm 6.8
	HSK2	32.6 \pm 9.5
	HSK3	45.6 \pm 12.0
	HSK4	53.5 \pm 11.3
	HSK5	58.3 \pm 13.2
	HSK6	66.6 \pm 12.8
	HSK7-9	73.6 \pm 15.3

Table 16: Vocabulary coverage score statistics for Qwen3-8B.

Lang.	Level	Simplified (Avg. \pm Std.)
EN	A1	33.9 \pm 11.0
	A2	54.6 \pm 11.8
	B1	69.5 \pm 11.2
	B2	80.4 \pm 9.8
	C1	78.9 \pm 8.8
	C2	83.0 \pm 7.7
JA	N5	24.7 \pm 7.5
	N4	36.5 \pm 9.5
	N3	57.1 \pm 10.8
	N2	61.1 \pm 9.8
	N1	72.7 \pm 10.3
KO	TOPIK1	23.8 \pm 9.7
	TOPIK2	38.6 \pm 10.5
	TOPIK3	49.7 \pm 11.0
	TOPIK4	62.4 \pm 10.6
	TOPIK5	68.0 \pm 10.5
	TOPIK6	71.6 \pm 10.5
ZH	HSK1	20.5 \pm 6.6
	HSK2	30.4 \pm 9.7
	HSK3	43.6 \pm 12.4
	HSK4	50.8 \pm 12.7
	HSK5	54.6 \pm 15.3
	HSK6	63.0 \pm 15.7
	HSK7-9	70.0 \pm 18.9

Table 17: Vocabulary coverage score statistics for Qwen3-14B.

Lang.	Level	Simplified (Avg. \pm Std.)
EN	A1	48.3 \pm 10.4
	A2	65.6 \pm 9.1
	B1	76.9 \pm 9.1
	B2	83.0 \pm 7.6
	C1	84.4 \pm 6.8
	C2	87.8 \pm 5.7
JA	N5	26.5 \pm 8.5
	N4	36.6 \pm 9.8
	N3	56.4 \pm 10.4
	N2	60.8 \pm 9.7
	N1	73.2 \pm 10.0
KO	TOPIK1	25.3 \pm 10.1
	TOPIK2	37.1 \pm 9.7
	TOPIK3	48.7 \pm 10.3
	TOPIK4	60.8 \pm 11.1
	TOPIK5	65.8 \pm 10.7
	TOPIK6	69.6 \pm 10.6
ZH	HSK1	20.7 \pm 7.0
	HSK2	29.8 \pm 9.8
	HSK3	41.8 \pm 12.6
	HSK4	49.5 \pm 13.1
	HSK5	52.9 \pm 15.0
	HSK6	60.6 \pm 16.1
	HSK7-9	67.9 \pm 19.8

Table 19: Vocabulary coverage score statistics for gemma-3-4b-it.

Lang.	Level	Simplified (Avg. \pm Std.)
EN	A1	38.3 \pm 9.6
	A2	59.6 \pm 9.4
	B1	73.9 \pm 10.2
	B2	83.2 \pm 8.9
	C1	84.1 \pm 8.0
	C2	87.2 \pm 7.0
JA	N5	28.0 \pm 7.7
	N4	41.4 \pm 9.6
	N3	61.7 \pm 10.3
	N2	65.2 \pm 9.6
	N1	74.9 \pm 9.9
KO	TOPIK1	24.3 \pm 9.8
	TOPIK2	39.3 \pm 9.4
	TOPIK3	51.1 \pm 10.7
	TOPIK4	63.3 \pm 10.6
	TOPIK5	68.4 \pm 10.1
	TOPIK6	71.6 \pm 11.2
ZH	HSK1	20.8 \pm 6.8
	HSK2	30.2 \pm 9.9
	HSK3	42.8 \pm 12.6
	HSK4	49.6 \pm 13.1
	HSK5	53.8 \pm 15.3
	HSK6	60.8 \pm 16.2
	HSK7-9	67.8 \pm 19.7

Table 18: Vocabulary coverage score statistics for Qwen3-32B.

Lang.	Level	Simplified (Avg. \pm Std.)
EN	A1	57.0 \pm 10.3
	A2	73.8 \pm 8.7
	B1	82.5 \pm 8.3
	B2	87.1 \pm 7.3
	C1	87.2 \pm 6.7
	C2	90.1 \pm 6.2
JA	N5	40.2 \pm 9.5
	N4	53.8 \pm 10.1
	N3	70.7 \pm 9.8
	N2	70.8 \pm 9.7
	N1	80.4 \pm 9.5
KO	TOPIK1	36.7 \pm 11.0
	TOPIK2	53.9 \pm 10.2
	TOPIK3	62.4 \pm 11.0
	TOPIK4	70.8 \pm 10.5
	TOPIK5	74.5 \pm 9.4
	TOPIK6	76.4 \pm 10.2
ZH	HSK1	34.8 \pm 10.7
	HSK2	47.0 \pm 13.9
	HSK3	57.6 \pm 14.4
	HSK4	64.1 \pm 13.3
	HSK5	67.1 \pm 15.0
	HSK6	73.4 \pm 15.0
	HSK7-9	75.3 \pm 18.1

Table 20: Vocabulary coverage score statistics for gemma-3-12b-it.

Lang.	Level	Simplified (Avg. \pm Std.)
EN	A1	53.4 \pm 10.6
	A2	72.4 \pm 9.0
	B1	82.5 \pm 8.4
	B2	87.8 \pm 6.7
	C1	87.3 \pm 6.2
	C2	89.7 \pm 5.6
JA	N5	40.7 \pm 9.6
	N4	55.2 \pm 10.0
	N3	71.8 \pm 9.5
	N2	71.6 \pm 9.6
	N1	80.6 \pm 9.2
KO	TOPIK1	38.4 \pm 11.6
	TOPIK2	53.3 \pm 10.5
	TOPIK3	64.4 \pm 11.4
	TOPIK4	71.6 \pm 9.9
	TOPIK5	75.2 \pm 9.4
	TOPIK6	76.8 \pm 9.8
ZH	HSK1	39.7 \pm 10.3
	HSK2	50.0 \pm 12.2
	HSK3	62.5 \pm 11.1
	HSK4	66.8 \pm 10.2
	HSK5	71.1 \pm 10.8
	HSK6	77.0 \pm 9.8
	HSK7-9	81.6 \pm 11.4

Table 21: Vocabulary coverage score statistics for gemma-3-27b-it.

Lang.	Level	Simplified (Avg. \pm Std.)
EN	A1	49.6 \pm 10.2
	A2	68.7 \pm 8.9
	B1	81.7 \pm 7.3
	B2	87.6 \pm 7.2
	C1	87.9 \pm 6.5
	C2	90.0 \pm 6.1
JA	N5	42.3 \pm 9.1
	N4	58.5 \pm 9.6
	N3	73.3 \pm 9.0
	N2	74.2 \pm 9.4
	N1	81.4 \pm 9.2
KO	TOPIK1	38.3 \pm 11.8
	TOPIK2	53.9 \pm 11.2
	TOPIK3	61.6 \pm 10.6
	TOPIK4	69.5 \pm 10.2
	TOPIK5	72.4 \pm 9.7
	TOPIK6	75.7 \pm 9.5
ZH	HSK1	42.6 \pm 10.0
	HSK2	54.1 \pm 10.8
	HSK3	62.8 \pm 9.5
	HSK4	66.2 \pm 9.3
	HSK5	68.8 \pm 10.2
	HSK6	75.5 \pm 9.4
	HSK7-9	79.7 \pm 12.5

Table 23: Vocabulary coverage score statistics for gemini-2.5-flash.

Lang.	Level	Simplified (Avg. \pm Std.)
EN	A1	45.1 \pm 9.8
	A2	59.7 \pm 9.7
	B1	72.5 \pm 9.6
	B2	81.2 \pm 8.5
	C1	81.5 \pm 7.7
	C2	84.7 \pm 8.7
JA	N5	38.7 \pm 8.4
	N4	50.5 \pm 9.9
	N3	65.6 \pm 9.8
	N2	66.0 \pm 9.7
	N1	74.4 \pm 10.2
KO	TOPIK1	32.4 \pm 10.8
	TOPIK2	45.7 \pm 10.6
	TOPIK3	53.7 \pm 11.6
	TOPIK4	64.0 \pm 10.0
	TOPIK5	67.2 \pm 12.0
	TOPIK6	71.1 \pm 11.9
ZH	HSK1	26.1 \pm 8.2
	HSK2	35.5 \pm 11.3
	HSK3	48.0 \pm 13.2
	HSK4	52.7 \pm 13.4
	HSK5	55.3 \pm 15.6
	HSK6	64.3 \pm 15.5
	HSK7-9	69.2 \pm 19.3

Table 22: Vocabulary coverage score statistics for gpt-5.2.

Lang.	Level	Simplified (Avg. \pm Std.)
EN	A1	39.3 \pm 13.0
	A2	52.6 \pm 12.0
	B1	70.2 \pm 10.1
	B2	79.8 \pm 8.5
	C1	81.7 \pm 7.5
	C2	84.9 \pm 7.2
JA	N5	39.4 \pm 10.3
	N4	46.1 \pm 10.7
	N3	62.5 \pm 11.0
	N2	62.8 \pm 9.6
	N1	73.5 \pm 10.2
KO	TOPIK1	36.1 \pm 11.9
	TOPIK2	44.9 \pm 11.0
	TOPIK3	55.4 \pm 11.1
	TOPIK4	64.7 \pm 10.6
	TOPIK5	69.0 \pm 10.0
	TOPIK6	72.1 \pm 11.2
ZH	HSK1	36.2 \pm 11.8
	HSK2	46.2 \pm 12.5
	HSK3	58.4 \pm 10.6
	HSK4	61.8 \pm 9.7
	HSK5	65.3 \pm 9.4
	HSK6	71.7 \pm 8.1
	HSK7-9	78.9 \pm 9.8

Table 24: Vocabulary coverage score statistics for claude-sonnet-4-5-20250929.

Lang.	Level	Voc.	Sem.	Coh.
EN	A1	57.7	78.2	76.4
	A2	76.0	78.1	77.7
	B1	84.6	81.8	82.1
	B2	89.8	81.9	84.4
	C1	89.3	81.1	87.8
	C2	91.2	83.8	88.6
	Total	81.6	80.8	82.9
JA	N5	49.4	79.4	78.9
	N4	63.9	78.6	80.2
	N3	79.8	78.3	81.5
	N2	80.7	75.0	83.1
	N1	86.5	79.4	83.4
	Total	71.9	78.0	81.4
KO	TOPIK1	45.0	87.6	80.6
	TOPIK2	61.0	85.7	82.0
	TOPIK3	70.0	86.9	83.8
	TOPIK4	77.9	86.8	85.9
	TOPIK5	81.4	88.3	86.8
	TOPIK6	83.8	86.9	86.7
	Total	70.4	87.1	84.4
ZH	HSK1	60.0	78.0	80.5
	HSK2	71.6	76.3	81.3
	HSK3	79.4	76.4	82.9
	HSK4	82.7	76.3	84.5
	HSK5	84.1	77.2	84.9
	HSK6	89.4	77.2	85.0
	HSK7-9	92.9	75.1	86.4
	Total	80.2	76.6	83.7

Table 25: Evaluation results of Re-RIGHT on Qwen3-4B-Instruct-2507 with Wikipedia test data.

Lang.	Level	Voc.	Sem.	Coh.
EN	A1	60.3	75.8	76.4
	A2	76.4	67.9	78.8
	B1	85.1	71.5	82.4
	B2	91.4	75.0	84.6
	C1	92.2	80.1	87.8
	C2	93.3	78.8	88.8
	Total	83.0	74.6	83.0
JA	N5	50.8	66.1	78.1
	N4	67.3	60.7	78.8
	N3	80.9	55.2	80.8
	N2	82.1	59.6	82.2
	N1	88.6	56.3	84.3
	Total	73.5	59.6	80.7
KO	TOPIK1	48.3	76.4	78.0
	TOPIK2	62.6	78.8	79.7
	TOPIK3	72.8	77.8	83.0
	TOPIK4	81.0	76.6	84.7
	TOPIK5	84.2	78.8	86.1
	TOPIK6	87.1	79.6	85.9
	Total	72.8	78.1	82.9
ZH	HSK1	56.2	63.5	81.5
	HSK2	73.0	68.2	82.1
	HSK3	79.6	69.6	84.1
	HSK4	83.1	61.7	85.5
	HSK5	87.2	72.7	87.0
	HSK6	89.3	69.1	85.7
	HSK7-9	92.7	64.5	87.5
	Total	80.8	67.1	84.8

Table 26: Evaluation results of Re-RIGHT on Qwen3-4B-Instruct-2507 with PGV test data.

Lang.	Level	Voc.	Sem.	Coh.
EN	A1	57.7	76.4	76.2
	A2	74.4	78.4	77.6
	B1	84.7	81.8	79.7
	B2	89.2	80.7	82.3
	C1	90.1	80.1	84.0
	C2	92.0	81.9	84.6
	Total	81.5	79.9	80.8
JA	N5	51.9	77.2	77.2
	N4	64.5	76.1	79.1
	N3	79.4	77.8	79.6
	N2	80.8	73.9	80.7
	N1	87.0	76.6	80.9
	Total	72.5	76.3	79.5
KO	TOPIK1	49.0	85.0	79.0
	TOPIK2	63.4	81.8	81.3
	TOPIK3	73.1	84.1	81.7
	TOPIK4	79.2	84.0	82.3
	TOPIK5	82.7	86.0	83.4
	TOPIK6	84.6	85.9	83.7
	Total	72.5	84.5	82.0
ZH	HSK1	51.1	73.5	78.3
	HSK2	60.2	73.8	80.0
	HSK3	72.9	72.9	79.6
	HSK4	79.0	74.6	80.3
	HSK5	81.5	73.3	81.1
	HSK6	87.2	75.9	81.2
	HSK7-9	92.2	71.5	81.6
	Total	75.1	73.6	80.3

Table 27: Evaluation results of Re-RIGHT on Qwen3-4B-Instruct-2507 with Wikipedia test data.

Lang.	Level	Voc.	Sem.	Coh.
EN	A1	58.7	76.1	77.2
	A2	75.6	67.6	78.0
	B1	85.8	67.7	79.9
	B2	91.0	74.4	82.4
	C1	92.4	75.4	84.4
	C2	92.9	71.2	85.0
	Total	82.7	72.1	81.1
JA	N5	51.5	67.1	77.2
	N4	66.5	64.4	78.5
	N3	80.9	60.7	79.2
	N2	81.4	64.3	80.1
	N1	88.8	65.1	81.5
	Total	73.4	64.3	79.2
KO	TOPIK1	54.1	76.5	78.1
	TOPIK2	64.8	81.7	80.1
	TOPIK3	73.6	73.0	81.4
	TOPIK4	82.6	74.5	82.0
	TOPIK5	85.7	78.0	83.2
	TOPIK6	87.8	76.9	83.2
	Total	74.9	76.6	81.4
ZH	HSK1	45.5	65.3	78.6
	HSK2	58.2	64.6	81.4
	HSK3	71.9	67.9	81.5
	HSK4	77.2	65.5	81.2
	HSK5	84.0	70.9	83.3
	HSK6	87.0	69.2	82.4
	HSK7-9	91.4	68.8	82.4
	Total	74.3	67.5	81.6

Table 28: Evaluation results of Re-RIGHT on Qwen3-4B-Instruct-2507 with Wikipedia test data.

Level	Simplified Text
Reference	Pigeon photography is an aerial photography technique invented in 1907 by the German apothecary Julius Neubronner, who also used pigeons to deliver medications. A homing pigeon was fitted with an aluminium breast harness to which a lightweight time-delayed miniature camera could be attached. Neubronner's German patent application was initially rejected, but was granted in December 1908 after he produced authenticated photographs taken by his pigeons. ...
CEFR A1	Pigeon photography is a way to take pictures from the air. It was made in 1907 by a man named Julius Neubronner from Germany. He also used pigeons to carry medicine. A kind of pigeon that can find its way home was given a small metal thing on its chest. On that chest, a small camera that did not weigh much and worked after a short time could be put. Neubronner wrote a plan in Germany, but first it was not accepted. In December 1908, it was allowed after he showed real pictures taken by his pigeons. ...
CEFR A2	Pigeon photography is a way of taking photos from the air. It was made in 1907 by a man named Julius Neubronner from Germany. He also used pigeons to carry medicine. A kind of pigeon that can find its way home was given a small metal thing on its chest. On that chest, a small camera that did not weigh much and worked after a short time could be put. Neubronner wrote a patent in Germany, but first it was not accepted. In December 1908, it was allowed after he showed real photos taken by his pigeons. ...
CEFR B1	Pigeon photography is a way of taking photos from the air that was first made in 1907 by a German man named Julius Neubronner. He also used pigeons to carry medicine. A type of pigeon that can find its way home was given a small metal piece on its chest, where a small camera that worked slowly and weighed very little could be put. Neubronner wrote a patent in Germany, but it was not accepted at first. In December 1908, it was finally allowed after he showed real photos taken by his pigeons. ...
CEFR B2	Pigeon photography is a way of taking photos from the air that was first created in 1907 by a German man named Julius Neubronner. He also used pigeons to carry medicine. A type of pigeon that can find its way home was given a small metal piece on its chest, where a very light camera — which worked after a short delay — could be placed. Neubronner's idea for a German patent was first turned down, but in December 1908, it was finally accepted after he showed real photos taken by his pigeons. ...
CEFR C1	Pigeon photography is a type of aerial photography that was first created in 1907 by a German man named Julius Neubronner, who also used pigeons to carry medicine. A pigeon capable of finding its way home was given an aluminium device on its chest, where a small, light camera with a delay feature could be placed. Neubronner's patent in Germany was first turned down, but later approved in December 1908 after he showed real photos taken by his pigeons. ...
CEFR C2	Pigeon photography is a method used in aerial imaging that was first created in 1907 by a German man named Julius Neubronner, who also used pigeons to carry medicines. A type of homing pigeon was equipped with an aluminium device on its chest, allowing a small, light camera with a delay feature to be attached. Neubronner's patent request in Germany was first turned down, but it was finally approved in December 1908 after he showed real photos taken by his pigeons. ...

Table 29: Re-RIGHT generation examples by proficiency levels in English. The trained policy model is Qwen3-4B-Instruct-2507.

Level	Simplified Text
Reference	<p>... 집회와 결사의 자유, 언론의 자유 등의 자유권의 보장과 독일 의회의 수립, 그리고 헌법의 제정을 요구했다. 비록 독일에서의 혁명은 뚜렷한 족적을 남기지는 않았지만, 프리드리히가 어릴 때 목도한 이 자유주의는 훗날 그의 일생에 걸쳐 큰 영향력을 발휘하게 된다.</p> <p>(Translation: ... They demanded the guarantee of civil liberties such as freedom of assembly and association, freedom of the press, the establishment of a German parliament, and the enactment of a constitution. Although the revolution in Germany did not leave a distinct mark, the liberalism Friedrich witnessed in his youth would exert a profound influence throughout his life.)</p>
TOPIK1	... 사람들은 모임과 함께 행동할 수 있고, 언론도 자유로이 쓸 수 있도록 하는 등의 자유를 받고, 독일의 회의를 만들고, 새로운 법을 만드는 것을 요구했다. 독일에서 일어난 혁명은 특히 큰 변화를 만들지는 않았지만, 프리드리히가 어릴 때 보았던 자유주의는 그 후 그의 삶 전체에서 큰 도움이 되었다.
TOPIK2	... 사람들이 모여 이야기할 수 있고, 언론도 자유롭게 쓸 수 있도록 하는 등의 자유를 받고, 독일의 회의를 만들고, 새로운 법을 만드는 것을 요구했다. 독일에서 일어난 혁명은 특히 큰 변화를 가져오지는 않았지만, 프리드리히가 어릴 때 보았던 자유주의의 생각은 그 후 그의 삶 전체에서 큰 영향을 주게 되었다.
TOPIK3	... 회의나 함께 행동할 수 있는 자유, 언론에서 자유롭게 말할 수 있는 권리 같은 중요한 권리들을 받고, 독일의 의회를 만들고, 새로운 헌법도 만들 것을 요구했다. 독일 내에서 일어난 혁명은 특히 큰 변화를 가져오지는 않았지만, 프리드리히가 어릴 때 보았던 자유주의의 생각은 그 후 그의 삶 전체에서 매우 큰 영향을 주게 되었다.
TOPIK4	... 회의나 결사, 언론 등 자유를 보장하고, 독일의 의회를 만들고, 헌법을 만들어야 한다고 요구했다. 독일 내에서 일어난 혁명은 특히 큰 변화를 가져오지는 않았지만, 프리드리히가 어릴 때 보았던 자유주의의 모습은 그 후 그의 삶 전체에서 매우 중요한 영향을 주게 되었다.
TOPIK5	... 회의나 단체 활동, 언론 사용 등 자유를 보장하고, 독일의 의회를 만들고 새로운 헌법을 만들어야 한다고 요구했다. 독일 내에서 일어난 혁명은 특히 명확한 결과를 가져오지는 않았지만, 프리드리히가 어릴 때 직접 보았던 자유주의의 개념은 그 후 그의 삶 전체에서 매우 중요한 영향을 주게 되었다.
TOPIK6	... 회의나 단체 활동, 언론 사용 등 자유를 보장받을 수 있는 조건과 독일의 의회를 만들고, 헌법을 만들어야 한다는 요구를 제시했다. 독일 내에서 일어난 혁명은 특히 명확한 결과를 가져오지는 못했지만, 프리드리히가 어릴 때 직접 보았던 자유주의의 사례는 그 후 그의 삶 전체에서 매우 큰 영향을 주게 되었다.

Table 30: Re-RIGHT generation examples by proficiency levels in a non-English language (Korean). The trained policy model is Qwen3-4B-Instruct-2507.

You are evaluating {language} text quality for a text simplification system.

Given [ORIGINAL_TEXT] and [SIMPLIFIED_TEXT], focus ONLY on how natural and fluent the [SIMPLIFIED_TEXT] reads as a rewrite of the [ORIGINAL_TEXT]. Rate the NATURALNESS of the [SIMPLIFIED_TEXT] as if it were written by a native speaker, strictly according to the following rules:

100 = indistinguishable from a native human-written well-edited text
80-99 = highly natural with only minor unnatural phrasing
60-79 = generally understandable but contains multiple awkward and unnatural expressions
30-59 = sounds clearly machine-generated, frequently unnatural or repetitive
0-29 = extremely incoherent or clearly broken language

Critical penalties:

- Strongly penalize repetitive template phrasing (e.g., repeating the same word/phrase many times to fill text).
- Strongly penalize awkward connective phrases or unnatural sentence patterns.
- Do NOT reward being 'simple' if it becomes unnatural. Simple but fully natural text should still receive a high score.

Use the full 0–100 range. Reflect even small differences in naturalness with 1-point precision.
Output only a single integer from 0 to 100, and say nothing else.

[ORIGINAL_TEXT]
{original_text}

[SIMPLIFIED_TEXT]
{simplified_text}

Table 31: Prompt for coherence scoring.