TaskGen at TSAR 2025 Shared Task: Exploring prompt strategies with linguistic knowledge

Juan Cruz Oviedo
juan.oviedo.869@unc.edu.ar
Universidad Nacional de Córdoba

Elisabet Comelles
elicomelles@ub.edu
Universitat de Barcelona

Laura Alonso Alemany
lauraalonsoalemany@unc.edu.ar
Universidad Nacional de Córdoba

Jordi Atserias Batalla
jordi.atserias@bsc.es
Barcelona Supercomputing Center

Abstract

TaskGen ranked as 6th best team in the TSAR 2025 shared task for English text adaptation to a target CEFR level. Our experiments consisted of prompting a Llama-3.1-8B-Instruct model with linguistic descriptors of the target level, examples of adaptations and multi-step approaches. Our best run, 13th in the overall ranking, applied an ensemble strategy using a voting mechanism to find the most adequate among 10 texts, each produced by a different prompting strategy.

1 Introduction

Text simplification (TS) aims to reduce the linguistic complexity of texts while preserving their meaning and coherence. This facilitates accessibility for readers with low literacy and individuals with cognitive challenges (Siddharthan, 2014), but it is also very useful to adapt texts for language learners. In that context, TS presents the additional requirement to target the learners' proficiency levels.

The TSAR 2025 shared task on Readability-Controlled Text Simplification (RCTS) (Alva-Manchego et al., 2025) directly addresses this need by requiring participants to simplify English texts at B2 level or higher to specified target CEFR levels (A2, B1). This setup not only evaluates systems on their ability to reduce complexity, but also on their capacity to produce texts adequate for language learners at different stages of proficiency.

2 Related work

Automatic Readability Assessment (ARA) "refers to the task of modeling the reading and comprehension difficulty of a given piece of text, for a given target audience" (Vajjala, 2022). It has been widely studied across NLP, education and psychology, and is often applied to support tasks such as selecting suitable materials for L2 learners (Heilman et al., 2007; Vajjala and Meurers, 2012).

Traditional es to ARA relied on readability formulae, such as Flesch-Kincaid Grade level (Kincaid et al., 1975), SMOG Grading (McLaughlin, 1969) and Coleman-Liau Index (Coleman and Liau, 1975), among others.

More recent work has explored feature-based machine learning and neural models (Hancke et al., 2012; Vajjala and Meurers, 2012; Chen and Meurers, 2018; Deutsch et al., 2020; Lee et al., 2021; Lee and Vajjala, 2022).

Although central to understanding textual complexity, ARA also serves as a foundation for downstream tasks such as text simplification (TS). Early es to TS were rule-based or statistical, but recent progress has been driven by neural models (Martin et al., 2022; Maddela et al., 2021; Sheang and Saggion, 2021). With the emergence of large language models (LLMs), TS research has increasingly swift to fine-tuning and prompt engineering (Kew et al., 2023; Nozza and Attanasio, 2023; Martínez et al., 2024).

Building on TS, Readability-Controlled Text Simplification (RCTS) introduces explicit control over the target readability level, making it particularly valuable in educational contexts where texts must be aligned with learners' proficiency (Scarton and Specia, 2018). Most work on this area has followed supervised training es (Nishihara et al., 2019; Spring et al., 2021; Rios et al., 2021; Yanamoto et al., 2022; Agrawal and Carpuat, 2023); however, the drawback to these es is the limited number of datasets available.

Recent advances in LLMs and in their capacity for text generation have opened new directions for RCTS, with studies exploring sentence-level (Barayan et al., 2025; Chi et al., 2023) and passage-level simplification (Farajidizaji et al., 2024; Imperial and Tayyar Madabushi, 2023) in zero-shot and few-shot settings.

Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 160–172

3 System Description

We explored a prompt-based approach to text adaptation, assessing the performance of different prompt variants:

- **Few-shot learning**. Including examples of pairs of original and adapted texts in the prompt and adding linguistic information, such as overall reading comprehension descriptors of the target proficiency level (Council of Europe, 2020), curated vocabulary lists, and examples of appropriate grammatical constructions. Also, explicit instructions were used to obtain only the adapted text. That is, requiring the model to avoid references to the adaptation task or to the output CEFR level.
- **Multi-step procedures**. An initial adaptation was followed by additional instructions for iterative refinement.
- **Ensemble-based strategy**. Multiple adaptations generated by ten distinct prompts were compared, and the most suitable version was selected through a voting mechanism.

All prompt different strategies used *Llama-3.1-8B-Instruct* model (Meta, 2024) with precision bfloat16 and decoding default parameters (nucleus sampling with top-p = 0.9, temperature = 0.6, and stochastic decoding enabled). We opted for *Llama-3.1-8B-Instruct* model because it is open-source, cost-effective and had performed well in CEFR-based text classification (Comelles et al., 2025). Also, a previous version of the *Llama 3 8B* family had proved successful for a similar task (Barayan et al., 2025).

3.1 Methodology and lessons learned

After systematically evaluating eleven different experiments with the metrics and trial data provided by the organizers, and with qualitative, manual inspection of the produced texts, we selected the three experiments described below for submission.

We wanted to compare approaches that required increasing levels of effort, so that the evaluation of the task served to assess the impact of more complex approaches in the resulting performance. Our first and second approaches were based on a prompt with extensive linguistic descriptors, guidelines and examples describing the target CEFR level (Figure 4). After analyzing the performance of this approach on trial data, we found that such lengthy instructions on the linguistic characterization of the level damaged the preservation of the content of the original text. To address this issue, for run 1 we added a follow-up prompt (Figure 5) that focused on the preservation of the content of the original text. Alternatively, for run 2 we applied a follow-up prompt (Figure 6) that removed all text where the language model described the task it was performing, with the aim of leaving only the adapted text, without meta information. Finally, for run 3, we obtained adaptations from 10 different prompts (see Appendix D). Then, each text was assessed with criteria that had been found useful in the analysis of performance on trial data.

3.2 Run 1: Complementary prompts with CEFR descriptors and content preservation

The first run consisted of a prompt (see Figure 4) with extensive linguistic information on the characteristics of CEFR levels, and a follow-up prompt (see Figure 5) to emphasize content adequacy.

The linguistic information in the first prompt was:

- CEFR overall reading comprehension descriptors for A2 and B1 (Council of Europe, 2020).
- A list of 1681 words taken from Oxford University Press (n.d.), which lists the most frequent and relevant words from A1 to B2. From this list, the words classified as A2 and B1 were used, totaling 872 words for A2 and 809 words for B1.
- A list of 5 examples of each morphosyntactic construction listed as characteristic of the target level in (North et al., 2010). A total of 190 sentences were included (105 A2 sentences, and 85 B1 sentences). They can be found at `https://github.com/juan-oviedo/Recursos-Linguisticos`.

The output of this prompt was taken as input for a subsequent prompt (see Figure 5) that aimed to preserve the content of the original text, prioritizing content over adequacy to the target CEFR level.

The first prompt in this strategy resulted quite lengthy, with slightly over 3,800 tokens in average. Excessive length is believed to impair the performance of language models. With this two-step approach we tried to address this limitation of LLMs by focusing the task first on CEFR adequacy and then on content preservation.

3.3 Run 2: Corrections on meta-information

TaskGen second run relied on the same first prompt as run 1 (Figure 4), that is, the lengthy prompt with linguistic information andto CEFR level descriptors. Then a second prompt was applied to remove meta information on simplification or CEFR level (Figure 6). Such meta-information appeared frequently in the LLM's outputs, despite having included explicit instructions to provide the simplified text only.

3.4 Run 3: Prompt output ensemble

For the third run we implemented an ensemble strategy. We prompted the same language model, *Llama-3.1-8B-Instruct*, with 10 different prompts. Then, the 10 output texts were ranked by a voting mechanism.

Prompts, shown in Appendix C, ranged from very simple instructions to adapt the text to a given target level, to instructions with extensive linguistic information, and to multi-step strategies with follow-up prompts, including few-shot approaches with examples of pairs of original - adapted texts.

The voting mechanism was manually constructed by inspecting errors in the trial data and assessing their impact on performance. The final submission consisted of the following simple heuristic, applied to each of the 10 adaptations, starting from 0 points:

- **Avoid Keyword [0, -10]:** ten points were subtracted if the output text contained one or more predefined sequences belonging to an explanation of the task, in addition to the adapted text itself (e.g., "CEFR", "This text is a simplification", "Here you have a simplified version", "Here is the simplified text").
- **CEFR level adequacy [0, -1, -2, -3, -4]**: classification in CEFR levels was computed using the three classifiers provided by the organizers in the evaluation script. A penalty was applied when a classifier identified a proficiency level in the generated text that differed from the target level: one point was subtracted if a single classifier indicated a mismatch, two points if two classifiers disagreed, and three points if all three disagreed. An additional penalty was applied proportional to the distance between the predicted and target levels, with one point subtracted for each level of discrepancy, so that the penalty for each classifier could substract up to 4 points. As will be discussed in the following section, this item may have been weighted more heavily to emphasize compliance with the target CEFR level.
- **Text length [0, -1]**: one point was subtracted if the adapted text exceeded 1.3 times the length of the original. This threshold was established after manual inspection of errors in trial data, which showed that lengthier adaptations included content that was not in the original text.
- **Semantic similarity[0, -5]**: semantic similarity was measured using the MeaningBERT script provided by the organizers to calculate Bert-based similarity to the original text. If the resulting similarity score fell below 0.6, a penalty of five points was applied. This threshold was established as a rule of thumb, aiming for the adaptation to preserve at least two thirds of the meaning of the original text.

After voting, the text with the highest score for each level was selected. In cases of ties, the text with the higher MeaningBERT score was preferred. If ties persisted, final selection was based on the performance of the prompts in the trial data: prompts 8, 4, 6, 3, 2.

As will be discussed in the following Section, this ensemble strategy achieved the best results of all *TaskGen* submitted runs, but there is still room for improvement in this line of work. The voting mechanism could likely be improved by assigning weights empirically, for example through regression on a larger dataset. Instead, for this shared task, the strengths of votes was assigned based on our analysis of performance on the trial partition of the dataset. Also, we expect that an approach based on agents can be more flexible than this heuristic, especially in ranking previously unseen cases.

4 Results and Discussion

As reported in Alva-Manchego et al. (2025), our third run, the ensemble with a voting strategy, ranked best among all the runs submitted by *TaskGen*, making for the 13th best submission and making *TaskGen* the 6th best team in the shared task. Although the RMSE with respect to the CEFR level was acceptable, with an average of 0.628, this run excelled in meaning preservation with respect to the original and the reference, with 0.856 and 0.826 respectively.

	RMSE	meaning-orig	meaning-ref
Run1	0.592	0.791	0.786
Run2	0.561	0.752	0.773
Run3	*0.628*	*0.856*	*0.826*

Table 1: *TaskGen* results in the TSAR shared task.

As can be seen in Table 1, our first run, consisting of a single prompt with linguistic information about the target CEFR level and a follow-up prompt addressing content adequacy, scored quite well in adequacy to the target level, with an RMSE of 0.592. However, it scored lower than other approaches in preservation of the meaning of the original text and similarity to the reference, with 0.791 and 0.786 respectively. As a result, this run ended up ranking in the lower half of the participating systems.

Our second run, aiming to remove meta-information generated by the LLM about the task itself, obtained lower scores for meaning preservation with respect to run 1. Despite this drawback, it was our best run with respect to adequacy to the target CEFR level, with an RMSE of 0.561. This shows that the follow-up prompt for content adequacy of run 1 harmed compliance to the target CEFR level.

The ensemble strategy, our third run, produced a significant improvement on preservation of the meaning with respect to the original and the reference, although at the expense of worse accuracy with respect to the target CEFR level, with the worst RMSE for our three runs, at 0.628. Nevertheless, high meaning preservation made this our best overall ranked submission.

Since adequacy to the target CEFR level was the main weakness of our approaches, we conducted a more detailed analysis of this aspect. As shown in Figure 1, all three runs —particularly the third— struggled to adapt texts to the B1 level, while systematically succeeding at the A2 level. Still, the errors were relatively minor, typically corresponding to a one-level difference, that is, texts that should have been adapted to B1 were adapted to A2 instead. We will examine these cases further to improve the system.

In contrast, meaning preservation was significantly better for B1 than for A2, as illustrated in Figure 2, showing the probability density function (a smoothed version of a histogram) of the score for meaning preservation with respect to the original. Thus, adaptation to lower levels seems to make

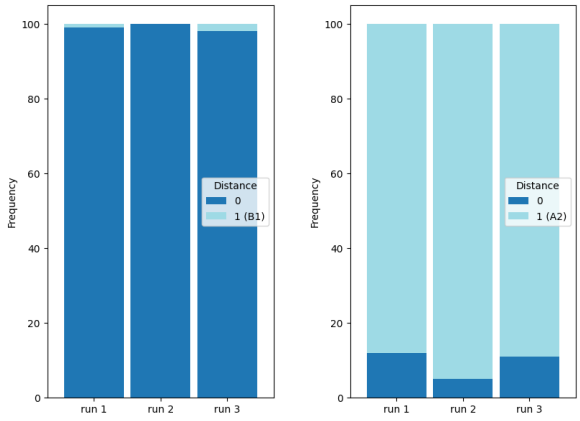

Figure 1: RMSE with target CEFR level for each of our three runs, for target level A2 (left) and B1 (right).

meaning preservation more difficult.

We found that run 3 was especially capable to adapt to different typologies of texts. For example, Figure 3 shows how runs 1 and 2 could not produce an adequate result for a text that was particularly challenging, because it could easily be interpreted as an instruction in itself by the language model. However, run 3 produced a better adaptation than the others, even if there is still room for improvement.

We believe that it is precisely the flexibility of the ensemble strategy to adapt to different typologies of texts that makes it our best ranked submission, ranking 13th best run and positioning *TaskGen* as the 6th best team in the shared task.

Finally, we find it useful to analyze how complex strategies improve performance upon simpler ones. In particular, we analyzed in trial data how different follow-up prompts improved performance on a base prompt. Table 2 displays the results obtained on the shared task evaluation data, applying the official evaluation scripts provided by Tsar organizers, for each of the individual prompts used in run 3. In particular, prompt 4, marked with an asterisk, was the base prompt for runs 1 and 2. As shown in Table 2, run 1 makes an important improvement on meaning preservation when compared to prompt 4 (from 0.75 to 0.79 on meaningbert-orig). As for run 2, it did not yield any improvement in meaning preservation but did improve in adequacy to the target CEFR level, achieving the best overall adequacy. Thus, it seems that both follow-up prompts did yield improvements, each in a different aspect. As future work, we plan to combine these complementary approaches to obtain texts that both preserve meaning and comply with the target level.

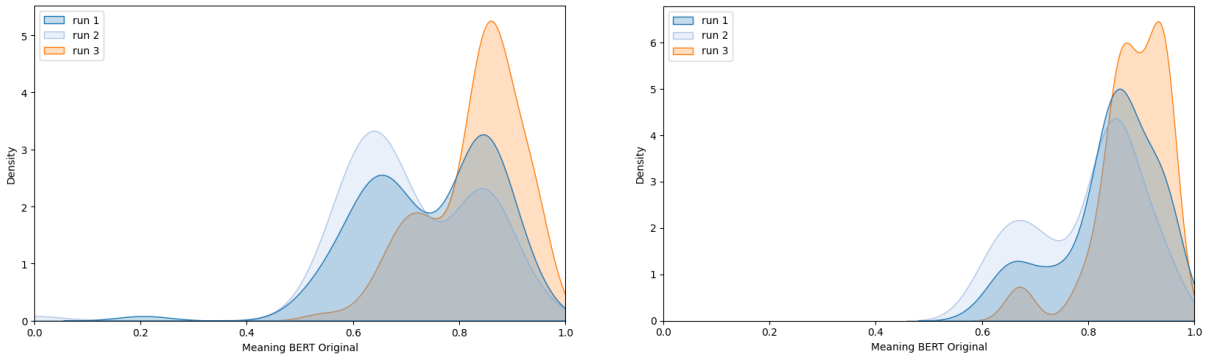

Figure 2: Density estimation of bert meaning preservation score to the original per target level (A2 left, B1 right).

exp.	prompt	weight_F1	rmse	meaningbert-orig	bertscore-orig	meaningbert-ref	bertscore-ref
01	1	0.6024	0.6403	0.7639	0.9274	0.7778	0.9330
02	2	0.5837	0.6633	0.7622	0.9264	0.7897	0.9321
03	3	0.6507	0.5958	0.7719	0.9261	0.7860	0.9308
04	4	0.6692	0.5745	0.7542	0.9267	0.7738	0.9314
05	5	0.6230	0.6708	0.7867	0.9248	0.7908	0.9269
06	6	0.6389	0.6364	0.7850	0.9283	0.7892	0.9292
07	7	0.6591	0.5874	0.7824	0.9318	0.7932	0.9362
08	8	0.6727	0.5958	0.7767	0.9283	0.7885	0.9325
09-run1	4+9	0.6633	0.5916	0.7906	0.9354	0.7863	0.9330
10-run2	4+10	**0.6826**	**0.5612**	0.7525	0.9262	0.7727	0.9311
11-run3	all	0.6353	0.6285	**0.8556**	**0.9378**	**0.8256**	**0.9371**

Table 2: *TaskGen* results of the experiments for the TSAR shared task on the final evaluation data.

5 Conclusions & future work

TaskGen submissions were the result of iterating on different prompting strategies, increasing their complexity based on the qualitative and quantitative analysis of performance on the trial dataset.

Results show that the ensemble with a voting strategy has proved effective. However, further research is required to improve adaptations to the B1 level, which were far worse than those for the A2 level. Also the proposed methods rely on manual prompt engineering and ensembling of different prompts results which may limit the applicability of the proposed solution to other problems or scale.

As future work, we aim to explore the specific contribution of the linguistic information used in the prompts, by analysing the individual effects of CEFR descriptors, CEFR-levelled wordlists and CEFR-levelled morphosyntactic structures on the LLM performance. We also plan to improve the voting heuristic by inferring the weight of each vote from empirical data, instead of manual inspection. To do that, we will work to obtain or build a bigger labelled corpus, so that empirically inferred votes are reliable. We also plan to examine more advanced prompting strategies, such as agentic and parallel prompting —where separate branches focus on CEFR adequacy and semantic semantic fidelity before combining. Finally, we intend to conduct fine-tuning experiments guided by automatic evaluation metrics, including CEFR-level classifiers and meaning preservation measures.

Acknowledgments

This research has been funded and made possible by Grant PID2023-149648NB-I00, funded by the Ministerio de Ciencia, Innovación y Universidades of the Spanish government, and the National Research Agency (AEI).

This work used computational resources from UNC Supercómputo (CCAD) – Universidad Nacional de Córdoba[1], which are part of SNCAD, República Argentina.

We also thank the reviewers for insightful comments and suggestions, which have contributed to improve this work.

[1] https://supercomputo.unc.edu.ar/

6 Lay Summary

This paper presents the taskGen submission to the TSAR 2025 Shared Task on Controlled-Readability Text Simplification. Text simplification (TS) is a process that makes texts easier to read and understand, while keeping their original meaning. It can help people readers with low literacy level and individuals with cognitive challenges (Siddharthan, 2014), but it is also very useful to adapt texts for language learners who need materials suited to their level of proficiency. The TSAR 2025 shared task on Readability-Controlled Text Simplification (RCTS) (Alva-Manchego et al., 2025) focuses on this idea: it asks participants to simplify English texts at an advanced level (B2 or higher) so that they match pre-intermediate and intermediate levels —CEFR A2 and B1 (the Common European Framework of Reference for Languages, or CEFR, is a widely used scale that describes language ability, where A1 is beginner and C2 is expert). The goal of the task was to see whether systems could not only make texts simpler, but also produce adapted versions that matched the right level for language learners. The taskGen submission includs three different methods, all based on Large Language Models (LLMs), that is AI systems trained to understand and generate text. Our first method uses two prompts (instructions to the AI): the first prompt describes in detail the language features typical of each CEFR level, and the second prompt emphasizes keeping the meaning of the original text. The second method also uses two prompts, but in the second one the AI is asked not to include explanations about why the text fits a certain level. Finally, our third method uses 10 different prompts and then a ranking system chooses the best simplified text. Our third method performed best. It ranked 13th overall among all the submitted methods and made our team the 6th best team overall in the shared task. This approach was particularly strong at keeping the meaning of the original text, but there is still room for improvement regarding compliance with target CEFR level.

References

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

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

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

Xiaobin Chen and Detmar Meurers. 2018. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510.

Alison Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee, and Jason S. Chang. 2023. Learning to paraphrase sentences to different complexity levels. *Transactions of the Association for Computational Linguistics*, 11:1332–1354.

M. Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Elisabet Comelles, Juan Cruz Oviedo, Laura Alonso, Roger Gilabert, and Joan Castellví. 2025. Exploring large language models for cefr-based text classification in foreign language education. In *Congreso Internacional De La Asociación Española de Lingüística Aplicada(AESLA)*.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing, Strasbourg.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

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

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India. The COLING 2012 Organizing Committee.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467, Rochester, New York. Association for Computational Linguistics.

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

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

Peter Kincaid, Robert P. Fishburne, 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.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813, Dublin, Ireland. Association for Computational Linguistics.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Paloma Martínez, Lourdes Moreno, and Alberto Ramos. 2024. Exploring large language models to generate easy to read content. *Preprint*, arXiv:2407.20046.

G Harry McLaughlin. 1969. Smog grading–a new readability formula. *The Journal of Reading*, 12(8):639–646.

Meta. 2024. Meta-Llama-3.1-8B-Instruct. `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`. Accessed: 2025-09-22.

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

Brian North, Angel Ortega, and Susan Sheehan. 2010. *A Core Inventory for General English*. British Council and EAQUALS.

Debora Nozza and Giuseppe Attanasio. 2023. Is it really that simple? prompting large language models for automatic text simplification in Italian. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 322–333, Venice, Italy. CEUR Workshop Proceedings.

Oxford University Press. n.d. The oxford 3000: The list of the 3000 most important words to learn in English, from A1 to B2 level. `https://www.oxfordlearnersdictionaries.com/external/pdf/wordlists/oxford-3000-5000/The_Oxford_3000.pdf`. Accessed: 2025-09-21.

Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161, Online and in Dominican Republic. Association for Computational Linguistics.

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

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Advaith Siddharthan. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics*, 165(2):259–298.

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

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

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.

A Description of linguistic resources

- CEFR overall reading comprehension descriptors for A2 and B1 (Council of Europe, 2020).
- List of A2 and B1 most relevant and frequent words, extracted from Oxford University Press (n.d.).
- A list 5 examples of each morphosyntactic construction listed as characteristic of A2 and B1, adapted from (North et al., 2010). The list of examples can be found in https://github.com/juan-oviedo/Recursos-Linguisticos.

B Example of the outputs using the different prompts

Original text:
Don't always follow positive feedback with negative feedback. When positive and negative feedback always appear to go hand in hand, the positives can become devalued and ignored. Ensure there are times when positive feedback is given for its own sake and resist the temptation to offer constructive criticism.
result of run 1 for level A2: You give good feedback. You say nice things to your students.
result of run 2 for level A2: You are a good teacher. You are kind to your students. Your students like you.
result of run 3 for level A2: Don't always say something is wrong. When we say something is wrong and then something is right, people start to ignore the good things. We should say something is good just because it is good.
For example, a student does a good job on a test. We should just say, "Well done!" and not add "But you could have done it faster." Or a friend helps us with something. We should say, "Thank you!" and not say "You could have done it better next time."
This way, people will appreciate the good things and feel happy when we give them positive feedback.

Figure 3: Results of our three runs for text 120 in the TSAR test dataset.

C Prompts for the submitted runs

Prompts for the submitted runs. The linguistic knowledge used in prompts is the following

- [DESCRIPTORS] Descriptors of the target CEFR level as provided by (Council of Europe, 2020).
- [LIST] Vocabulary list for each level as provided by Oxford University Press (n.d.).
- [GRAMMAR] 5 examples illustrating the morphosyntactic constructions of each level specified in (North et al., 2010), found at https://github.com/juan-oviedo/Recursos-Linguisticos.
- [LEVEL] and [CEFR LEVEL] The target level, that is, A2 or B1.
- [INPUT] The input text to be adapted. For TSAR 2025 shared task, each of the texts provided by the organizers.
- [ORIGINAL TEXT] The original in a pair of <original, adapted> texts.
- [SIMPLIFIED TEXT] The text adapted to the target level in a pair of <original, adapted> texts.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task.
Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners.
To simplify you may use the CEFR [LEVEL] [DESCRIPTORS], the [CEFR LEVEL] vocabulary in [LIST] and the [CEFR LEVEL] grammar structures in [GRAMMAR].
Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates.
Keep the order of the events.
Please avoid adding extra information to the text.
Omit any introduction or conclusion.
Only provide the simplified version of the text.
Make sure you simplify the text until the end.

Figure 4: Base prompt for runs 1 and 2.

You are given two texts:
Original Text → contains the full meaning, details, and factual information.
Adapted Text → written at the [LEVEL] CEFR EFL level, but it loses important meaning from the original.
Your task is to rewrite the Original Text so that:
The meaning, facts, and order of events from the Original Text are fully preserved.
The grammatical structures and vocabulary complexity should match those in the Adapted Text. Use the Adapted Text as the reference for CEFR level.
Do not add new information or remove important details from the Original Text.
Do not copy sentences directly from the Adapted Text if they distort the meaning of the Original.
Keep proper names, places, times, and dates unchanged.
Omit any introductions, conclusions, or justifications. Provide only the simplified text.
Input:
Original Text: [ORIGINAL TEXT]
Adapted Text: [SIMPLIFIED TEXT]
Output:
Simplified version of the Original Text that preserves its meaning but matches the [LEVEL] CEFR EFL grammar and vocabulary style shown in the Adapted Text.

Figure 5: Follow-up prompt for run 1, which takes as input the output of the base prompt.

Please, make sure you remove all the introductions, conclusions and justifications related to the CEFR level and/or the task of text simplification from the following text, and do not introduce any new justifications or explanations regarding this instruction, but just reproduce the core text of the original: [SIMPLIFIED TEXT]

Figure 6: Follow-up prompt for run 2, which takes as input the output of the base prompt.

D Prompts for the ensemble strategy

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text.

Figure 7: Prompt 1 for run 3.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. To simplify you may substitute difficult words for simpler ones. You may break down long complex sentences into shorter ones. Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates. Keep the order of the events. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text. Make sure you simplify the text until the end.

Figure 8: Prompt 2 for run 3.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. To simplify you may use the CEFR [LEVEL] [DESCRIPTORS] and the [CEFR LEVEL] grammar structures in [GRAMMAR] . Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates. Keep the order of the events. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text. Make sure you simplify the text until the end.

Figure 9: Prompt 3 for run 3.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. To simplify you may use the CEFR [LEVEL] [DESCRIPTORS], the [CEFR LEVEL] vocabulary in [LIST] and the [CEFR LEVEL] grammar structures in [GRAMMAR]. Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates. Keep the order of the events. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text. Make sure you simplify the text until the end.

Figure 10: Prompt 4 for run 3.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. To simplify you may substitute difficult words for simpler ones. You may break down long complex sentences into shorter ones. Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates. Keep the order of the events. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text. Make sure you simplify the text until the end. Here are examples of two complex texts and their [CEFR LEVEL] simplified adaptations. Examples: [ORIGINAL TEXT 1] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 1]; [ORIGINAL TEXT 2] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 2]

Figure 11: Prompt 5 for run 3.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. To simplify you may use the CEFR [LEVEL] [DESCRIPTORS], the [CEFR LEVEL] vocabulary in [LIST] and the [CEFR LEVEL] grammar structures in [GRAMMAR]. Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates. Keep the order of the events. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text. Make sure you simplify the text until the end. Here are examples of two complex texts and their [CEFR LEVEL] simplified adaptations. Examples: [ORIGINAL TEXT 1] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 1] [ORIGINAL TEXT 2] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 2]

Figure 12: Prompt 6 for run 3.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. To simplify you may substitute difficult words for simpler ones. You may break down long complex sentences into shorter ones. Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates. Keep the order of the events. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text. Make sure you simplify the text until the end. Here are examples of two complex texts and their [CEFR LEVEL] simplified adaptations. Examples: [ORIGINAL TEXT 1] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 1]; [ORIGINAL TEXT 2] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 2]

Figure 13: Prompt 7 for run 3.

I'm a teacher of English as a Foreign Language (EFL) and I'm preparing a reading task. Please, simplify this text so that it can be understood by [LEVEL] CEFR EFL learners. To simplify you may use the CEFR [LEVEL] [DESCRIPTORS], the [CEFR LEVEL] vocabulary in n

d the [CEFR LEVEL] grammar structures in [GRAMMAR]. Make sure you keep the core content of the source text. Keep all the factual information, proper names (the names of people, places, etc.), times and dates. Keep the order of the events. Please avoid adding extra information to the text. Omit any introduction or conclusion. Only provide the simplified version of the text. Make sure you simplify the text until the end. Here are examples of two complex texts and their [CEFR LEVEL] simplified adaptations. Examples: [ORIGINAL TEXT 1] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 1]; [ORIGINAL TEXT 2] and its simplified [CEFR LEVEL] [SIMPLIFIED TEXT 2]

Figure 14: Prompt 8 for run 3.

follow-up prompt for PROMPT 4:
You are given two texts:
Original Text → contains the full meaning, details, and factual information.
Adapted Text → written at the [LEVEL] CEFR EFL level, but it loses important meaning from the original.
Your task is to rewrite the Original Text so that:
The meaning, facts, and order of events from the Original Text are fully preserved.
The grammatical structures and vocabulary complexity should match those in the Adapted Text. Use the Adapted Text as the reference for CEFR level.
Do not add new information or remove important details from the Original Text.
Do not copy sentences directly from the Adapted Text if they distort the meaning of the Original.
Keep proper names, places, times, and dates unchanged.
Omit any introductions, conclusions, or justifications. Provide only the simplified text.
Input:
Original Text: [ORIGINAL TEXT]
Adapted Text: [SIMPLIFIED TEXT]
Output:
Simplified version of the Original Text that preserves its meaning but matches the [LEVEL] CEFR EFL grammar and vocabulary style shown in the Adapted Text.

Figure 15: Prompt 9 for run 3.

follow-up prompt for PROMPT 4:
Please, make sure you remove all the introductions, conclusions and justifications related to the CEFR level and/or the task of text simplification from the following text, and do not introduce any new justifications or explanations regarding this instruction, but just reproduce the core text of the original: [SIMPLIFIED TEXT]

Figure 16: Prompt 10 for run 3.