STARLING at TSAR 2025 Shared Task: Leveraging Alternative Generations for Readability Level Adjustment in Text Simplification

Piotr Przybyła
TALN Group, Universitat Pompeu Fabra;
Institute of Computer Science, Polish Academy of Sciences
piotr.przybyla@upf.edu

Abstract

Readability adjustment is crucial in text simplification, as it allows to provide generations appropriate to the needs of a particular group of readers. Here we present a method for simplifying a text fragment that aims for a given CEFR level, e.g. A2 or B1. The proposed approach consists of two stages, executed for each sentence: (1) obtaining several candidate simplification through prompting a large language model and (2) choosing a candidate that maximises the compliance with the desired readability level. Several variants of our approach are evaluated within the framework of TSAR 2025 shared task, showing a trade-off between precise readability adjustment and faithful meaning preservation.

1 Background

Text simplification (TS) promises to make language content accessible to various vulnerable groups that might find typical text difficult, such as people with age-related impairments, learning difficulties, dyslexia, autism etc., but also non-native speakers (Shardlow, 2014; Paetzold and Specia, 2017). This causes an obvious challenge: it is impossible to satisfy the needs of all these groups with *the same* simplification. Some aspects of text might be difficult for one reader, but not for another one (Tamor, 1981). For example, even quite long and complicated words might appear straightforward for a non-native speaker, as long as they know similar lexemes from their mother tongue.

Therefore, text simplification can never be successful as a generic task and it needs to be anchored in a specific target audience. However, approaching the challenge for each group separately is impeded when taking a classic machine-learning approach. The datasets with simplifications prepared by human experts for specific readers are scarce (Cistola et al., 2021; Alarcon et al., 2023), so many TS solutions rely on generic data anyway. This limitation remains true for large language models (LLMs) – since they are unlikely to have been provided with examples of such tailored simplifications in their pretraining, they shouldn't be expected to generate them when prompted.

Here we present an approach to the problem of providing text simplifications for readers with a given language proficiency, expressed as a CEFR level. While there are some resources for the non-native speakers according to their level (Scarton et al., 2018), these are not sufficient for model training from scratch.

Therefore, our approach is a hybrid one. Firstly, we generate simplifications using generic LLM capabilities, but we ask the model to produce several variants of the output. Then, we use CEFR labellers to find the rephrasing that best corresponds to the desired level. This process is repeated for all sentences in a given text, and the final output is a concatenation of the simplified variants of all sentences.

2 Task

Our solution was prepared within the framework of the TSAR 2025 Shared Task on Readability-Controlled Text Simplification (Alva-Manchego et al., 2025). The task participants were provided with a series of pairs, each including a text fragment in English (paragraph-length) and a CEFR level, and their goal was to provide simplifications that maintain the meaning of the original fragment, but at the same time possess the provided difficulty level (A2 or B1).

The submitted fragments were evaluated using the following measures:

- Difference between the CEFR levels (requested and observed in text) according to three fine-tuned language models (Imperial et al., 2025), quantified as weighted F1, adjusted accuracy and RMSE.

Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 155–159

- Meaning preservation measured as text similarity (original and provided simplification) according to MeaningBERT (Beauchemin et al., 2023) and BERTScore (Zhang et al., 2020),
- Closeness to reference simplification, also according to MeaningBERT and BERTScore.

Of these, the RMSE CEFR and both MeaningBERT similarities were used to establish the final ranking.

For details regarding the previous work in the domain, evaluation details and results of all participants, see the overview article (Alva-Manchego et al., 2025).

3 Methods

The submitted solution was prepared using elements of STARLING (**S**implifying **T**ext **Acr**oss **L**anguages Us**in**g **G**enerative Models): a TS system under construction, which uses LLM prompting for obtaining robust simplifications across several languages, including some with low NLP support. The multilingual capabilities were however not used here, since the task is performed in English, a language with sufficient monolingual resources.

Broadly, the solution involves the following steps:

1. Splitting the input (complex) text into sentences,
2. For each sentence:
   - Producing variants of simplification using a prompted language model,
   - Selecting the variant that is the closest to the desired CEFR level,
3. Concatenating the obtained sentences.

3.1 Splitting

We split the given paragraph into individual sentences with LAMBO (Przybyła, 2022) 2.3 segmenter[1], using the `LAMBO-UD_English-EWT` model trained on English dependency parsing corpus[2] in Universal Dependencies (de Marneffe et al., 2021), version 2.13. The sentence splitting is motivated by (1) preliminary experiments showing that LLMs, when tasked with rephrasing a longer paragraph, often omit some of the details, and (2) the intention to obtain a wide range of candidates by reformulating sentences independently.

3.2 Simplifying

For simplification, we use the following prompt:

> Please rewrite the following complex sentence in order to make it easier to understand by non-native speakers of the language. You can do so by replacing complex words with simpler synonyms (i.e. paraphrasing), deleting unimportant information (i.e. compression), and/or splitting a long complex sentence into several simpler ones. The final simplified sentence needs to be grammatical, fluent, and retain the main ideas of its original counterpart without altering its meaning. Make sure the output is in the same language as the original.
> Return five different rephrasings, separated by newline. Do not generate any text except the reformulations.
> INPUT: ⟨input sentence⟩

This prompt is inspired by a formulation obtaining good results in BLESS benchmark (Kew et al., 2023) – prompt 2[3]. However, it was modified to include a request of five different rephrasings. Note how it does not include any mention of the desired CEFR level – our preliminary experiments showed it not to be useful for a general-purpose model pretrained without such specialised data.

Our baseline model was Gemma 3 (Gemma Team et al., 2025) 27 B, implemented on *HuggingFace Transformers* 4.38.1 (Wolf et al., 2020) (model `google/gemma-3-27b-it`). The computations were performed on double-GPU configuration with NVIDIA A100.

3.3 Choosing a variant

We create the list of variants by splitting the model output into individual options, removing anomalies (outputs shorter than 1/3 of the input[4]) and adding the original complex sentence. Each variant on the

[1] https://gitlab.clarin-pl.eu/syntactic-tools/lambo

[2] https://universaldependencies.org/treebanks/en_ewt/index.html

[3] Note that BLESS experiments were performed in a few-shot setting, but in the shared task no training data were available, so we use the same prompt in a zero-shot setting.

[4] These usually come from the LLM adding extra text to the output: enumeration markers, comments, etc.

method	CEFR compliance			Original similarity		Reference similarity	
	F1	Acc.	RMSE	M-BERT	BERT-S	M-BERT	BERT-S
Gemma-5v-best	0.5107	0.9500	0.8216	0.8075	0.9281	**0.7584**	**0.9124**
Gemma-5v-random	0.4895	0.8500	1.0000	**0.8204**	**0.9335**	0.7567	0.9109
Gemma-10v-best	**0.6921**	**0.9750**	**0.6325**	0.7675	0.9190	0.7458	0.9062
Gemma-10v-random	0.3888	0.9500	0.8803	0.7780	0.9194	0.7441	0.9038
original	0.1288	0.5250	1.6125	*1.0000*	*1.0000*	*0.7901*	*0.9265*

Table 1: Evaluation results on the trial dataset, showing CEFR compliance results, similarity to the original text and reference simplification, both measured using MeaningBERT (M-BERT) and BERTScore (BERT-S). The best values are in boldface; the submitted solutions are underlined.

list is then assessed with respect to readability level. Specifically, we apply the three CEFR labellers[5] provided by the organisers (section 2). We then check which of the variants (including the complex original) has been assigned the desired CEFR level with the highest probability, according to any for the labellers. In case of ties, the order on the variant list decides. Therefore, for uncertain ratings, the original complex sentence is used (to maintain meaning preservation).

3.4 Concatenation

After the CEFR-optimised variant is chosen for each sentence, they are all concatenated together to create continuous text. Note that this risks breaking some discourse links, which is a weakness of our approach (section 5).

4 Evaluation

The solution described above is our baseline approach to the problem. However, in order to better understand its strengths and weaknesses, we check the result of modifying some aspects:

- using a higher number of requested variants: 10 instead of 5.
- using a simpler heuristic for selecting a variant: random choice or keeping the original sentence.

All the solutions are tested using the evaluation code provided by the organisers (see section 2) on the 40 instances of the trial set. The full results on the test dataset, including our submission, can be found in the shared task overview article (Alva-Manchego et al., 2025).

Table 1 illustrates the results of the evaluation. We can see that the best CEFR compliance is achieved by generating 10 variants and choosing the one that corresponds to the desired level. This proves that the variant selection mechanism fulfils its purpose. However, the same approach is performing the poorest in content preservation. Instead, generating 5 variants and randomly selecting one of them delivers the best similarity to the original text (except returning the original itself). We can therefore see a clear trade-off between delivering the expected CEFR level and preserving the original content. Finally, our basic approach (5 variants, guided selection) achieves the best performance in terms of similarity to the reference simplification, indicating it overall usefulness. It is interesting to note that the original text is more similar to the reference simplifications than any of our approaches, indicating that the LLM is too aggressive in its rewriting. This mirrors a similar phenomenon in text rewriting for the purpose of a different task, namely adversarial example generation (Przybyła et al., 2025).

5 Discussion

Judging by the evaluation performed in the previous section, our approach performs well: it allows to adjust for the desired readability level without sacrificing too much of the original meaning. However, this is clearly a prototype solution and several limitations remain.

Firstly, a more intensive adjustment to readability level is possible. Since generating 10 variants allows better compliance than 5, one could also try 20 or 100 – though in case of most sentences it would be unrealistic to expect that many different reformulations. It is therefore an open question on

[5] https://huggingface.co/AbdullahBarayan/ModernBERT-base-doc_en-Cefr, https://huggingface.co/AbdullahBarayan/ModernBERT-base-doc_sent_en-Cefr and https://huggingface.co/AbdullahBarayan/ModernBERT-base-reference_AllLang2-Cefr2

whether this wouldn't cause a loss of meaning or general decrease in quality of further generations. We leave this for future work.

Secondly, we limited our prompting experiment to adapting a formulation that has been found to work well in previous research, but that usecase did not include readability adjustment. Therefore, we expect that experimenting with the prompt can help to guide the model towards producing reformulations with more utility for the current setup, e.g. by encouraging diveristy in the variants. Recent research in TS indicates great room for improvement in tuning the prompt (Guidroz et al., 2025).

Thirdly, we performed the simplification on the sentence level to obtain fine-grained control over the variants chosen and avoid the loss of details that we observe in full paragraph rewriting. However, this introduces a limitation: when each sentence is reformulated independently, we risk breaking discourse links between them, resulting in less coherent text. This is a known problem in text simplification, but it requires further research to deliver satisfactory solutions (Vásquez-Rodríguez et al., 2023). One bypass could be to provide whole paragraphs as input text and look for other ways to avoid information loss, e.g. through various prompts.

Finally, we have to emphasise that our evaluation is based solely on automatic measures, which for some time have been known to poorly reflect human judgement of simplified text (Alva-Manchego et al., 2021). This final step of TS evaluation should involve human evaluators, especially when a solution is claimed to be adjusted to specific user group.

6 Conclusions

To sum up, the proposed approach to readability-adjusted text simplification achieves positive results, delivering output that is in agreement with the desired CEFR level (95% accuracy) and maintains the overall meaning (93% BERTScore). Depending on which configuration we choose, a trade-off between simplicity and similarity to the original can be struck differently, but overall the results indicate this to be a promising direction for future investigation.

Acknowledgments

This work was done with the project IDEAL (Inclusive Democratic Engagement and Language Technologies in Europe), which received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101178191. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018019.

Lay Summary

When a text is changed to make it simpler for some readers, it is important to take into account who these readers are. For example, if the text is in English, but the readers are not native speakers, we should take into account their level of knowledge of the language. In this article we show how we simplify text in English so that learners at some level (for example A2 or B1) can understand it. First, we divide the text into separate sentences. Second, a language model changes each sentence, producing several possible modifications. We then choose the modification that is closest to the needed difficulty level. Finally, all the changed sentences are connected back together. Our approach is tested at the TSAR 2025 workshop. The results show that sometimes you have to choose between keeping the meaning unchanged and arriving at the difficulty level you want.

References

Rodrigo Alarcon, Lourdes Moreno, and Paloma Martínez. 2023. EASIER corpus: A lexical simplification resource for people with cognitive impairments. *PLOS ONE*, 18(4):1–23.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4):861–889.

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.

David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. MeaningBERT: assessing meaning

preservation between sentences. *Frontiers in Artificial Intelligence*, Volume 6 - 2023.

Giorgia Cistola, Mireia Farrús, and Ineke van der Meulen. 2021. Aphasia and acquired reading impairments: What are the high-tech alternatives to compensate for reading deficits? *International Journal of Language & Communication Disorders*, 56(1):161–173.

Marie-Catherine de Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

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.

Theo Guidroz, Diego Ardila, Jimmy Li, Adam Mansour, Paul Jhun, Nina Gonzalez, Xiang Ji, Mike Sanchez, Sujay Kakarmath, Mathias M J Bellaiche, Miguel Ángel Garrido, Faruk Ahmed, Divyansh Choudhary, Jay Hartford, Chenwei Xu, Henry Javier Serrano Echeverria, Yifan Wang, Jeff Shaffer, Eric, and 8 others. 2025. LLM-based Text Simplification and its Effect on User Comprehension and Cognitive Load. *Preprint*, arXiv:2505.01980.

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.

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*.

Gustavo H. Paetzold and Lucia Specia. 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593.

Piotr Przybyła. 2022. LAMBO: Layered Approach to Multi-level BOundary identification.

Piotr Przybyła, Euan McGill, and Horacio Saggion. 2025. Attacking Misinformation Detection Using Adversarial Examples Generated by Language Models. *Preprint*, arXiv:2410.20940.

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. Text Simplification from Professionally Produced Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation ({LREC} 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Matthew Shardlow. 2014. A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications*, 4(1).

Lynne Tamor. 1981. Subjective Text Difficulty: An Alternative Approach to Defining the Difficulty Level of Written Text. *Journal of Reading Behavior*, 13(2):165–172.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia.