HOPE at TSAR 2025 Shared Task: Balancing Control and Complexity in Readability-Controlled Text Simplification

Sujal Maharjan
Taylor's University
Subang Jaya, Malaysia
sujalmaharjan007@gmail.com

Astha Shrestha
Taylor's University
Subang Jaya, Malaysia
aasthashrestha688@gmail.com

Abstract

This paper describes our submissions to the TSAR 2025 Shared Task on Readability-Controlled Text Simplification. We present a comparative study of three architectures: a rule-based `Baseline`, a heuristic-driven `Expert` system, and a zero-shot generative T5 `Pipeline` with a semantic guardrail. Our analysis of the official results shows a clear trade-off between the controllability of rule-based systems and the fluency of generative models. We detect that in this zero-shot instance, our simpler, confined systems achieved superior meaning preservation scores compared to the powerful but less predictable generative model. We present a diagnostic failure analysis centered in our actual system outputs, illustrating how different architectural choices result distinct error patterns, such as under-simplification, information loss via heuristics, and semantic drift.

1 Introduction

The Shared Task on Readability-Controlled Text Simplification at the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025) (Alva-Manchego, Fernando et al., 2025) requires systems to simplify a given text to a specified Common European Framework of Reference (CEFR) level while preserving meaning. This creates a conflict between reducing linguistic complexity and maintaining semantic fidelity.

To evaluate the trade-off between generative power and controllability, we engineered three systems: a deterministic `Baseline`, a heuristic-driven `Expert` system, and a zero-shot `Generative Pipeline` based on T5. Our core empirical observation is that in this zero-shot setting, constrained and interpretable approaches perform better than a powerful generative pipeline on official meaning preservation metrics. We explore the specific failure modes of each system, illustrating a definite trade-off between generative power and semantic controllability, and propose directions for future hybrid approaches.

2 Related Work

Text Simplification (TS) has transformed from early rule-based systems primarily focused on lexical and syntactic transformations (Siddharthan, 2014) to the current paradigm which is dominated by neural sequence-to-sequence models (Nisioi et al., 2017). Models like T5 (Raffel et al., 2019) and BART (Lewis et al., 2020), pre-trained on vast text corpora, have become the de facto standard, achieving state-of-the-art fluency when fine-tuned on task-specific data.

However, a key challenge in modern TS is **controllability** (Maddela et al., 2021). While large language models are proficient at fluent paraphrasing, guiding them to simplify to a precise readability level without sacrificing semantic fidelity remains an open problem. Researchers have explored techniques like explicit target-level prompting, but models can still "hallucinate" or deviate from the underlying meaning. Our work directly address the issue, questioning whether a powerful generative model utilized in a zero-shot setting is the effective tool for a task with strict semantic constraints, echoing findings in other domains where simpler models can be more robust (Rudin, 2019).

3 System Descriptions

We implemented three systems with increasing complexity to explore the trade-off between control and generative power.

3.1 System 1: Baseline

Our baseline is a deterministic pipeline serving as a high-precision, low-recall benchmark. It performs three operations: (1) lowercasing the input text, (2) applying a curated 10-word substi-

Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 261–265

tution dictionary (e.g., `'approximately' -> 'about'`), and (3) capitalizing the first letter of the output. This system performs minimal, safe edits designed to maximize meaning preservation.

3.2 System 2: Expert

The Expert system extends the baseline with two key features. First, it uses a more intricate, two-tiered substitution lexicon of approximately 100 entries, governed by resources such as the English Vocabulary Profile (EVP) to ensure CEFR-appropriateness. For lower proficiency targets (A1/A2), a more robust set of substitutions is used (`e.g., 'substantial' -> 'big'`). For higher levels (B1+), a more conservative lexicon is utilized to maintain nuance.

Second, it implements a heuristic for structural simplification: for texts targeted at A1/A2 levels exceeding 50 words, the system truncates the output to the first three sentences. This rule is a direct, interpretable method to regulate output length, a crucial aspect of lower-level texts, though it comes with the risk of information loss.

3.3 System 3: Generative Pipeline

This multi-stage pipeline was used in a zero-shot setting, as the shared task provided no official training data. The stages are:

1. **Lexical/Syntactic Preprocessing:** The input text is first simplified utilizing the same non-truncating rules as the Expert system.
2. **CEFR-Aware Prompting:** A T5-base model is guided by a dynamic instructional prompt. For example, to simplify a text for a B1 target, the prompt is: 'Simplify this text using clear language for intermediate level: [original text]'.
3. **Semantic Guardrail:** An embedding-based check is performed. We compute the cosine similarity between the original and T5-generated text embeddings using the 'all-MiniLM-L6-v2' model. If the similarity is below a threshold of 0.75, the T5 output is rejected, and the system reverts to the preprocessed text from stage (1). This mechanism is a countermeasure against significant semantic drift.

4 Experimental Setup

Data: We present results on the official TSAR 2025 test set (200 instances; targets A2/B1). The human-simplified 'reference' texts were employed in the official scoring and in our diagnostic analysis.

Official Metrics: We present the official AUTORANK composite score and its components as provided by the organizers.

- **AUTORANK:** Official composite metric used by the shared task (lower is better).
- **MB-orig/ref:** MeaningBERT score against the original and reference texts, respectively (higher is better).
- **RMSE:** Root-mean-square error for CEFR level prediction (lower is better).

Implementation: Appendix A contains reproducibility notes.

5 Results and Analysis

Table 1: Official final results on the TSAR 2025 test set. The official AUTORANK score is a composite metric where lower is better. Best scores for each metric are in bold.

System	RMSE ↓	MB-orig ↑	MB-ref ↑	AUTORANK ↓
`Baseline`	1.428	**0.945**	**0.815**	**12.230**
`Expert`	**1.402**	0.919	0.797	13.260
`Generative Pipeline`	1.600	0.841	0.730	19.030

Table 1 presents the official final scores for our three systems. The results demonstrate a clear pattern: while the `Expert` system achieved the best CEFR compliance (lowest RMSE), the simpler `Baseline` system was superior on both meaning preservation metrics (MB-orig, MB-ref) and, consequently, the final AUTORANK composite score. The `Generative Pipeline` performed worst across all official metrics.

5.1 Diagnostic Failure Analysis

To understand the trade-offs revealed by these scores, we performed a diagnostic failure analysis by comparing system output against the original text and the human-written reference. Table 2 provides a representative example that illustrates the distinct failure modes of each system.

Failure Spectrum: Our analysis confirms different failure modes. The `Baseline` makes safe but inadequate modifications. The `Expert` system's aggressive heuristics improve readability metrics but can severely hamper fidelity. The `Generative Pipeline`, despite its potential for fluency, is vulnerable to semantic drift through "content omission", a common failure mode for

Table 2: Diagnostic failure analysis for a representative example (`text_id: 22-a2`, Target: A2).

System & Input	Output Text (Actual)	Error Type and Analysis
Original	One of the most interesting cases of wild animals living in a city are the wild dogs of Moscow... They have learned that it is safer to cross the street with people...	A descriptive, multi-fact text of 97 words.
Human Reference	There are many wild dogs in Moscow, about 35,000. Some were born wild, and some were pets... They found that the dogs have learned how to live in the city...	Simplifies vocabulary and rephrases for clarity while retaining all key facts. Serves as the gold standard for simplification.
`Baseline`	One of the most interesting cases... Moscow. In Moscow, there are about 35,000 wild dogs... they have learned that it is safer to cross the street...	**Under-simplification:** Only performed lexical substitution (`'approximately' -> 'about'`). Failed to simplify sentence structure. This explains its high MB-orig score (0.945) but lower MB-ref score (0.815).
`Expert`	One of the most interesting cases... pets that have been abandoned by their owners.	**Information Loss via Heuristic:** Our sentence truncation rule for long A2 texts deleted all facts about the dogs' adaptation to urban life. This aggressive rule explains the system's improved RMSE (1.402) but also its drop in meaning preservation scores.
`Generative Pipeline`	One of the most interesting cases of wild animals living in a city are the wild dogs of Moscow. In Moscow, there are around 35,000 wild dogs living on the streets.	**Content Omission:** The T5 model fluently paraphrased the first two sentences but omitted all subsequent information about the dogs' origins and learned behaviors. This explains why it received the worst scores on all meaning preservation metrics (MB-orig: 0.841, MB-ref: 0.730).

zero-shot models that we incorrectly labeled "hallucination" in a preliminary analysis. This confirms the reviewer's observation and highlights the difficulty of ensuring factual consistency with generative models.

6 Semantic Guardrail: Design and Limitations

We implemented an embedding-cosine guardrail (model: 'all-MiniLM-L6-v2'; cutoff = 0.75). Its advantages are speed and ease of deployment, and it successfully screens out many gross semantic deviations. However, our analysis reveals its limitations: cosine similarity is an imperfect proxy for entailment. As seen in the example in Table 2, a text can remain topically similar while missing critical facts, a subtlety that the guardrail can miss. This indicates that thresholds are dataset-specific and that more robust verification techniques are needed for high-stakes applications.

7 Conclusion

Our comparative analysis on the TSAR 2025 Shared Task shows a clear trade-off between control and fluency in zero-shot readability-controlled text simplification. Our findings empirically illustrates that in the absence of fine-tuning data, simpler, interpretable approaches can be more robust for semantic fidelity. A simple `Baseline` preserved meaning but was insufficient for structural

simplification. A heuristic-driven `Expert` system enhanced readability metrics but caused significant information loss. Finally, a `Generative Pipeline` offered fluent paraphrasing but was highly vulnerable to content omission, resulting in the lowest meaning preservation scores. This implies that for tasks with stringent semantic constraints, the controllability of simpler systems provides a distinct advantage.

Limitations and Future Work

Limitations: The primary limitation of this study is the lack of a large-scale human evaluation to confirm whether the automated metrics, including the official AUTORANK, fully align with human judgments of simplification quality. While our analysis uses human-written references for scoring, it does not include direct human ratings of our systems' outputs. Furthermore, our analysis revealed that the embedding-based guardrail, while effective at catching major deviations, is an imperfect proxy for fine-grained semantic fidelity.

Future Work: We propose two main directions. First, exploring hybrid systems that balance control and fluency, for instance, through lexicon-constrained decoding to guide generative models away from factual errors. Second, developing stronger, automated fidelity checks. Our analysis showed that cosine similarity can be insufficient; future work should investigate using Natural Language Inference (NLI) or Question-Answering (QA) models to verify the preservation of key facts. Validating these more advanced automated metrics against targeted human evaluation will be a critical next step for the field.

Lay Summary

Making complicated text easier to read is important for everything from education to making public information more accessible. This process is called text simplification. The challenge is not just to make text simpler, but to simplify it for a specific reading levellike for a beginner versus an intermediate learner—without changing the original meaning.

Scientists use different tools for this task. Some use simple, strict rules, like swapping a hard word for an easy one. Others use powerful Artificial Intelligence (AI) models, similar to ChatGPT, which can fluently rewrite entire sentences. We wanted to find out which approach works best for this controlled simplification task, especially when there is no specific training data available. Is the most powerful AI always the best choice when preserving the original meaning is critical?

To answer this, we built and compared three systems: a `Baseline` system with just a few word-swapping rules, a smarter `Expert` system with more rules (including one to shorten long texts), and a powerful AI `Generative Pipeline`. Our study found that the simpler, rule-based systems were surprisingly better at keeping the original meaning of the text. The powerful AI, while often producing fluent sentences, made significant errors by deleting important informationa problem we call 'content omission'. Our `Expert` system also lost information when its rule to shorten long texts was too aggressive. The safest system was the simplest `Baseline`, which made only minor changes but never altered the core message.

Our findings are important for developers building tools for education and accessibility. They show that for tasks where accuracy is crucial, relying on simple, predictable rules can be more reliable than using a complex AI that you can't fully control. The best path forward may be to create hybrid systems that combine the safety of rules with the fluency of modern AI.

References

Alva-Manchego, Fernando, Stodden, Regina, Imperial, Joseph Marvin, Barayan, Abdullah, North, Kai, and Tayyar Madabushi, Harish. 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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. 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.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Preprint*, arXiv:1811.10154.

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

A Reproducibility Notes

A.1 Code and Data Availability

To facilitate full replication, the complete source code for all three systems, the final system outputs, and the analysis scripts utilized in this paper are publicly available in a GitHub repository under an MIT License. The repository can be accessed at: `https://github.com/SUJAL390/hope-tsar-emnlp2025`

A.2 Dependencies and Hyperparameters

- **General:** Python 3.8+, `numpy`, `pandas`.
- **Libraries:** `transformers` (v4.55+), `evaluate`, `scikit-learn`, `torch`.
- **T5 decoding:** model `t5-base`; decoding parameters: temperature = 0.7, top-p = 0.9.
- **Semantic guardrail:** model `sentence-transformers/all-MiniLM-L6-v2`; cutoff = 0.75.

A.3 Recommended Guardrail Validation

To validate a guardrail's effectiveness, one should sample N accepted and N rejected outputs, annotate them for meaning preservation (binary) and then compute precision and recall. We recommend N=100 for an initial check.