Improving Estonian Text Simplification through Pretrained Language Models and Custom Datasets

Eduard Barbu
Institute of Computer Science
University of Tartu
Tartu, Estonia
eduard.barbu@ut.ee

Meeri-Ly Muru
Institute of Computer Science
University of Tartu
Tartu, Estonia
meerimuru@gmail.com

Sten Marcus Malva
Institute of Computer Science
University of Tartu
Tartu, Estonia
sten.malva@gmail.com

Abstract

This paper presents a method for text simplification based on two neural architectures: a neural machine translation (NMT) model and a fine-tuned large language model (LLaMA). Given the scarcity of existing resources for Estonian, a new dataset was created by combining manually translated corpora with GPT-4.0-generated simplifications. OpenNMT was selected as a representative NMT-based system, while LLaMA was fine-tuned on the constructed dataset. Evaluation shows LLaMA outperforms OpenNMT in grammaticality, readability, and meaning preservation. These results underscore the effectiveness of large language models for text simplification in low-resource language settings. The complete dataset, fine-tuning scripts, and evaluation pipeline are provided in a publicly accessible supplementary package to support reproducibility and adaptation to other languages.

1 Introduction

Text simplification converts complex text into simpler forms while preserving meaning. Typical operations include sentence splitting, shortening, lexical substitution, and syntactic simplification. Simplified texts help language learners, people with cognitive impairments, and readers with low literacy, and are useful in education and public communication (Alva-Manchego et al., 2020; Shardlow, 2014a).

Automatic text simplification (ATS) uses NLP models to perform these edits. Proprietary LLMs (e.g., GPT-4.0, Claude) can do zero/few-shot simplification, but often fail to reach the lowest readability levels while preserving meaning (Barayan et al., 2025) and raise concerns about transparency and long-term reproducibility.

Open, language-specific systems are an alternative: they allow control over training data and domain goals, and avoid dependence on commercial APIs—especially important for low-resource languages like Estonian.

We study two architectures for Estonian ATS: OpenNMT (simplification as monolingual MT) and a fine-tuned LLaMA model trained on a newly built Estonian Simplification Dataset. We complement automatic evaluation using BLEU and SARI scores with human evaluation by native speakers. **Contributions.**

- A new Estonian Simplification Dataset combining translated, GPT-4.0–generated, and manually validated pairs.
- A head-to-head comparison of OpenNMT and fine-tuned LLaMA for Estonian.
- Evidence from human ratings that LLaMA trained on Estonian data outperforms OpenNMT.
- Released scripts/configs for easy adaptation to other low-resource languages.

The remainder of the paper is organized as follows. Section 3 describes the Estonian simplification dataset, Section 4 benchmarks English models and motivates OpenNMT as a baseline, Section 5 details LLaMA fine-tuning, and the final sections present evaluation and conclusions.

2 Related Work

Research on ATS has been dominated by English, with surveys charting the progression from handcrafted rules to neural models and large-scale pretraining (Saggion, 2017; Espinosa-Zaragoza et al., 2023; Shardlow, 2014b; Paetzold and Specia, 2016).

Rule-based approaches. Early work relied on handcrafted rules for lexical substitution and grammatical restructuring. Foundational systems by Chandrasekar et al. (1996) and Inui et al. (2003)

Proceedings of Recent Advances in Natural Language Processing, pages 133–142
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-016

introduced syntactic transformations to aid reading comprehension. These methods were interpretable but brittle and difficult to scale across domains.

User-group oriented projects. Several initiatives highlighted the value of tailoring simplification to specific audiences. The HAPPI project simplified dialogue systems for people with aphasia (Devlin and Unthank, 2006). PorSimples targeted Brazilian Portuguese for readers with low literacy (Aluisio et al., 2010). The FIRST project adapted texts for individuals with autism spectrum disorder and cognitive disabilities (Barbu et al., 2015; Štajner and Saggion, 2018), while CLEAR focused on supporting second-language learners (Gala et al., 2020). These projects emphasized that simplification is not one-size-fits-all, but must reflect user needs and linguistic context.

Neural models. The rise of deep learning shifted ATS toward Seq2Seq architectures with attention. The DRESS model (Zhang and Lapata, 2017) pioneered reinforcement learning for simplification, optimizing grammaticality, meaning preservation, and simplicity simultaneously. Later work integrated semantic parsing (Zhao et al., 2018), while controllable models introduced explicit knobs for simplification strength (Sheang and Saggion, 2021; Martin et al., 2020). These methods improved fluency and structural edits compared to rule-based systems.

Pretrained language models. Transformer-based pretrained models pushed performance further. T5 (Raffel et al., 2020b) framed simplification as a text-to-text problem, enabling broad transfer. Proprietary models like GPT-4 (OpenAI, 2023) achieve strong zero/few-shot results, and in-context learning methods extend this flexibility (Agrawal et al., 2023). However, they remain closed-source, expensive, and less transparent. Open-access LLaMA models (Touvron et al., 2023) offer a promising alternative: large multilingual pretraining combined with accessibility for fine-tuning in low-resource languages.

Datasets. English ATS benefited from multiple corpora. For example, WikiSmall (Zhu et al., 2010) aligned Wikipedia with Simple English Wikipedia. TurkCorpus (Xu et al., 2016) provided multiple human simplifications per sentence, improving evaluation robustness. Newsela (Xu et al., 2015) offered professional multi-level rewrites, though restricted to licensed access. Unsupervised methods (Alva-Manchego et al., 2021; Jiang et al., 2020) explored automatic sentence alignment and pseudo-parallel generation to overcome data scarcity.

Low-resource languages. Outside English, ATS research is limited. For Latvian, Virk et al. (2021) emphasized lack of corpora and tools. Slovenian work (Štajner et al., 2022) similarly identified data bottlenecks. Estonian ATS has so far been confined to two undergraduate theses: one on WordNet-based lexical substitution and another on template-driven syntactic rules. No large-scale, publicly available dataset previously existed.

While ATS has matured for English, low-resource languages remain underserved. We address this gap by releasing the first large Estonian simplification dataset and benchmarking OpenNMT against fine-tuned LLaMA, showing that reproducible ATS is feasible for underrepresented languages.

3 Building the Estonian Simplification Dataset

In this section, we describe the construction of the Estonian Simplification Dataset. The goal was to create a resource large and diverse enough to train deep neural models. Building such a dataset from scratch would have been prohibitively time-consuming, given the absence of prior Estonian simplification data.

We aligned the dataset with English corpora and established clear guidelines for Estonian-specific phenomena. This dual alignment with international simplification efforts and local linguistic norms ensures both relevance and linguistic grounding.

The first source for our dataset is the Turk corpus (Xu et al., 2016), which contains 2,359 original Wikipedia sentences, each accompanied by eight simplified versions crowdsourced via Amazon Mechanical Turk. This corpus captures a variety of simplification strategies, including lexical substitution, paraphrasing, and sentence restructuring. We manually translated relevant portions of this corpus into Estonian, but only retained sentence pairs where a clear simplification relationship was preserved after translation. In some cases, simplification evident in the English original did not carry over naturally into Estonian due to syntactic or lexical differences, and such pairs were excluded.

The second source is the Wikipedia Data Set 2.0 (Kauchak, 2013), which includes 167,689 aligned sentence pairs between standard and Simple English Wikipedia. Small subsets of this dataset were machine-translated into Estonian and subsequently corrected by native annotators. However, despite this effort, the resulting translations were often semantically inconsistent or lacked coherence, rendering the dataset less useful for training.

To address these challenges and ensure linguistic consistency, we developed the Estonian Simplification Guidelines, partly inspired by the methodology and goals of the Newsela Corpus (Xu et al., 2015). Newsela includes professionally edited news articles rewritten at multiple reading levels. Although this corpus is not openly available, it provides a valuable reference for developing high-quality simplification guidelines.

The simplification guidelines consist of a detailed 10-page document that outlines specific strategies adapted to the morphosyntactic properties of Estonian. The guidelines concentrate on three main areas:

- **Grammatical complexity:** Simplifying inflectional morphology, verb tense usage, and mood markers.
- **Syntactic complexity:** Splitting long or embedded clauses, simplifying coordination, and reordering sentence constituents for clarity.
- **Lexical simplification:** Replacing low-frequency words with more common synonyms, using hypernyms, and avoiding domain-specific jargon.

Following best practices in prior English-language simplification work, the primary corpus source for large-scale simplification is Estonian Wikipedia. Articles were extracted from the latest Wikipedia dumps, segmented into sentences, and filtered to include only those longer than 15 words.

To generate a large volume of simplification examples efficiently, we employed GPT-4.0 (OpenAI, 2023), a state-of-the-art proprietary language model that ranks highly on the Chatbot Arena LLM Leaderboard. Although relying on GPT-4.0 incurs some costs, these are significantly lower than the expense of hiring human annotators for full-scale manual simplification.

GPT-4.0's flexibility allows it to assume tailored roles and styles through prompting. We leveraged this feature by designing templates that embedded persona-driven behavior to guide simplification style and focus. Prior research has shown that persona-based prompting outperforms general prompting in specific tasks (Pataranutaporn et al., 2021; Wang et al., 2024).

Throughout development, we tested multiple prompting strategies across GPT-4.0 instances. Initial templates used no persona framing, while later iterations embedded role-based behaviors targeting syntactic or lexical operations. After extensive experimentation, the most effective results came from sequential prompting: first performing lexical simplification, followed by syntactic restructuring. Although recent studies have questioned the general effectiveness of persona-based prompting (Zheng et al., 2024), in our task setting, the agent-style prompts proved superior. By agent-style prompts, we refer to role-based instructions instantiating a simplification assistant, such as: "You are a lexical simplification assistant tasked with simplifying Estonian text."

The breakdown of simplified sentences generated using GPT-4.0 is shown in Figure 1. Out of 47,112 simplified sentence pairs, 28,479 were produced using our initial generic prompt, and 18,633 were generated using agent-style prompts for lexical and syntactic tasks.

3,992 Articles Simplified

↓

47,112 Simplified Sentence Pairs

↙ ↘

28,479 **18,633**

↓

9,913 Sentence Pairs Corrected by Annotators

Figure 1: Breakdown of Estonian Dataset Pairs

Below is an excerpt from the lexical simplification agent used in our GPT-4.0 prompts (see 3). The agent uses a few-shot learning strategy with example-based guidance to consistently apply Estonian lexical simplification techniques.

Prompt Example: Lexical Simplification Agent

Instruction:
You are a lexical simplification assistant tasked with simplifying Estonian text. Your role is to receive Estonian sentences and transform them by simplifying complex words and phrases while maintaining the original meaning.

Lexical Simplification Guidelines:

- Replace difficult words with simpler,

more common alternatives.
- Carefully consider context to preserve the original meaning.
- Avoid simplifying proper nouns, technical terms without simpler equivalents, and words essential to the sentence's meaning.
- Remove adjectives that do not add critical meaning.
- Use simple, common verbs.
- Prefer simpler tenses to enhance clarity.

Persona:
You are an expert in lexical simplification, focused on making Estonian text more accessible. Your approach balances reducing complexity with maintaining meaning, ensuring the text remains clear and true to the original.

Example:
Original: Epidemioloogia uurib nakkushaigusi ja nende tõkestamist.
Simplified: Epidemioloogia uurib nakkushaigusi ja nende peatamist.

The final dataset statistics are presented in Table 1. In total, we compiled 50,416 sentence pairs across three primary sources. This corpus forms the backbone of our experiments and is used to train the models presented in the following sections.

Source	Sentence Pairs
Turk	1,896
Wikipedia Data Set 2.0	1,408
GPT-4.0	47,112
Total Pairs	**50,416**

Table 1: Composition of the Estonian Simplification Dataset

This dataset enables the fine-tuning of deep neural network architectures tailored for Estonian text simplification, as described in the following sections.

4 Benchmarking Simplification Models for English Language

To select an appropriate baseline model for Estonian text simplification, several established English simplification systems with publicly available code were benchmarked. The evaluation protocol followed the methodology outlined by Zhang and Lapata (2017).

The first system evaluated applies machine translation (MT) techniques to the simplification task. While MT systems are traditionally designed to translate across languages, their sequence-to-sequence (Seq2Seq) architectures, including both recurrent and transformer-based models, can be repurposed to translate complex sentences into simpler ones in the same language.

One foundational model is built using OpenNMT (Klein et al., 2018), a widely used neural MT framework. The text simplification system described in Nisioi et al. (2017) is implemented with an encoder-decoder architecture, typically using Recurrent Neural Networks (RNNs) such as LSTMs or GRUs. An attention mechanism is integrated to help the decoder focus on the most relevant parts of the input sentence.

OpenNMT-py, the PyTorch-based implementation, was selected for its flexibility and ease of integration. The configuration used in the experiments included:

- Two encoder-decoder layers with 500 hidden units per layer,
- A dropout rate of 0.3 to reduce overfitting,
- Fifteen training epochs with Stochastic Gradient Descent (SGD) and a learning rate of 0.1.

The second model evaluated is DRESS (Zhang and Lapata, 2017), a simplification system trained using reinforcement learning. Unlike standard supervised learning approaches, DRESS optimizes simplification quality through a reward-based signal, encouraging a balance between grammaticality, meaning preservation, and simplicity. A pretrained encoder-decoder model is used, and reinforcement learning is applied to iteratively improve output quality based on the designed reward function.

The third model is T5 (Text-To-Text Transfer Transformer) (Raffel et al., 2020c), a unified framework in which all NLP tasks are cast as text generation problems. T5 was pretrained on the C4 dataset (Raffel et al., 2020a) and has demonstrated strong generalization capabilities. It uses a transformer-based encoder-decoder architecture and supports task-specific prompting.

For the simplification task, T5 was fine-tuned over five epochs using the Adam optimizer. Input sentences were prefixed with the instruction "simplify." Control tokens encoding structural properties (e.g., character length ratio, Levenshtein distance) were incorporated to enhance simplification quality.

All models were trained and evaluated on the WikiSmall corpus (Zhu et al., 2010), which consists of aligned sentence pairs from English Wikipedia and Simple English Wikipedia. The training set includes 89,042 sentence pairs, while the test set comprises 100 pairs.

Three standard metrics were used for evaluation:

- **BLEU** (Papineni et al., 2002), which measures n-gram overlap between system output and reference simplifications. While informative, BLEU does not reward deletion or simplification explicitly.
- **SARI** (Xu et al., 2016), which evaluates the system output against both input and reference sentences, rewarding appropriate additions, deletions, and retention.
- **FKGL** (Flesch-Kincaid Grade Level) (Flesch, 1948), which estimates readability by calculating sentence length and syllable count. Lower scores correspond to simpler and more accessible text.

Evaluation results are shown in Table 2.

Model	BLEU	SARI	FKGL
DRESS	**47.93**	13.61	11.35
OpenNMT	44.61	35.82	9.97
T5	30.88	**41.21**	**7.23**

Table 2: Evaluation of the simplification models

The results indicate that T5 achieved the highest SARI and lowest FKGL scores, suggesting superior simplification quality and readability. However, its BLEU score was lower, indicating less lexical overlap with the reference. DRESS attained the highest BLEU score but scored poorly on SARI and FKGL, reflecting a tendency to copy input sentences with minimal simplification. OpenNMT demonstrated balanced performance across all metrics, excelling particularly in structure simplification and lexical reduction.

Given its robustness and ease of deployment, OpenNMT was selected as the baseline model for Estonian text simplification, especially considering its suitability for low-resource scenarios and independence from large-scale pretraining.

5 Simplification with Fine-Tuned LLaMA

LLaMA (Touvron et al., 2023) is a family of open-access large language models developed by Meta, designed to offer high performance while maintaining computational efficiency. Unlike some of the larger proprietary models, LLaMA models are optimized for accessibility and can be fine-tuned and deployed in resource-constrained environments. These models are trained on diverse, publicly available multilingual corpora and support a broad range of NLP tasks, including generation, summarization, translation, and simplification.

LLaMA 3.1, introduces enhanced efficiency and accuracy compared to previous versions. It is available in various parameter configurations, with sizes up to 65 billion. The training data includes multiple languages, and while the exact sources for Estonian remain undisclosed, it is known that Estonian was included in the pretraining corpus. This multilingual pretraining provides a suitable foundation for adapting the model to Estonian-specific tasks.

To enable sentence-level simplification in Estonian, both LLaMA 3.0 and LLaMA 3.1 with 8 billion parameters were fine-tuned using the Estonian Simplification Dataset (Section 3). The fine-tuning process was organized into the following steps:

1. **Pretrained Model Initialization**: The LLaMA models and tokenizer were loaded from their base checkpoints. Due to their prior exposure to multilingual content, these models serve as strong initializations for low-resource language tasks such as Estonian simplification.

2. **Integration of the Unsloth Library**: The Unsloth library[1] was used to optimize fine-tuning. This framework leverages Low-Rank Adaptation (LoRA) and QLoRA to reduce memory consumption and computational cost. It also supports mixed-precision training and efficient memory allocation, making it well suited for fine-tuning LLaMA models on modern GPUs.

3. **Training Configuration**: Hyperparameters were adjusted to balance efficiency and convergence. Key parameters included learning

[1] https://github.com/unslothai/unsloth

rate, batch size, and total number of steps. Mixed precision (FP16) and gradient accumulation were employed to stabilize training. The resulting model learned to generate simplified Estonian sentences while preserving meaning and grammaticality.

After training, both the fine-tuned model and tokenizer were serialized for downstream use in inference pipelines. The resulting model serves as an Estonian-specific simplification engine that is fully self-hosted and independent of external proprietary APIs.

6 Evaluation

The evaluation of the Estonian simplification models was conducted using a combination of automatic metrics and human judgments. This dual approach ensures quantitative and qualitative assessment, since automatic metrics alone often miss fluency, meaning, and simplification quality.

A subset of 100 complex sentences was randomly sampled from the Estonian Simplification Dataset. These sentences were manually simplified by two native Estonian linguists to form a high-quality gold standard. The resulting parallel dataset was then used to evaluate the performance of both the OpenNMT and LLaMA-based models.

6.1 Automatic Evaluation

To address the scarcity of Estonian simplification data, the English WikiSmall corpus was translated into Estonian using the Tartu NLP machine translation service (Korotkova and Fishel, 2024). This enabled training the OpenNMT model entirely on Estonian text, albeit derived from English simplification pairs, resulting in a larger and more diverse dataset than the Estonian Simplification Dataset.

For LLaMA, both the 3.0 and 3.1 versions were fine-tuned on the Estonian Simplification Dataset, excluding the 100-sentence test set. Qualitative analysis showed that LLaMA 3.1 consistently outperformed its predecessor, and it was therefore selected for final evaluation.

Training for OpenNMT was performed using the OpenNMT-py library with a shared vocabulary of 60,000 tokens. Vocabulary sharing was enabled to improve token alignment between source and target. The training process saved 12 checkpoints, and the best-performing checkpoint (based on dev loss) was selected to avoid overfitting.

For LLaMA 3.1, fine-tuning was conducted over 500 training steps, with a warmup of 100 steps. A batch size of 8 with gradient accumulation over 2 steps was used. The AdamW optimizer was configured with a learning rate of 0.00005, weight decay of 0.01, and cosine learning rate scheduling. Mixed precision (FP16) training was used to accelerate computation and reduce memory usage. A fixed seed (42) was applied for reproducibility. The resulting model was serialized and deployed for testing.

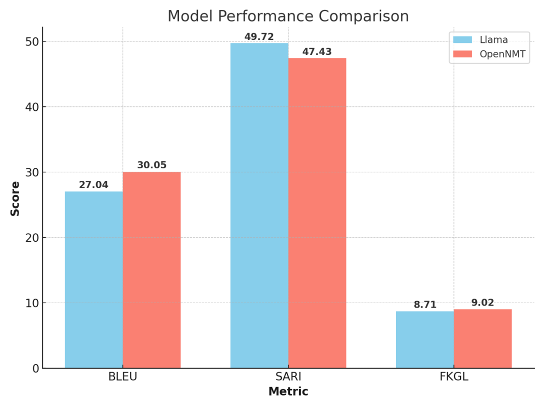

Figure 2: Comparison of Model Performance on BLEU, SARI, and FKGL metrics for LLaMA 3.1 and OpenNMT

Evaluation was carried out using three standard metrics:

- **BLEU** (Papineni et al., 2002): Measures n-gram overlap between model output and reference simplifications. It rewards fluency but penalizes deletion.
- **SARI** (Xu et al., 2016): Designed specifically for simplification, this metric compares system output to both the input and reference, rewarding appropriate edits.
- **FKGL** (Flesch, 1948): The Flesch-Kincaid Grade Level estimates the readability of the generated output.

As shown in Figure 2, OpenNMT achieved a higher BLEU score (30.05 vs. 27.04), indicating stronger n-gram similarity with the reference texts. However, LLaMA 3.1 outperformed OpenNMT on SARI (49.72 vs. 47.43), suggesting it made more effective simplification edits. The FKGL scores (8.71 for LLaMA 3.1 vs. 9.02 for OpenNMT) show that both systems improved readability, with a slight advantage for LLaMA. These mixed results highlight the limitations of relying solely on automated metrics.

6.2 Manual Evaluation

To complement the automatic evaluation, a human assessment was conducted by two independent native Estonian linguists. Fifty randomly selected sentence pairs, simplified by each model, were rated across four distinct dimensions: grammaticality (G), readability (R), preservation of meaning (M), and reduction effort (R_{eff}). Each criterion was scored on a 0–4 scale, with 4 denoting excellent performance and 0 indicating a complete failure. The rating guidelines were collaboratively developed and piloted to ensure reliability. Annotator disagreements were resolved through discussion to produce consensus scores.

Grammaticality (G) assesses the syntactic correctness of the output. A score of 4 corresponds to a grammatically flawless sentence with native-level fluency. Lower scores indicate increasingly frequent or severe errors, including agreement mismatches, fragmentary structures, or incoherent phrasing.

Readability (R) measures how easy the simplified sentence is to comprehend. This includes factors such as sentence length, lexical familiarity, and word order clarity. A top score of 4 reflects a well-structured, fluent sentence that is easy to read without effort.

Preservation of Meaning (M) captures the extent to which the simplified sentence retains the key information and intent of the original. A score of 4 indicates full semantic preservation, while lower scores reflect omission, distortion, or misinterpretation.

Reduction Effort (R_{eff}) refers specifically to the degree of surface-level simplification applied—such as shortening, word substitution, and structural flattening. It does not measure quality or appropriateness in isolation but indicates whether any simplification effort was actually made. This criterion is separated from fluency and faithfulness to avoid conflating simplification activity with overall simplification quality. It is important to note that sentence simplification as a task ideally balances all four criteria rather than optimizing for reduction alone.

Model	G	R	M	R_{eff}	Overall
LLaMA 3.1	**3.46**	**3.26**	**3.24**	**2.16**	**3.03**
OpenNMT	2.26	2.04	1.76	0.94	1.60

Table 3: Mean human ratings: Grammaticality (G), Readability (R), Meaning (M), Reduction Effort (R_{eff}), and Overall Average

The results in Table 3 indicate that LLaMA 3.1 consistently outperformed OpenNMT across all dimensions. The largest gap appeared in preservation of meaning (3.24 vs. 1.76), confirming that LLaMA maintained semantic integrity more reliably. It also exhibited significantly higher grammatical fluency and readability. Although both models attempted simplification, LLaMA applied more effective and deliberate reduction strategies (2.16 vs. 0.94), including sentence segmentation, lexical replacement, and omission of unnecessary modifiers.

These findings reinforce the importance of human evaluation in ATS research. While automatic metrics provide useful benchmarks, they often fail to capture deeper qualitative aspects—such as nuance, clarity, or subtle meaning loss—that are crucial for practical deployment in real-world, user-sensitive applications.

7 Conclusions

This study explored two approaches to Estonian text simplification: a neural machine translation model using OpenNMT and a fine-tuned large language model, LLaMA. Given the limited resources for Estonian ATS, we created the Estonian Simplification Dataset by combining translated data and GPT-4.0-generated simplifications. The experimental results show that the LLaMA model, fine-tuned on this dataset, consistently outperforms OpenNMT across key criteria, including readability, grammaticality, and meaning preservation.

The evaluation showed that standard metrics, such as BLEU and SARI, were insufficient to determine a clear winner between OpenNMT and LLaMA. While BLEU scores marginally favored OpenNMT, reflecting closer alignment with reference texts, SARI scores suggested that LLaMA might better capture the simplification process by adding, deleting, or altering content for readability. However, neither metric alone fully encapsulated critical aspects of simplification quality, such as meaning preservation and readability. These findings underscore the limitations of automated metrics and the necessity of manual evaluation. Despite these limitations, BLEU and SARI are widely adopted in ATS research, enabling comparison with prior work.

The manual evaluations highlighted LLaMA's superior performance, particularly in maintaining the original meaning and applying effective simplification techniques, such as sentence splitting and lexical substitution. These findings underscore the potential of LLMs for handling low-resource languages, with fine-tuning proving effective in adapting pre-trained models to the specific linguistic and structural features of Estonian.

Importantly, the fine-tuning methodology presented in this work is not limited to Estonian. The approach, based on openly available LLaMA models and lightweight tuning via the Unsloth framework, can be readily applied to other low-resource languages with comparable syntactic or morphological complexity. By adapting the data collection, annotation guidelines, and persona-based prompting strategies introduced here, researchers can build domain- and language-specific simplification systems tailored to their needs. To support this, we share scripts, configuration templates, and model inference examples in the supplementary section to facilitate replication and reuse in future work.

This research contributes to the underexplored area of Estonian ATS, demonstrating that LLMs, supported by targeted persona prompting and data resources, can achieve meaningful simplifications in a low-resource language. It also supports practical applications, such as creating educational tools and improving accessibility for Estonian speakers with cognitive disabilities, aligning with user groups like those in the HAPPI, FIRST, and CLEAR projects. Future work will focus on expanding the dataset and incorporating additional human corrections of the simplified sentences. Furthermore, this research lays the groundwork for exploring document-level simplification for Estonian.

8 Supplementary Material

We release all datasets, models, and tools developed in this study to support reproducibility:

- **Dataset:** Estonian Text Simplification Dataset (50,416 sentence pairs), combining GPT-4.0 generations, manual corrections, Wikipedia data, and machine-translated English corpora. `https://huggingface.co/datasets/vulturuldemare/Estonian-Text-Simplification`
- **Models:**
  - Fine-tuned LLaMA 3.1: `https://huggingface.co/datasets/vulturuldemare/Estonian-Text-Simplification/resolve/main/llama31-model.zip`
  - OpenNMT model: `https://huggingface.co/datasets/vulturuldemare/Estonian-Text-Simplification/resolve/main/openNMT-SimplificationModel.pt`
- **Applications:**
  - LLaMA-based web app: `https://github.com/SoimulPatriei/webapp-llama`
  - OpenNMT-based web app: `https://github.com/SoimulPatriei/webapp-opennmt`

All resources are openly available under permissive licenses.

Acknowledgments

This work was supported by the EKTB55 project *Text Simplification for Estonian* and by PRG2006 *Language Technology for Low-Resource Finno-Ugric Languages and Dialects*. We thank the annotators who contributed to the Estonian Simplification Dataset and the three anonymous reviewers for their valuable feedback.

References

Sweta Agrawal, Marine Carpuat, Guy Feigenblat, and Ashish Mishra. 2023. In-context learning for text simplification with large language models.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles, California. Association for Computational Linguistics.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2021. Unsupervised sentence simplification via dependency parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 1327–1337.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 2025 International Conference on Computational Linguistics (COLING)*, pages 6762–6774.

Eduard Barbu, M. Teresa Martín-Valdivia, Eugenio Martínez-Cámara, and L. Alfonso Ureña-López. 2015. Language technologies applied to document simplification for helping autistic people. *Expert Systems with Applications*, 42(12):5076–5086.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Siobhan Devlin and Gary Unthank. 2006. Helping aphasic people process online information. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '06, page 225–226, New York, NY, USA. Association for Computing Machinery.

Isabel Espinosa-Zaragoza, José Abreu-Salas, Elena Lloret, Paloma Moreda, and Manuel Palomar. 2023. A review of research-based automatic text simplification tools. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 321–330, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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

Núria Gala, Thomas François, Eva Ziegler, and Meritxell Bernal. 2020. The clear project: Simplifying texts for better comprehension in second language learners. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with Reading Difficulties*, pages 98–105.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: a project note. In *Proceedings of the Second International Workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, page 9–16, USA. Association for Computational Linguistics.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural crf model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7953.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M. Rush. 2018. Opennmt: Neural machine translation toolkit.

Elizaveta Korotkova and Mark Fishel. 2024. Estonian-centric machine translation: Data, models, and challenges. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 647–660, Sheffield, UK. European Association for Machine Translation (EAMT).

Louis Martin, Angela Fan, Éric V de la Clergerie, Antoine Bordes, and Benoît Sagot. 2020. Controllable sentence simplification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 704–712.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *ACL (2)*. The Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. Accessed: 2024-10-04.

Gustavo H Paetzold and Lucia Specia. 2016. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 55:549–593.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. 2021. AI-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. C4: The colossal clean crawled corpus. *arXiv preprint arXiv:1910.10683*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020c. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Horacio Saggion. 2017. *Automatic Text Simplification*. Synthesis Lectures on Human Language Technologies. Springer, Cham, Switzerland.

Matthew Shardlow. 2014a. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014*, 4(1).

Matthew Shardlow. 2014b. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4.

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable text simplification with explicit rewriting and editing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2605–2617.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Shafqat Virk, Jan Niehues, Christian Bentz, and Marta R Costa-jussà. 2021. Challenges in text simplification for baltic languages: A case study on latvian. *Baltic Journal of Modern Computing*, 9(3):321–335.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2024. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

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.

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.

Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. 2018. Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3173, Brussels, Belgium. Association for Computational Linguistics.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

Sanja Štajner, Maja Popović, and Horacio Saggion. 2022. Text simplification for low-resource languages: A case study on slovenian. *Frontiers in Artificial Intelligence*, 5:837360.

Sanja Štajner and Horacio Saggion. 2018. Can text simplification help people with autism read better? a review. *Universal Access in the Information Society*, 17(2):361–372.