The Two Towers for Estonian-Centric and Finno-Ugric Machine Translation

Mark Fishel and **Lisa Yankovskaya**
Institute of Computer Science
University of Tartu, Estonia
{mark.fisel, lisa.yankovskaya}@ut.ee

Abstract

We present two open-weight translation models for Estonian and its low-resource "relatives" in the Finno-Ugric language family. The training data includes 12 languages paired with Estonian as well as 23 more Finno-Ugric languages and varieties, ranging from mid-resource examples with tens of thousands of speakers to extremely low-resource critically endangered languages with less than a hundred speakers. The translation models use Unbabel Tower+ 2B and 9B as their starting point. We compare their performance on two benchmarks to DeepL and GPT-5.2 and show that in most cases we surpass the quality of DeepL and match or nearly match the quality of GPT-5.2's output with just a fraction of the parameters. Among other contributions we also restore the paragraph structure of a massive synthetic multiparallel corpus for Estonian translation and use it in training the models. The resulting models, training scripts and training data are released openly.

1 Introduction

The discrepancy in the level of support between high-resource languages like English or Chinese and lesser-resourced languages is a well-known issue in natural language processing (NLP). Despite the emergence of massively multilingual language models that cover dozens, hundreds, or even thousands of languages (Martins et al., 2025; Jiang et al., 2023; Ji et al., 2025; Apertus et al., 2025; Communication et al., 2023), the output quality has high variance between the languages and the models often fail to properly support language- and culture-centric content (Ortega and Church, 2023; Mager et al., 2023). As Joshi et al. (2020) emphasize in their taxonomy of language resource disparities, the vast majority of the world's languages remain systematically neglected.

Here we describe the process of creating generative translation models for Estonian and its "relatives": the under-resourced languages from the Finno-Ugric (F-U) family. We pair Estonian with 12 high-resource languages based on their regional and global relevance. We also include in training the parallel data for 23 F-U languages and varieties, several of which are critically endangered and extremely under-resourced. The goal here is to provide these languages with additional support via cross-lingual transfer learning from a strong multilingual model that already includes their higher-resourced relatives, Estonian and Finnish.

A significant contribution of our work is the preparation of a diverse set of translation examples for the included translation directions. A large portion of the data was sentence-level synthetic back-translation data and had to be restored to paragraph-level grouping via alignment with the original document-level source data. This as well as creating paragraph pairs had a number of technical challenges, listed in the sections below. We make all resulting prepared training data openly available[1].

With two Tower+ translation models (Rei et al., 2025), sized 2B and 9B as base models, we de-

[1] https://huggingface.co/datasets/tartuNLP/SynEstParallel

scribe the fine-tuning process and evaluate the results on two benchmarks: FLORES-200 (Team et al., 2022) and WMT24++ (Deutsch et al., 2025). The low-resource F-U language quality is evaluated using SMUGRI-FLORES (Yankovskaya et al., 2023; Pashchenko et al., 2025). Our evaluation results show that, although the models are relatively small (2B and 9B parameters), our larger model generally outperforms DeepL[2], a widely used translation system for high-quality machine translation, and mostly matches the quality of translations produced by GPT-5.2[3]. We release models [4] and scripts used in this work[5].

Our contributions are thus rather technical than research-oriented. We prepare and release a massive parallel dataset and show its usefulness via also training and releasing translation models that achieve high output quality.

The remainder of the paper is organized as follows. Section 2 provides background and reviews related work on Estonian and low-resource Finno-Ugric languages. Section 3 describes the data, its preprocessing, and fine-tuning of the Tower+ models. Section 4 presents the automatic evaluation of both models and analyzes a specific discourse-level phenomenon: gender-ambiguous references in Estonian.

2 Background and Related Work

Our focus is on Estonian and under-resourced Finno-Ugric languages. Estonian has a fair share of support (Kuulmets et al., 2025) and it is added in language models like Llama 3 (Grattafiori et al., 2024) and EuroLLM (Martins et al., 2025) as well as commercial services like DeepL[2] and Google Translate[6]. Furthermore, in 2025 it was simultaneously added to three public evaluation campaigns: WMT general translation shared task (Kocmi et al., 2025), IWSLT offline speech translation task (Abdulmumin et al., 2025) and NLP4CALL MultiGEC grammatical correction shared task (Masciolini et al., 2025).

There are some efforts on providing language-specific models for Estonian, including EstLLM (Dorkin et al., 2026) and TildeOpenLLM (Bergmanis et al., 2026), neither of which are translation models or include tuning to translation instructions. Some Estonian-centric translation models were previously developed (Bergmanis et al., 2022; Korotkova and Fishel, 2024) but all belong to the previous generation of translation models. Previous massively multilingual models like NLLB (Team et al., 2022), SeamlessM4T (Communication et al., 2023) and MADLAD (Kudugunta et al., 2023) all include Estonian but lose in quality to the Estonian-centric models.

Lower-resourced Finno-Ugric languages have been studied extensively in the context of NLP, including rule-based translation work for Sámi languages (Antonsen et al., 2016; Trosterud and Unhammer, 2012) and neural translation models for 19 low-resource Finno-Ugric languages (Yankovskaya et al., 2023; Purason et al., 2024)[7]. In this work, we include several languages that were not supported in these earlier works, namely Pite Sámi, Kildin Sámi, Ingrian and Votic.

3 Model Training Details

The included translation directions were selected to cover both regionally relevant and globally prevalent languages. Regionally relevant languages include Finnish, Latvian, Lithuanian, Swedish, Russian, and Ukrainian; globally prevalent languages include English, German, French, Spanish, Chinese, and Arabic. All 12 languages were included as input and output, resulting in 24 translation directions.

3.1 Training Data

Below we describe the preparation of the training data and the way the models were fine-tuned. One of the goals of our work was to enable the resulting translation models to handle multiple sentences in a single request. This is especially important for genderless languages like Estonian and its low-resource F-U relatives due to genderless pronouns and the need to resolve them outside of the sentence (Pashchenko et al., 2025).

A large portion of the training data, the SynEst corpus (Korotkova and Fishel, 2024) consists of back-translation data that was created with older sentence-level translation models. We restored the paragraph-level alignment of the data based on the original source material, which presented some

[2]https://www.deepl.com/en/products/translator
[3]https://openai.com/index/introducing-gpt-5-2/
[4]https://huggingface.co/tartuNLP/Tigutorn2B, https://huggingface.co/tartuNLP/Tahetorn9B
[5]https://koodivaramu.eesti.ee/tartunlp/translate
[6]https://translate.google.com/

[7]Though the authors state "20 low-resource F-U languages", we consider Proper Karelian and Livvi Karelian as two dialects of the same language.

Corpus	Original Sentences	Deleted	Missing (Initial)	Missing (Final)	Recovered (%)	Final Translated
OpenSubtitles	441.4M	161.9k	1.6M	45.8k	97.1%	441.2M
UNPC	34.5M	196.2k	1.3M	632.7k	51.3%	33.7M
ENC	2.2B	18.5M	175.8M	161.2M	8.3%	2.0B
Total	**2.6B**	**18.9M**	**277.7M**	**161.9M**	**41.7%**	**2.4B**

Table 1: Document-level reconstruction statistics for the synthetic corpora. "Missing (Initial)" and "Missing (Final)" represent alignment failures before and after applying text normalization techniques. "Recovered (%)" denotes the proportion of initially missing sentences rescued through this processing.

challenges – below we describe the way these were tackled.

Restoring SynEst Paragraphs

The SynEst corpus consists of several monolingual corpora that were machine-translated to serve as synthetic (back-translated) data. While some of these have the original sentence order scrambled (Newscrawl, Paracrawl), other parts are based on corpora with the original order of sentences preserved (OpenSubtitles, UNPC and ENC, the Estonian National Corpus). This naturally leads to the possibility of restoring the document and paragraph structure in the SynEst corpus source-side and projecting it to the generated translations: even though the translations were done with a sentence-level system, the original texts can support learning inter-sentential phenomena.

Still, turning it into a parallel corpus with paragraph-level alignments presented a number of challenges. As the original corpus was created with sentence pairs in mind, some source sentences were skipped during translation (Korotkova and Fishel, 2024) – leading to possible gaps in the documents.

Moreover, the corpus was composed of *processed* source sentences instead of the original ones – which, for instance, resulted in the translation models occasionally replacing unrecognized symbols with `<unk>` tokens, which prevented exact string matching. Additionally, errors such as unescaped ampersand symbols in the UNPC dataset caused crashes during XML tree parsing. To maximize sentence matching and bypass these artifacts, we applied targeted text normalization: we systematically removed the `<unk>` tokens and temporarily stripped all non-ASCII characters and numbers from the source material and source-side SynEst corpus sentences during the alignment phase to ensure consistent matching.

This data processing and normalization significantly minimized data loss and rescued a massive amount of paragraph context. Out of the 2.6 billion total sentences in the original source corpora, 277.7 million sentences were initially missing due to alignment mismatches. By applying our normalization and replacement strategies, we successfully recovered 115.8 million of these sentences, representing a 41.7% rescue rate of the missing data. As detailed in Table 1, the final document-level reconstructed corpus retains 2.4 billion translated sentences with only a 7% overall data loss, providing a robust, context-rich dataset for fine-tuning our generative translation models.

Other Training Data

We also included native paragraph-level parallel data from DocHPLT (O'Brien et al., 2025) as well as sentence-level parallel data from several corpora in the OPUS collection (Tiedemann, 2012).

Parallel data for low-resource F-U languages was taken from a public repository[8]. This constituted only 7% of the total training data, when counting by the ratio of output tokens. The included languages were:

Finnic: Võro (Southern Estonian), Livonian, Izhorian, Votic, Karelian (including Proper Karelian and Livvi Karelian), Ludian and Veps

Sámi: Northern Sámi, Inari Sámi, Southern Sámi, Skolt Sámi, Lule Sámi, Kildin Sámi and Pite Sámi

Permic: Komi Zyrian, Komi Permyak, Udmurt

Mari: Meadow Mari, Hill Mari

Mordvin: Erzya, Moksha

Ugric: Mansi, Khanty

Several of these languages actually consist of multiple dialects, which were kept in the input to inform the translation models of which dialect they are taught to generate. For instance, Khanty splits into Kazym Khanty, Surgut Khanty, etc.

[8] https://huggingface.co/datasets/tartuNLP/smugri4-data

	Native (human) data, directions from Estonian to:											
	Arabic	**Chinese**	**English**	**Finnish**	**French**	**German**	**Latvian**	**Lithuanian**	**Russian**	**Spanish**	**Swedish**	**Ukrainian**
Segments	1.8M	15k	8.6M	1.3M	991K	2.6M	703k	725k	1.2M	1.4M	1.3M	763K
Source Words	8.9M	148K	113.0M	12.4M	15.1M	41.2M	9.0M	7.4M	7.6M	20.7M	15.3M	3.8M
Target Words	8.3M	21K	156.7M	11.6M	22.7M	53.5M	9.9M	7.8M	8.2M	31.1M	19.1M	3.9M
	Native (human) data, directions into Estonian from:											
	Arabic	**Chinese**	**English**	**Finnish**	**French**	**German**	**Latvian**	**Lithuanian**	**Russian**	**Spanish**	**Swedish**	**Ukrainian**
Segments	1.8M	15K	8.6M	1.3M	991K	2.6M	703K	725K	1.2M	1.4M	1.3M	763K
Source Words	8.3M	21K	156.7M	11.6M	22.7M	53.5M	9.9M	7.8M	8.2M	31.1M	19.1M	3.9M
Target Words	8.9M	148K	113.0M	12.4M	15.1M	41.2M	9.0M	7.4M	7.6M	20.7M	15.3M	3.8M
	Synthetic (back-translation) data, synthetic data back-translated from Estonian to:											
	Arabic	**Chinese**	**English**	**Finnish**	**French**	**German**	**Latvian**	**Lithuanian**	**Russian**	**Spanish**	**Swedish**	**Ukrainian**
Segments	14.0M	32.8M	84.1M	55.9M	50.5M	64.2M	40.1M	39.8M	75.5M	52.6M	56.3M	60.3M
Source Words	132.0M	47.2M	1361.3M	690.5M	925.7M	967.6M	557.5M	544.7M	1071.6M	945.3M	834.7M	875.4M
Target Words	116.9M	338.0M	994.8M	731.5M	631.8M	751.4M	518.7M	522.4M	929.4M	650.5M	688.1M	775.9M

Table 2: Amount of native (human) and synthetic (back-translated) parallel data used for training Tigutorn and Tähetorn. An additional 16.2M segments (390.3M source words, 563.5M target words) was back-translated from English to Estonian. This makes a total of 804.1M segments (11.3B source words, 10.2B target words), of which about 84% is synthetic.

Training Data Filtering

Finally, we filtered the resulting mix of data heurisitcally, using OpusFilter (Aulamo et al., 2020) and then using COMET-Kiwi (Rei et al., 2022) as a source of quality estimates and setting a quality threshold of 0.85. The resulting data sizes are shown in Table 2. 84% of the data (8.58×10^9 target words) consisted of synthetic translations (back-translation and pivot-generated F-U data) and the remaining 16% (1.62×10^9 target words) was native parallel data.

3.2 Model Fine-tuning

The Tower+ models (Rei et al., 2025) were selected as base models due to their strong translation performance. These models do not officially support Estonian or other F-U languages and had to be fine-tuned. We focused on the smaller versions of Tower+, namely the 2B and 9B versions. We transferred the naming of the original model to Estonian and called the resulting models Tigutorn (2B) and Tähetorn (9B)[9].

Since the base models are instruction-tuned, we performed supervised fine-tuning, matching the Tower+ instruction format. Computational resources for the fine-tuning were provided by the LUMI Supercomputer[10]. We ran the process until the model had seen all of the high-resource training examples once and the low-resource F-U data twice, adjusting data sampling accordingly. With a batch size of 8192 segments (adjusted to 256 GPUs / 32 nodes via gradient accumulation) this resulted in 98'000 update steps. The learning rate was set to 10^{-4} via a preliminary grid search and selection of the most quickly reducing loss value.

4 Evaluation

In this section, we evaluate the translation quality of both systems using automatic evaluation metrics and examine whether the systems can resolve gender-ambiguous references in Estonian.

We first report the results for translation between Estonian and other high-resource languages in both directions. We then present the evaluation results for low-resource languages. This separation is necessary because, for high-resource languages, more reliable neural network-based metrics can be applied, here we used reference-less COMETKiwi-XL (Rei et al., 2023) and reference-based xCOMET-XL (Guerreiro et al., 2024). Since modern neural network-based metrics do not support the low-resource languages in our setting, we use ChrF++ (Popović, 2017) for their evaluation.

In addition to automatic evaluation, we analyze a specific discourse-level phenomenon: gender-ambiguous references in Estonian. This analysis focuses on cases where the source text does not explicitly encode gender, but the target language requires gender-marked forms, such as pronouns or verb forms.

[9]Both Tähetorn and Tigutorn are well-known tower landmarks in Tartu, Estonia.

[10]As one of the most eco-efficient data centers in the world, LUMI reports a carbon-negative footprint: https://csc.fi/en/news/lumi-europes-most-powerful-supercomputer-is-solving-global-challenges-and-promoting-a-green-transformation/

	est → x												x → est											
	ara	zho	eng	fin	fra	deu	lav	lit	rus	spa	swe	ukr	ara	zho	eng	fin	fra	deu	lav	lit	rus	spa	swe	ukr
FLORES																								
Tigutorn	0.533	0.654	0.832	0.761	0.691	0.743	0.656	0.579	0.726	0.773	0.731	0.648	0.594	0.738	0.850	0.767	0.756	0.826	0.673	0.637	0.760	0.805	**0.772**	0.610
Tähetorn	0.599	0.671	0.836	0.775	0.697	0.760	0.681	0.637	0.740	0.780	0.742	0.670	0.613	0.743	0.851	**0.770**	0.754	**0.829**	0.673	**0.646**	0.769	0.807	**0.772**	0.611
GPT-5.2	**0.641**	**0.689**	**0.839**	**0.785**	**0.698**	**0.768**	**0.690**	**0.674**	**0.747**	**0.785**	**0.746**	**0.678**	**0.618**	**0.757**	**0.857**	0.768	**0.763**	0.827	**0.679**	0.641	**0.770**	**0.810**	0.771	**0.617**
DeepL	0.622	0.654	0.832	0.765	0.687	0.756	0.667	0.652	0.730	0.773	0.732	0.658	0.593	0.713	0.844	0.743	0.741	0.815	0.650	0.615	0.746	0.788	0.754	0.597
WMT24++																								
Tigutorn	0.452	0.529	0.750	0.628	0.545	0.632	0.520	0.431	0.587	0.657	0.594	0.510	0.423	0.629	0.739	0.645	0.607	0.722	0.549	0.520	0.655	0.705	0.629	0.508
Tähetorn	0.489	0.550	**0.757**	**0.643**	**0.567**	**0.659**	0.543	0.493	**0.602**	**0.665**	**0.610**	**0.531**	0.468	**0.646**	0.751	**0.652**	**0.623**	**0.729**	**0.556**	**0.532**	**0.659**	**0.713**	**0.638**	**0.514**
GPT-5.2	0.493	**0.556**	0.751	**0.643**	0.554	0.645	**0.549**	**0.514**	0.551	0.606	0.595	0.528	**0.510**	0.615	**0.760**	0.651	0.620	0.724	0.547	0.524	0.650	0.707	0.628	0.506
DeepL	**0.495**	0.535	0.749	0.629	0.547	0.649	0.522	0.494	0.589	0.654	0.601	0.518	0.475	0.596	0.720	0.628	0.586	0.702	0.529	0.509	0.618	0.675	0.615	0.480

Table 3: CometKiwi-XL scores for Estonian↔x translation across 12 language pairs on FLORES and WMT 24++. Within each dataset block, the best score in each column is shown in bold and the second-best distinct score is underlined.

	est → x												x → est											
	ara	zho	eng	fin	fra	deu	lav	lit	rus	spa	swe	ukr	ara	zho	eng	fin	fra	deu	lav	lit	rus	spa	swe	ukr
FLORES																								
Tigutorn	0.642	0.810	0.960	0.919	0.839	0.937	0.820	0.720	0.901	0.899	0.917	0.868	0.799	0.903	0.948	0.934	0.916	0.945	0.896	0.876	0.930	0.927	0.936	0.896
Tähetorn	0.755	0.840	0.967	0.944	0.858	0.949	0.871	0.821	0.920	0.912	0.936	0.897	0.842	0.921	0.950	**0.942**	**0.925**	**0.951**	0.906	0.891	**0.938**	**0.937**	**0.943**	**0.908**
GPT-5.2	**0.842**	**0.863**	**0.972**	**0.955**	**0.871**	**0.958**	**0.896**	**0.894**	**0.929**	**0.921**	**0.943**	**0.910**	**0.868**	**0.926**	**0.952**	0.940	0.923	0.948	**0.908**	**0.897**	0.936	0.933	0.942	**0.908**
DeepL	0.833	0.842	0.957	0.937	0.857	0.950	0.870	0.874	0.917	0.902	0.930	0.896	0.843	0.889	0.941	0.919	0.906	0.938	0.884	0.866	0.920	0.917	0.924	0.891
WMT24++																								
Tigutorn	0.484	0.653	0.872	0.807	0.654	0.831	0.670	0.557	0.731	0.755	0.791	0.697	0.494	0.763	0.836	0.829	0.776	0.837	0.761	0.721	0.790	0.811	0.816	0.767
Tähetorn	0.563	0.681	0.890	0.849	0.686	0.859	0.715	0.653	0.763	0.783	0.824	0.741	0.591	0.796	0.850	0.849	0.803	0.851	0.775	0.751	0.808	0.829	0.832	0.783
GPT-5.2	**0.623**	**0.742**	**0.913**	**0.872**	**0.729**	**0.877**	**0.758**	**0.728**	**0.787**	**0.807**	**0.844**	**0.762**	**0.723**	**0.828**	**0.880**	**0.865**	**0.821**	**0.866**	**0.807**	**0.786**	**0.819**	**0.856**	**0.854**	**0.809**
DeepL	0.595	0.680	0.873	0.835	0.672	0.857	0.700	0.695	0.759	0.767	0.805	0.740	0.635	0.739	0.810	0.816	0.762	0.821	0.750	0.725	0.764	0.795	0.805	0.750

Table 4: xCOMET-XL scores for Estonian↔x translation across 12 language pairs on FLORES and WMT24++. Within each dataset block, the best score in each column is shown in bold and the second-best distinct score is underlined.

4.1 Evaluation of High-Resource Languages

For the evaluation of high-resource languages, we used two datasets: the devtest split of FLORES-200, which contains 1012 segments from Wikimedia (Team et al., 2022) and WMT24++, which consists of 998 source texts from four domains: literary, news, social, and speech (Deutsch et al., 2025).

In addition to comparing our two models, we include DeepL[2] and GPT-5.2[3] as strong baselines[11].

Both metrics show similar overall trends. First, WMT24++ appears to be more challenging benchmark than FLORES-200. Second, according to both metrics, translation into Estonian is systematically easier than translation out of Estonian.

According to COMETKiwi-XL (Table 3), GPT-5.2 achieves the best average performance on FLORES, but this advantage does not carry over to WMT24++ dataset, where Tähetorn ranks first overall. On WMT24++ dataset DeepL and GPT-5.2 show similar performance. Though the differences between all systems are relatively small.

According to xCOMET-XL (Table 4), GPT-5.2 achieves the best overall performance across all Estonian-centric translations and both datasets. Its advantage is particularly strong on WMT24++ dataset, where it ranks first on all language pairs. Tähetorn, the larger of our two models, consistently outperforms Tigutorn, indicating that scaling from 2B to 9B yields gains for nearly all language pairs. DeepL performs between Tähetorn and Tigutorn.

Overall, the results do not identify a single system that dominates across both metrics and datasets, although GPT-5.2 and Tähetorn generally outperform DeepL and Tigutorn.

4.2 Gender Ambiguity in Paragraph-level Translation

Here, we examine whether our systems can resolve gender-ambiguous references using discourse context. The issue is particularly relevant for Estonian and other Finno-Ugric languages, where personal pronouns, surnames and verb forms do not encode gender. As a result, without sufficient context, it is impossible to determine whether a masculine or feminine pronoun, adjective, verb, etc. should

[11] We did not evaluate our systems against the Tower models, as these models do not officially support Estonian. Although they may have some knowledge of Estonian, we tested both Tower models on the WMT24++ dataset using English, the most common language. Both systems, Tower+ 2B and Tower+ 9B, obtained low CometKiwi-XL scores when translating into Estonian: 0.398 and 0.399, respectively. Translation from Estonian yielded better results, with scores of 0.672 and 0.668, respectively. Still, after fine-tuning the scores go up significantly: to 0.739 and 0.751 for English-Estonian and to 0.75 and 0.757 for Estonian-English (for 2B and 9B).

Original paragraph in Estonian
Professionaalseks fotograafiks võeti **Lembit/Veera** Peegel tööle alles 1987. aastal Maalehte. Spordifotosid **ta** aga ei jätnud. "Sport oli minu teine armastus, nagu öeldakse. Kui ma töötasin Maalehes, käisin ikka spordivõistlustel." Spordilehte kutsuti Peegel tööle 1989. aastal. "Palk oli väiksem, tööd oli kindlasti rohkem, honorari ka maksti. Siis ma küsisin abikaasa käest, kas minna või mitte minna. Abikaasa ütles, et kui sa armastad mind ja sporti, siis palun mine," meenutas **ta**. Ka nüüd, olles 88-aastane, pole Peegel sporti jätnud. "Sel aastal lähen ka kindlasti jalgpalli, korvpalli ja kergejõustikku pildistama," on **ta** kindel. "Aga nüüd ma valin muidugi ilma ja temperatuuri," muigas **ta**.
Translation with a female first name
Veera Peegel was hired as a professional photographer only in 1987 at Maaleht. However, **she** did not stop taking sports photos. "Sports was my second love, as they say. When I worked at Maaleht, I still went to sports competitions." Peegel was invited to work at Spordileht in 1989. "The salary was smaller, there was definitely more work, and I was paid a fee. Then I asked my husband whether to go or not to go. My husband said that if you love me and sports, then please go," **she** recalled. Even now, at the age of 88, Peegel has not given up on sports. "This year I will definitely go to take photos of football, basketball and track and field," **she** is sure. "But now I choose the weather and temperature, of course," **she** smiled.
Translation with a male first name
Lembit Peegel was hired as a professional photographer only in 1987 at Maaleht. However, **he** did not give up sports photography. "Sports was my second love, as they say. When I worked at Maaleht, I still went to sports competitions." Peegel was invited to work at Spordileht in 1989. "The salary was smaller, there was definitely more work, and I was paid a fee. Then I asked my wife whether to go or not to go. My wife said that if you love me and sports, then please go," **he** recalled. Even now, at the age of 88, Peegel has not given up sports. "This year, I will definitely go to photograph football, basketball and track and field," **he** is sure. "But now I choose the weather and temperature, of course," **he** smiled.

Table 5: Example output of Tähetorn for a gender-contrastive paragraph pair, showing consistent translation of gendered pronouns and context-dependent translation of Estonian "abikaasa" ("spouse") as "husband" or "wife" according to the referent's gender.

be used when translating into languages that encode grammatical gender (e.g. pronouns in English and several parts of speech in German, Russian, French, and others). To explore this question, we conducted two experiments.

Gender Ambiguity with Names as Cue

In the first experiment, we extracted Estonian paragraphs from local news media in order to evaluate the systems on naturally occurring discourse contexts. The selected paragraphs contained at least two sentences and mentioned a person by first name. Subsequent sentences either contained the pronoun "tema" – corresponding to both "she" and "he" in English – or referred to the same person by surname only, for example "Tamm said". In Estonian, such references remain gender-ambiguous.

For each original paragraph, we created an additional version of it with a first name from the other gender. Thus, when the original contained a female first name, we created the male version by replacing only the first name, and vice-versa. All other parts of the paragraph were kept unchanged. Together, these two versions formed one gender-contrastive paragraph pair. This setup tests whether the system can use the gender information provided by the first name to resolve gender-ambiguous references elsewhere in the paragraph.

We then translated both versions into English and Russian. These two target languages differ in how gender ambiguity affects translations. In English, the ambiguity is relevant only when the translation requires a gender pronoun ("he" or "she"). In cases where no pronoun is required, for example "Tamm said", the sentence remains grammatically correct regardless of whether "Tamm" refers to a woman or a man. In Russian, however, the translator needs to know the person's gender, because past-tense verbs are gender-marked. This difference in gender-marking requirements led to different numbers of paragraphs in the English and Russian evaluations.

We evaluated the translations at the level of paragraph pairs. A pair was counted as correct only if both the female-name and male-name versions reflected the intended gender. If either version in the pair contained an incorrect gender assignment, the entire pair was counted as containing a gender-related error.

The results show that both systems are often able to infer the gender associated with the first name and use this information to translate gender-marked forms consistently. Overall, the larger model (Tähetorn) performs more reliably than the smaller system (Tigutorn) both in terms of gender consistency and in its lower tendency to omit or

input ↓ output →	Estonian	Finnish	Hungarian	English	Latvian	Russian	Võro	Livonian	Veps	Ludic	Livvi Karelian	Proper Karelian	Erzya	Moksha	Hill Mari	Meadow Mari	Komi Zyrian	Udmurt	Mansi	Kazym Khanty
Estonian		53.2	50.5	66.5	52.0	50.9	47.5	36.6	37.6	35.1	36.2	45.5	35.1	38.8	34.3	38.8	41.9	37.3	21.9	16.0
Finnish	53.8		48.5	61.5	49.7	48.2	36.1	32.2	37.8	36.5	38.3	53.5	34.9	34.0	34.2	37.9	38.1	35.9	21.8	16.9
Hungarian	52.9	49.4		61.7	50.3	47.8	35.3	31.7	36.2	32.3	35.4	44.1	34.6	34.4	34.4	38.5	37.6	36.1	21.3	16.6
English	61.4	55.7	54.6		56.5	55.5	37.5	34.1	37.5	33.8	36.5	46.6	35.4	35.3	33.8	40.3	40.7	37.5	22.1	14.5
Latvian	56.1	52.3	50.9	64.4		50.3	36.1	33.1	38.3	33.9	36.8	44.1	36.0	36.0	36.4	40.5	40.4	38.0	23.0	16.6
Russian	54.6	49.5	48.3	63.7	51.2		35.7	33.3	43.5	36.5	41.1	45.7	43.5	36.9	43.0	48.7	48.7	45.9	25.7	20.0
Võro	70.3	47.9	45.8	55.9	45.9	44.1		34.8	35.3	32.2	34.2	41.7	33.0	37.3	32.9	35.0	37.3	33.8	21.5	17.7
Livonian	58.8	45.8	44.1	54.3	44.6	42.2	37.7		34.7	32.5	33.9	40.9	34.0	35.8	32.7	34.1	37.5	33.7	20.9	16.4
Veps	44.8	44.5	40.7	47.7	42.7	47.8	31.0	30.0		34.0	38.1	40.2	37.4	34.5	36.8	39.1	40.7	37.5	24.0	19.0
Ludic	49.8	52.9	43.5	50.9	44.8	46.4	32.4	31.9	39.4		39.9	44.1	35.7	35.1	35.1	37.4	39.0	35.9	23.1	18.2
Livvi Karelian	46.8	50.3	43.0	51.6	44.5	49.2	30.8	29.5	40.2	34.8		43.1	37.3	34.4	36.4	39.1	40.4	36.7	22.8	17.7
Proper Karelian	51.3	60.9	47.5	57.1	47.1	47.0	34.1	31.5	38.7	35.5	39.1		36.0	35.4	35.9	37.7	39.9	36.6	21.7	17.9
Erzya	43.0	44.0	42.2	51.1	43.4	56.3	30.7	29.6	38.8	30.9	35.8	39.3		34.2	38.3	43.0	42.6	39.5	25.3	17.5
Moksha	49.4	43.5	42.1	49.5	42.0	41.1	33.0	29.8	34.4	30.6	32.9	37.4	34.4		32.2	36.3	37.9	34.4	23.0	17.6
Hill Mari	44.1	44.4	43.1	52.4	44.2	52.2	31.6	29.3	37.4	31.5	35.2	38.9	38.5	33.3		46.4	41.2	41.1	25.1	19.5
Meadow Mari	44.1	45.0	44.0	53.2	45.3	56.9	31.4	29.8	38.3	32.1	35.5	39.6	39.0	35.0	43.8		42.3	41.8	26.6	20.7
Komi-Zyrian	47.8	45.1	42.8	52.1	45.7	53.9	32.8	30.8	39.1	32.7	36.3	40.6	39.4	36.2	39.2	42.8		40.5	25.2	20.8
Udmurt	45.2	44.9	44.3	52.3	45.0	57.3	31.9	30.0	39.1	33.1	36.5	40.2	38.7	34.5	41.4	44.9	43.5		26.2	20.9
Mansi	37.9	36.8	34.4	39.7	36.0	39.7	26.5	24.4	31.3	26.2	29.1	32.0	31.8	28.4	31.0	35.5	33.6	33.1		17.1
Kazym Khanty	34.0	34.1	31.7	36.5	31.6	35.5	25.2	23.5	30.0	24.9	28.6	29.8	29.0	26.5	27.5	31.2	32.3	30.1	21.2	

Table 6: ChrF++ scores for 20 languages on the subset of FLORES-200 dataset, 14 of which are low-resource Finno-Ugric languages or dialects. These are grouped according to their respective subfamilies. The color scheme used in the table ranges from yellow to green for lower to higher scores.

avoid the relevant gender-marked reference.

For Estonian-Russian, both systems were evaluated on 43 paragraphs. Tigutorn left the relevant sentence untranslated in one case, leaving 42 evaluable outputs. Of these, seven contained a gender-related error. In six cases, the system inferred the person's gender incorrectly and used gender-marked forms consistent with the wrong gender. In one additional case, the system used a feminine noun form but later referred to the same person as male. Tähetorn compressed two two-sentence paragraphs into single-sentence outputs, however, these outputs still preserved the correct gender marking and were therefore included in the gender evaluation. Overall, it produced one gender-related error, where a masculine noun form was used instead of the required feminine form.

For Estonian-English, both systems were evaluated on 39 paragraphs. Tigutorn failed to translate the final relevant sentence in four cases and rephrased two outputs in a way that avoided the use of a gendered pronoun, leaving 33 evaluable outputs. Among them, four outputs contained an incorrect gender assignment. Tähetorn left the relevant sentence untranslated in two cases. It also merged two two-sentence paragraphs into single-sentence outputs, but in both cases the gendered pronoun was translated correctly; these outputs were therefore counted as correct for gender assignment. Among the 37 evaluable pairs, one contained an incorrect gender assignment.

Table 5 presents an example from Tähetorn output for one gender-contrastive paragraph pair. The example shows that the system translates gendered pronouns consistently and also resolves the gender-neutral Estonian noun "abikaasa" ("spouse") according to the referent's gender.

Gender Ambiguity with Keywords as Cue

The second experiment used a similar gender-contrastive setup, but differed in the type of gender cue. Whereas the first experiment tested whether the systems could infer gender from first names, the second tested whether they could use explicit lexical gender cues in the discourse. These cues included words such as "sister/brother", "woman/man", "mother/father".

For this experiment, we created 15 gender-contrastive pairs of short segments, each consisting of two or three sentences. The two segments in each pair were identical except for the gender of the referent. Unlike the first experiment, which used naturally occurring news paragraphs, these examples were designed specifically for this evaluation.

As in the first experiment, the segments were translated into English and Russian, and the evaluation was conducted at the level of gender-contrastive pairs. A pair was counted as correct only if both versions preserved the intended gender consistently.

Tähetorn, the large system, preserved the in-

tended gender across all segments and both target languages. Tigutorn, the small system, made one gender-related error: when translating a segment containing "vanaisa" ("grandfather") to Russian, it used female pronouns. In some outputs, the systems combined all sentences into a single sentence; however, the gender marking remained correct and consistent with the intended referent.

Overall, both experiments demonstrated that both systems handle gender ambiguity well.

4.3 Evaluation of Low-Resource Languages

To evaluate low-resource F-U languages, we used SMUGRI-FLORES (Yankovskaya et al., 2023; Pashchenko et al., 2025). It consists of 250 segments taken from the devtest split of FLORES-200 and covers 14 language varieties: Mansi, Kazym Khanty (modern orthography), Komi Zyrian, Udmurt, Hill Mari, Meadow Mari, Moksha, Erzya, Livonian, Veps, Võro, Ludic, Karelian Proper, and Livvi Karelian. We aligned these with six high-resource languages: Estonian, Finnish, and Hungarian, as languages from the same language group; Russian and Latvian, as supported languages for some low-resource languages; and English, as one of the main world languages.

Table 6 shows the expected patterns: the model performs better when translating from low-resource languages into high-resource languages (46.1 ± 6.7) than in the opposite direction (35.5 ± 7.8), while the lowest quality is observed in translation between low-resource languages (33.2 ± 6.6). However, some low-resource to low-resource directions still achieve relatively strong results, particularly when the languages are closely related, as in Hill Mari into Meadow Mari. More generally, performance is usually best when translating into a closely related language, such as Võro into Estonian, or into a language with training support, like Udmurt into Russian. Translation from low-resource languages into English usually yields the second-best score, except for Livvi Karelian and Kazym Khanty, for which translation into English achieves the best result.

Although several languages in our study are also covered in previous work (Purason et al., 2024), they do not report individual metrics for separate translation pairs, only average values, making a direct comparison per translation pair impossible. Comparison between the average scores of their work and Tähetorn, computed in the same way, shows noticeable improvement (see Table 7).

	chrF++
`low-low`	34.9 (+2.5)
`low-high`	46.2 (+2.4)
`high-low`	35.8 (+1.9)

Table 7: Average chrF++ scores across different language pair clusters for nine low-resource languages Komi, Udmurt, Hill and Meadow Mari, Erzya, Moksha, Livonian, Mansi, Livvi Karelian; numbers in brackets indicate the improvement of the Tähetorn over the previous work (Purason et al., 2024). Low-low - translations from low-resource to low-resource, low-high - from low-resource to high-resource, high-low - from high-resource to low-resource languages.

5 Conclusions

We presented the newly developed Estonian-centric generative MT models Tigutorn (2B) and Tähetorn (9B). The models were trained on data for translation directions from/into Estonian, covering 12 other high-resource languages, as well as 23 more low-resource Finno-Ugric languages. We also described the details of data preparation and model training. Finally, we presented results of automatic evaluation, which compared the two models to a commercial translation service (DeepL) and a commercial language model (GPT-5.2) and studied how gender-ambiguous Estonian references are translated into English and Russian, where gender must often be inferred from context.

Results show that it is possible, albeit via investing significant computational resources and data, to match or surpass the performance of commercial language and translation models with a much smaller-scale open-weight model. The models, data and all associated scripts are released openly.

One of limitations of the presented study is that manual evaluation was limited to gender-ambiguous references in Estonian. While this captures one discourse-level aspect of paragraph-level translation, the overall comparison of model performance relies on automatic metrics. Broader qualitative analysis is therefore needed to better assess the strengths and weaknesses of the developed models.

Also, paragraph-level translation is only partially evaluated. Other discourse-level phenomena remain outside the scope of the evaluation, largely due to the lack of document-level benchmarks and suitable metrics for Estonian.

It would also be interesting to compare the per-

formance of our models on the few low-resource Finno-Ugric languages supported elsewhere: for instance, Northern Sami and Komi Zyrian in Google Translate, as well as any of the F-U languages supported in GPT-5.2.

Finally, given the scale of the present study, we conclude that further significant improvement of translation between Estonian and other high-resource languages can be achieved at relatively great costs: by switching to significantly larger models and/or performing massive-scale back-translation with more monolingual data.

Acknowledgements

This work was supported by the National Program for Estonian Language Technology Program under project EKTB67: "Estonian Machine Translation: Adding New Functionality and Languages", funded by the Estonian Ministry of Education and Research, as well as by the Estonian Research Council grant PRG2006 (Language Technology for Low-Resource Finno-Ugric Languages and Dialects). All computations were performed on the LUMI Supercomputer through the University of Tartu's HPC center.

References

Abdulmumin, Idris, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, Fortuné Kponou, Mateusz Krubiński, Tsz Kin Lam, Danni Liu, Evgeny Matusov, Chandresh Kumar Maurya, John P. McCrae, Salima Mdhaffar, Yasmin Moslem, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Atul Kr. Ojha, John E. Ortega, Sara Papi, Pavel Pecina, Peter Polák, Piotr Połeć, Ashwin Sankar, Beatrice Savoldi, Nivedita Sethiya, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Marco Turchi, Alex Waibel, Patrick Wilken, Rodolfo Zevallos, Vilém Zouhar, and Maike Züfle. 2025. Findings of the IWSLT 2025 evaluation campaign. In Salesky, Elizabeth, Marcello Federico, and Antonis Anastasopoulos, editors, *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481, Vienna, Austria (in-person and online), July. Association for Computational Linguistics.

Antonsen, Lene, Trond Trosterud, and Francis Tyers. 2016. A north saami to south saami machine translation prototype. *Northern European Journal of Language Technology*, 4:11–27, 03.

Apertus, Project, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, Inés Altemir Mariñas, Mohammad Hossein Amani, Matin Ansaripour, Ilia Badanin, Harold Benoit, Emanuela Boros, Nicholas Browning, Fabian Bösch, Maximilian Böther, Niklas Canova, Camille Challier, Clement Charmillot, Jonathan Coles, Jan Deriu, Arnout Devos, Lukas Drescher, Daniil Dzenhaliou, Maud Ehrmann, Dongyang Fan, Simin Fan, Silin Gao, Miguel Gila, María Grandury, Diba Hashemi, Alexander Hoyle, Jiaming Jiang, Mark Klein, Andrei Kucharavy, Anastasiia Kucherenko, Frederike Lübeck, Roman Machacek, Theofilos Manitaras, Andreas Marfurt, Kyle Matoba, Simon Matrenok, Henrique Mendonça, Fawzi Roberto Mohamed, Syrielle Montariol, Luca Mouchel, Sven Najem-Meyer, Jingwei Ni, Gennaro Oliva, Matteo Pagliardini, Elia Palme, Andrei Panferov, Léo Paoletti, Marco Passerini, Ivan Pavlov, Auguste Poiroux, Kaustubh Ponkshe, Nathan Ranchin, Javi Rando, Mathieu Sauser, Jakhongir Saydaliev, Muhammad Ali Sayfiddinov, Marian Schneider, Stefano Schuppli, Marco Scialanga, Andrei Semenov, Kumar Shridhar, Raghav Singhal, Anna Sotnikova, Alexander Sternfeld, Ayush Kumar Tarun, Paul Teiletche, Jannis Vamvas, Xiaozhe Yao, Hao Zhao, Alexander Ilic, Ana Klimovic, Andreas Krause, Caglar Gulcehre, David Rosenthal, Elliott Ash, Florian Tramèr, Joost VandeVondele, Livio Veraldi, Martin Rajman, Thomas Schulthess, Torsten Hoefler, Antoine Bosselut, Martin Jaggi, and Imanol Schlag. 2025. Apertus: Democratizing open and compliant llms for global language environments.

Aulamo, Mikko, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In Celikyilmaz, Asli and Tsung-Hsien Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156, Online, July. Association for Computational Linguistics.

Bergmanis, Toms, Marcis Pinnis, Roberts Rozis, Jānis Šlapiņš, Valters Šics, Berta Bernāne, Guntars Pužulis, Endijs Titomers, Andre Tättar, Taido Purason, Hele-Andra Kuulmets, Agnes Luhtaru, Liisa Rätsep, Maali Tars, Annika Laumets-Tättar, and Mark Fishel. 2022. MTee: Open machine translation platform for Estonian government. In Moniz, Helena, Lieve Macken, Andrew Rufener, Loïc Barrault, Marta R. Costa-jussà, Christophe Declercq, Maarit Koponen, Ellie Kemp, Spyridon Pilos, Mikel L. Forcada, Carolina Scarton, Joachim Van den Bogaert, Joke Daems, Arda Tezcan, Bram

Vanroy, and Margot Fonteyne, editors, *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 309–310, Ghent, Belgium, June. European Association for Machine Translation.

Bergmanis, Toms, Martins Kronis, Ingus Jānis Pretkalniņš, Dāvis Nicmanis, Jeļizaveta Jeļinska, Roberts Rozis, Rinalds Vīksna, and Mārcis Pinnis. 2026. Tildeopen llm: Leveraging curriculum learning to achieve equitable language representation.

Communication, Seamless, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. Seamlessm4t: Massively multilingual multimodal machine translation.

Deutsch, Daniel, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria, July. Association for Computational Linguistics.

Dorkin, Aleksei, Taido Purason, Emil Kalbaliyev, Hele-Andra Kuulmets, Marii Ojastu, Mark Fišel, Tanel Alumäe, Eleri Aedmaa, Krister Kruusmaa, and Kairit Sirts. 2026. Estllm: Enhancing estonian capabilities in multilingual llms via continued pretraining and post-training.

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu,

Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Guerreiro, Nuno M., Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Ji, Shaoxiong, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schuetze, Jörg Tiedemann, and Barry Haddow. 2025. EMMA-500: Enhancing massively multilingual adaptation of large language models.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock,

Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July. Association for Computational Linguistics.

Kocmi, Tom, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China, November. Association for Computational Linguistics.

Korotkova, Elizaveta and Mark Fishel. 2024. Estonian-centric machine translation: Data, models, and challenges. In Scarton, Carolina, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Rachel Bawden, Víctor M Sánchez-Cartagena, Patrick Cadwell, Ekaterina Lapshinova-Koltunski, Vera Cabarrão, Konstantinos Chatzitheodorou, Mary Nurminen, Diptesh Kanojia, and Helena Moniz, editors, *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 647–660, Sheffield, UK, June. European Association for Machine Translation (EAMT).

Kudugunta, Sneha, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kuulmets, Hele-Andra, Taido Purason, and Mark Fishel. 2025. How well do LLMs know Finno-Ugric languages? A systematic assessment. In Johansson, Richard and Sara Stymne, editors, *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 340–353, Tallinn, Estonia, March. University of Tartu Library.

Mager, Manuel, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada, July. Association for Computational Linguistics.

Martins, Pedro Henrique, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G.C. de Souza, Alexandra Birch, and André F.T. Martins. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62. Proceedings of the Second EuroHPC user day.

Masciolini, Arianna, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina, and Robert Östling. 2025. The MultiGEC-2025 shared task on multilingual grammatical error correction at NLP4CALL. In Muñoz Sánchez, Ricardo, David Alfter, Elena Volodina, and Jelena Kallas, editors, *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 1–33, Tallinn, Estonia, March. University of Tartu Library.

O'Brien, Dayyán, Bhavitvya Malik, Ona de Gibert, Pinzhen Chen, Barry Haddow, and Jörg Tiedemann. 2025. DocHPLT: A massively multilingual document-level translation dataset. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 286–300, Suzhou, China, November. Association for Computational Linguistics.

Ortega, John E and Kenneth Church. 2023. A research-based guide for the creation and deployment of a low-resource machine translation system. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 813–823.

Pashchenko, Dmytro, Lisa Yankovskaya, and Mark Fishel. 2025. Paragraph-level machine translation for low-resource Finno-Ugric languages. In Johansson, Richard and Sara Stymne, editors, *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 458–469, Tallinn, Estonia, March. University of Tartu Library.

Popović, Maja. 2017. chrF++: words helping character n-grams. In Bojar, Ondřej, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, *Proceedings of the Second Conference on Machine*

Translation, pages 612–618, Copenhagen, Denmark, September. Association for Computational Linguistics.

Purason, Taido, Aleksei Ivanov, Lisa Yankovskaya, and Mark Fishel. 2024. SMUGRI-MT - machine translation system for low-resource Finno-Ugric languages. In Scarton, Carolina, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Mikel Forcada, and Helena Moniz, editors, *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 31–32, Sheffield, UK, June. European Association for Machine Translation (EAMT).

Rei, Ricardo, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Rei, Ricardo, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore, December. Association for Computational Linguistics.

Rei, Ricardo, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms.

Team, NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Trosterud, Trond and Kevin Brubeck Unhammer. 2012. Evaluating North Sámi to Norwegian assimilation RBMT. In España-Bonet, Cristina and Aarne Ranta, editors, *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 13–26, Gothenburg, Sweden, June 13-15.

Yankovskaya, Lisa, Maali Tars, Andre Tättar, and Mark Fishel. 2023. Machine translation for low-resource Finno-Ugric languages. In Alumäe, Tanel and Mark Fishel, editors, *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 762–771, Tórshavn, Faroe Islands, May. University of Tartu Library.