

ReMova: Fine-tuning LLMs for English to Belarusian translation*

Mikita Pilinka¹

Aliaksandr Kliujeŭ²

David Samuel¹

Yves Scherrer¹

¹ University of Oslo

Department of Informatics

{mikitap,davisamu,yvessc}@uio.no

² Independent Researcher

orcid.org/0009-0005-0720-8175

katoshrodingera@protonmail.ch

Abstract

This paper presents a Belarusian-specific data-cleaning pipeline and fine-tuning for English-Belarusian machine translation. Our cleaning pipeline distinguishes itself from others by employing a correction tool that addresses the issue of the two orthographies of the Belarusian language, noise in the training data, interference from other languages and other misspelling issues common in Belarusian on the internet. A matched ablation on unfiltered training data shows substantial benefits from filtering for all fine-tuned models, with the LLM-based models gaining roughly twice as much from filtering as the dedicated encoder-decoder MT system, supporting the view that for Belarusian MT one of the primary bottlenecks is data quality.

1 Introduction

The uneven distribution of languages across the web and the wide variation in available linguistic resources have left many languages, including Belarusian, under-prioritised in the training and evaluation of machine translation systems. We present a Belarusian-specific data-cleaning pipeline and use the resulting corpus to fine-tune translation models from two neural MT paradigms: NLLB (NLLB Team et al., 2022) as an encoder-decoder system, Gemma-4 (Gemma Team et al., 2026) and TranslateGemma (Finkelstein et al., 2026) as decoder-only LLMs. We release our primary model, the fine-tuned TranslateGemma-12B-IT, under the Gemma Terms of Use.¹

1.1 Belarusian variation

Contemporary written Belarusian displays genuine orthographic variation, having two active graphical codified systems: cyrillic (mostly) and latin

(rarely), and featuring two competitive normative systems: the official standard orthography (often informally called *Narkamaŭka*) and the alternative orthography, *Taraškievica*, characterized by lesser codification and the lack of common standardizing authority. Although these names originally referred to different orthographic systems, they subsequently came to denote broader normative systems, thus their differences extend beyond spelling – concerning phonetics, morphology, word formation, vocabulary and syntax. The opposition is not entirely categorical, however, because many grammatical forms, allegedly distinctive for *Taraškievica*, are also admitted, stylistically marked, or used as variants within the official standard (Ramza, 2018). Aparovich et al. (2025, p. 513) found that most digital examples in their sentiment dataset followed the official standard, fewer than 10% used *Taraškievica*, and a small minority used the Latin alphabet *lacinka*. This work focuses exclusively on standard Belarusian orthography. Extending the model to other Belarusian varieties will require dedicated evaluation and training data to ensure adequate, variety-specific performance.

Language interference. The existence of several orthographic traditions in Belarusian complicates the formation of coherent training data. As a result, a model may learn from heterogeneous linguistic material, creating noise in its outputs. As Nazaranka (2025, pp. 18–19, 24–27) notes, this may manifest itself, for example, in the coexistence of elements from two orthographies within a single text, the use of erroneous constructions at different levels of language, or unstable switching between different norms. An additional problem is interference from other closely related languages, particularly Russian and, according to our observations, Ukrainian and Polish. These languages may fill gaps resulting from the limited amount of

*The name combines the prefix *re-* with Belarusian *мова* ('language').

¹<https://huggingface.co/mineralsfree/ReMova-TranslateGemma-en-be>

Belarusian-language training data. Such interference may occur not only at the lexical level but also at the phonetic, morphological, syntactic, and stylistic levels.

It is worth noting that a significant portion of the modern Belarussophone community consists of “new speakers” raised in Russian-speaking families (Woolhiser, 2013, pp. 63–65). In Woolhiser’s pilot sample, performed in 2013, most of these speakers reported shifting to Belarusian in peer-group communication during adolescence or early adulthood, between the ages of 13 and 22 (Woolhiser, 2013, pp. 86–87). Their usage coexists with the two above-mentioned normative systems and displays variation associated with speakers’ differing preferences for forms perceived as more or less divergent from Russian (Woolhiser, 2013, pp. 107–111). Considering current trends in Belarusian language, we suspect that the proportion of “new speakers”, that learned Belarusian as the second language, has only grown since the date of Woolhiser’s work.

Misspellings. Orthographic diversity and language interference should be distinguished from misspellings, pseudo-Belarusian forms, and unreviewed machine translation.

Koshchanka and Martysiuk (2025) identify the low level of proficiency for literary Belarusian, together with the desire to produce texts with minimal effort and an insufficient understanding of automatic translation, as factors contributing to the growing presence of machine-translated language in texts of different genres. According to their analysis, artificial and unnatural forms generated by the early Russian-to-Belarusian translation system *Belazar* entered media texts and, through repetition, gradually began to be perceived as correct. The subsequent use of *Google Translate* and *Yandex Translate* intensified this process, with traces of automatic translation reportedly increasing rapidly across genres. The effects described are not limited to isolated spelling errors. Corpus research reports that machine-translated texts introduce artificially formed lexical items and grammatical constructions into ordinary usage and may produce semantic shifts. This calls into question the value of some Belarusian-language textual data and the extent to which such language can be considered representative for research purposes (Koshchanka and Martysiuk, 2025, p. 146–147) and, by extension, for creating a proper Belarusian dataset.

Liasovich (2024, p. 56) offers a different view, arguing that media texts published online reflect the current state of the Belarusian language’s lexical resources and their ongoing transformation, thereby revealing the condition of contemporary living Belarusian.

1.2 Importance of data cleaning

Looking at the available parallel data, we also noticed the high degree of variability of modern Belarusian language, which emphasizes the need of a proper data cleaning process that distinguishes normative systems, and cleans obvious errors and language interference issues. We propose a balanced approach that does not step into the internal processes of the language change – not imposing subjective views on how the language should change in the future. In our work we try to stabilise the lexical, orthographical and typographical features of the text, according to the standard Belarusian orthography, while accepting new word-forms, without interfering at the syntactical level of the written texts, although we want to mark that syntax should be the matter of future works in this field.

2 Methodology

In this study, we use a custom filtering pipeline, which we refer to as ReMova,² that combines current tooling and quality metrics to extract reliable parallel segments. The pipeline is organised into stages, each of which writes its output to a file and records summary statistics. Some stages apply to both sides of the parallel data, such as *ftfy*,³ NFKC normalisation,⁴ and GlotLID-based language identification,⁵ while others address the aforementioned issues specific to the Belarusian side (*Wikificator*⁶ and *LanguageTool*⁷). A schematic representation of the pipeline is shown in Figure 1.

2.1 Belarusian-specific normalisation

A central element of our approach is the strict handling of misspellings and *Taraškievica* forms. We

²ReMova github: <https://github.com/mineralsfree/ReMova>

³*ftfy* (fixes text for you): <https://github.com/rspeer/python-ftfy>.

⁴NFKC: Unicode Normalisation Form Compatibility Composition; see <https://unicode.org/reports/tr15/>.

⁵GlotLID: <https://github.com/cisnlp/GlotLID>.

⁶Wikificator: <https://be.wikipedia.org/wiki/Вікіпедыя:Вікіфікатар>.

⁷LanguageTool: <https://languagetool.org/>.

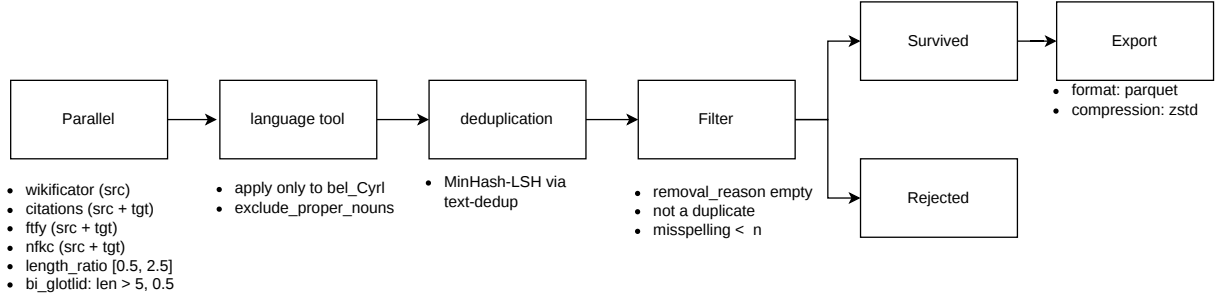


Figure 1: The multiple stages of the ReMoVA cleaning pipeline employed to ensure the quality of the English-Belarusian parallel dataset.

rely on Wikificator (Belarusian Wikipedia contributors, 2026), a set of regular expressions maintained by the Belarusian Wikipedia community to detect and correct the most common orthographic errors and *Taraškievica* variants found in Belarusian text. We complement this with LanguageTool (LanguageTool, 2026), an open-source proof-reading platform whose Belarusian module provides a spellchecker based on the grammatical base of the National Corpus of the Belarusian Language (Koshchanka, 2018). The National Corpus of the Belarusian Language, which is compiled by the Institute of Belarusian Language, currently recommends LanguageTool as the main spellchecker solution for the Belarusian language and contributes to the project (National Corpus of the Belarusian Language, 2026). We therefore considered it the most advanced and reliable solution for our spell-checking workflow. While Wikificator targets a fixed set of high-frequency patterns, LanguageTool offers broader dictionary-based coverage, catching misspellings and grammatical inconsistencies that fall outside the regex rules.

The different stages of the pipeline have distinct roles. Most either normalise text in place (ftfy, NFKC, Wikificator) or annotate each segment with metrics that downstream stages consume: GlotLID adds language-identification confidence, LanguageTool tags the Belarusian side with match counts and match density (matches per non-proper-noun word), and deduplication marks near-duplicates. Removal is concentrated at two places. A dedicated filter stage applies thresholds on the accumulated metrics. Before running the pipeline, we discard trivially short pairs (two words or fewer).

Table 1 shows the effects of the different stages on one of the utilised corpora, Tatoeba. It can be seen that the filter stage dominates the rest:

Stage	In	Out	Notable effect
prep	8,980	7,899	1,081 pairs with ≤ 2 words dropped
parallel	7,899	7,899	271 typography, 201 mojibake, 101 NFKC
languagetool	7,899	7,899	1,051 rows with LT matches (13.3%)
dedup	7,899	7,899	near-duplicates marked (signal-only)
filter	7,899	6,953	88% survived; 692 LT errors, 231 dup, 23 lang/len
BLASER	6,953	6,749	204 low-quality pairs dropped (2.9%)
export	6,749	6,749	~285 KB parquet (ZSTD)

Table 1: Tatoeba (Belarusian–English) pipeline execution. 8,980 raw pairs $\rightarrow$ 6,749 training pairs (75% overall yield).

about 12% of the pairs are dropped there, mostly because they exceed the manually selected LanguageTool density threshold (692 pairs, a proxy for *Taraškievica* influence and chronic misspelling) or were flagged as duplicates (231 pairs). Because the density metric is sensitive to LanguageTool’s coverage of Belarusian, we iteratively curated its dictionary: the most frequent matches were either added with their correct spelling or marked as ignorable where the match was a false positive.

2.2 Training data

We selected the training dataset through both human review and pipeline-based evaluation. Rather than relying on a single corpus, we combined five sources with per-corpus BLASER (Communication et al., 2023) cutoffs calibrated to their register:

- HPLT web-crawled bitext (Burchell et al., 2025),
- OpenSubtitles-derived paragraph alignments (upsampled $2\times$ to enrich the spoken register; Lison and Tiedemann, 2016),

- Tatoeba crowdsourced pairs (upsampled $2\times$; Tiedemann, 2020),
- Wikimedia aligned paragraphs (Schwenk et al., 2021a), and
- back-translated segments from a curated subset of HPLT’s monolingual Belarusian, itself derived from Common Crawl and the Internet Archive (upsampled $2\times$).

Table 2 summarises the composition. Per-corpus thresholds were necessary because BLASER’s written-text calibration systematically under-scores subtitle-style compression even for faithful translations; a uniform cutoff would have discarded most of the spoken-register signal. Other OPUS subcorpora such as WikiMatrix were excluded, as many of their text segments are misaligned.

To strengthen the coverage of informal Belarusian, we augmented the training set with back-translated pairs from selected Common Crawl web resources. To improve the ability of our models to work with spoken registers we created a whitelist-curated subset of the HPLT monolingual Belarusian collection (2,300 segments after cleaning) in two particular registers: ID (Interactive Discussion) and SP (Spoken). The resulting segments were back-translated into English twice with a large language model DeepSeek-V4-Flash (DeepSeek-AI et al., 2026), yielding 4,600 BE-EN pairs whose two targets are textually distinct in 84% of cases and thus function as paraphrastic augmentation.

3 Experiments

3.1 Identifying strongest baseline

To get a fuller picture, we fine-tune models from both paradigms: an encoder-decoder system and a decoder-only LLM. We target the constrained track which requires releasing a model under 20B parameters under a licence permitting unrestricted non-commercial use (WMT26 General Machine Translation Task Organizers, 2026). We compare a subset of the suggested base models (Gemma-3 (Gemma Team et al., 2025), Aya (Üstün et al., 2024), Qwen-2.5 (Yang et al., 2025)) along with Llama-3.2 (Grattafiori et al., 2024) before fine-tuning to identify the strongest baseline for Belarusian translation. As the encoder-decoder counterparts, we use both NLLB and MADLAD-400 (Kudugunta et al., 2023), a comparably positioned Google model covering 400 languages.

Corpus	BLASER $\geq$	Rows	$\times$	Share
HPLT (web)	3.5	83,770	1	49.4%
OpenSubtitles	3.0	29,849	2	35.2%
Tatoeba	3.5	6,749	2	8.0%
Wikimedia	3.7	7,923	1	4.7%
ID/SP (HPLT)	–	2,300	2	2.7%
Total				169,489

Table 2: Composition of the fine-tuning corpus. Per-corpus BLASER thresholds are calibrated to each source’s register; the $\times$ column reports the upsampling factor and *Share* is the resulting share of the combined training set. ID/SP (HPLT) denotes back-translated in-domain parallel text pairs generated from a whitelist-curated subset of HPLT monolingual Belarusian segments.

An initial experiment compared base models under 4B parameters, a cap chosen to allow rapid iteration within the available compute budget. Note that no fine-tuning was applied at this stage; the goal was to identify the strongest base model under the parameter budget before committing compute to adapter training.

3.2 Fine-tuning

The strongest models were chosen for fine-tuning and further evaluation on the same metrics and dataset: full fine-tuning for the encoder-decoder (NLLB) via Seq2SeqTrainer, and LoRA-based supervised fine-tuning of the LLMs (Gemma-4-12B-IT and TranslateGemma-12B-IT) via TRL’s SFTTrainer. Both regimes use AdamW with warmup ratio 0.05; weight decay is 0.01 for NLLB and 0.0 for the LLMs. LoRA is applied to all-linear modules for Gemma-4; TranslateGemma restricts the target set to `q,k,v,o,gate,up,down` projections to avoid adapting the vision encoder. The parameters used for fine-tuning selected models are listed in Table 3. Our training corpus mixes sentence pairs (Tatoeba) with paragraph-level and multi-sentence alignments (Wikimedia, OpenSubtitles, HPLT); the chosen `max_length` budgets (128 per side for NLLB, 256 for Gemma-4 and TranslateGemma) cover the vast majority of pairs, and BOUQuET is evaluated at its native sentence granularity.

3.3 Filtering ablation

To quantify the contribution of the ReMoVA filtering pipeline, we run an ablation in which the selected models are re-fine-tuned on unfiltered data

	NLLB FT	LLM LoRA SFT [†]
Base model	nllb-200-3.3B	gemma-4-12b-it
Epochs	3	2
Eff. batch size	32	16
Learning rate	3×10^{-5}	1×10^{-4}
Max seq length	128	256
LoRA ($r / \alpha / \text{drop}$)	—	16 / 32 / 0.05

Table 3: Hyperparameters for fine-tuning. [†]Both Gemma-4-12B-IT and TranslateGemma-12B-IT share this LoRA recipe.

drawn from the same source corpora. The ablation run matches the cleaned run in both total volume and per-source proportions, so that any difference in downstream translation quality can be attributed to filtering rather than to changes in training data quantity or composition. All hyperparameters (Table 3) and the evaluation protocol are kept identical to the main run.

3.4 Evaluation

We use the recent BOUQuET evaluation dataset (Andrews et al., 2025) for its cross-domain sentence coverage and human-verified translations, which lets us track per-domain performance improvements from fine-tuning. The test split contains 854 unique sentences (198 paragraphs), written in the standard Belarusian orthography, across eight domains: comments, conversation, how-to/instructions, narration, other miscellaneous, reflection, social posts, and web miscellaneous, matching the row labels in Table 5. The dev split was used for best-checkpoint selection; all reported metrics come from the held-out test set.

To capture multiple dimensions of translation quality, we report BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017) as surface-level metrics. ChrF++ is more informative for morphologically rich languages, but reporting both is preferable in low-resource settings, as the two metrics expose different translation artefacts (Kumar et al., 2026). We additionally report COMET (Rei et al., 2020), a neural metric that correlates more closely with human judgements. Base LLMs were evaluated with a fixed 3-shot prompt of simple, short demonstration pairs, identical across all base-LLM evaluations.

Model	Size	BLEU	chrF++	COMET
<i>Dedicated MT models</i>				
NLLB-200	3B	27.2	51.4	86.7
MADLAD-400	3B	21.8	41.6	81.2
<i>Base LLMs</i>				
Gemma-4 [†]	12B	25.2	49.1	84.8
TranslateGemma ^{†‡}	12B	25.0	51.7	89.6
Gemma-3	4B	22.9	46.4	84.5
Llama-3.2	3B	10.4	30.7	69.6
Aya	3B	4.6	19.8	50.5
Qwen-2.5	3B	3.6	18.5	48.0

Table 4: English→Belarusian on BOUQuET test; best per family in bold. Base LLMs use a fixed 3-shot prompt. [†]Above the 4B screening cap. [‡]Native chat template (not 3-shot).

4 Results and discussion

4.1 Base models

Base model results are summarised in Table 4. Within the dedicated MT family, NLLB-200 outperforms MADLAD-400 by a wide margin at comparable scale (27.2 vs. 21.8 BLEU), consistent with NLLB’s targeted Belarusian coverage. Among base LLMs, the Gemma family stands apart: Gemma-3 reaches 22.9 BLEU at 4B. The remaining LLMs trail far behind: Llama-3.2 (10.4 BLEU) recovers only lexical fragments, while Aya and Qwen-2.5 (3.6–4.6 BLEU) are essentially non-functional despite Aya and Qwen-2.5 being positioned as multilingual models, suggesting scale alone does not compensate for limited Belarusian pretraining exposure. Only NLLB-200 and the Gemma family clear the bar for productive fine-tuning; we advance NLLB-200 to full fine-tuning and Gemma-4-12B-IT together with TranslateGemma-12B-IT to the LoRA track.

4.2 Fine-tuned models

Fine-tuning results for NLLB, Gemma-4, and TranslateGemma are presented in Table 5, alongside a matched ablation in which the same models are re-fine-tuned on unfiltered data from the same source corpora (§3.3, Table 3).

All three fine-tuned models cluster within a narrow band on aggregate metrics, with TranslateGemma leading both BLEU (32.63) and chrF++ (55.36), Gemma-4 close behind on BLEU (31.40), and NLLB slightly ahead of Gemma-4 on chrF++ (Table 5). The close finish is itself notable: a general-purpose LLM (Gemma-4-12B)

		BLEU			chrF++			COMET		
Domain	<i>n</i>	base	unfilt.	+FT	base	unfilt.	+FT	base	unfilt.	+FT
<i>NLLB-200-3.3B</i>										
Comments	73	25.22	31.18	32.83	48.17	51.23	52.72	83.07	85.76	86.40
Conversation	186	26.83	34.92	35.18	49.43	53.62	54.66	85.87	89.27	90.07
How-to/instr.	107	25.77	24.29	27.68	50.65	50.42	53.19	84.10	83.66	85.33
Narration	98	27.41	30.95	32.57	51.68	56.08	56.04	85.33	87.29	88.56
Other misc.	95	28.98	30.03	32.38	53.46	54.70	55.46	88.99	89.79	91.13
Reflection	88	25.95	24.42	27.43	52.39	51.86	52.91	89.09	88.78	89.56
Social posts	104	30.35	30.58	32.30	52.87	53.26	53.93	88.39	88.26	89.88
Web misc.	103	24.20	24.67	26.30	51.17	52.82	53.86	88.65	89.08	90.77
<i>Overall</i>	854	27.24	29.23	31.22	51.35	53.09	54.19	86.67	87.90	89.12
<i>Gemma-4-12B-IT</i>										
Comments	73	26.49	28.47	28.66	48.06	49.87	49.77	83.58	82.33	86.65
Conversation	186	32.54	31.61	34.74	50.64	52.12	52.67	86.63	87.48	87.78
How-to/instr.	107	21.60	23.81	28.88	48.56	51.13	52.88	82.89	85.04	86.37
Narration	98	24.41	27.97	33.46	49.31	53.55	54.92	82.25	87.09	88.80
Other misc.	95	24.64	28.29	30.17	48.75	53.07	53.95	85.35	89.19	90.41
Reflection	88	22.20	23.45	29.73	48.24	50.65	53.03	84.49	86.98	90.29
Social posts	104	26.09	30.75	32.64	50.37	53.57	55.69	86.22	87.59	90.17
Web misc.	103	19.22	23.83	28.05	47.53	51.80	54.56	85.06	88.73	90.93
<i>Overall</i>	854	25.23	27.78	31.40	49.08	52.11	53.64	84.80	86.99	88.85
<i>TranslateGemma-12B-IT</i>										
Comments	73	25.33	26.77	30.93	50.65	49.50	53.08	88.84	83.16	87.57
Conversation	186	28.78	32.82	37.90	50.18	51.60	55.19	88.83	87.90	89.01
How-to/instr.	107	24.47	25.66	32.65	52.69	52.30	57.45	87.90	84.49	88.57
Narration	98	24.54	29.42	34.09	52.09	54.96	57.67	88.73	87.41	89.64
Other misc.	95	27.79	31.49	31.40	54.05	56.41	53.88	90.95	89.33	91.79
Reflection	88	20.29	24.29	27.80	50.33	51.34	53.38	90.52	88.29	91.17
Social posts	104	24.98	28.52	32.15	51.96	52.60	55.34	90.69	86.84	90.75
Web misc.	103	21.49	23.62	28.52	51.59	52.14	55.67	91.24	89.92	92.21
<i>Overall</i>	854	24.98	28.25	32.63	51.73	52.73	55.36	89.63	87.33	90.03

Table 5: Per-domain English→Belarusian test scores. Columns: *base* (pre-fine-tune baseline), *unfilt.* (fine-tuned on unfiltered data from the same sources), *+FT* (fine-tuned on the ReMoVA-cleaned corpus). Best per row in bold.

with LoRA fine-tuning on cleaned data reaches parity with a fully fine-tuned dedicated MT model (NLLB-3.3B) without any translation-specific pre-training objective.

For TranslateGemma specifically, fine-tuning yields a large lexical improvement (+7.65 BLEU, +3.63 chrF++). Per-domain scores, presented in Table 5, show that fine-tuning improves BLEU in every domain, but by markedly uneven margins: gains range from +9.55 BLEU on Narration and +9.12 on Conversation down to +3.61 on Other misc., the only domain where chrF++ does not improve (−0.17).

Per-domain BLEU gains from fine-tuning vary across models, though Narration improves substantially for all three (+5.2 for NLLB, +9.1 for Gemma-4, +9.6 for TranslateGemma). This is plausibly driven by OpenSubtitles, which contributes 35% of the training data and was up-sampled 2× to enrich the spoken register (§2). The distribution of smaller gains differs by model: NLLB benefits least on Reflection and How-to;

Gemma-4 on Comments and Conversation and TranslateGemma on Other miscellaneous. Diversifying the stylistic range of training data is a natural direction for future work.

4.3 Filtering ablation

To measure the contribution of the ReMoVA cleaning pipeline in isolation, we compare each model’s cleaned fine-tune against a matched fine-tune on unfiltered data from the same source corpora (Table 5, “unfilt.” column). Cleaning delivers +1.99 BLEU / +1.10 chrF++ for NLLB, +3.62 / +1.53 for Gemma-4, and +4.38 / +2.63 for TranslateGemma (COMET deltas follow the same ordering). The pattern appears to split along the encoder-decoder-MT versus decoder-only-LLM divide rather than along model size or degree of translation specialisation: NLLB, whose pretraining objective is already close to supervised parallel translation, gains roughly two times less from cleaning than either decoder-only LLM. Notably, TranslateGemma’s translation-specialisation does

not close this gap; despite reaching the highest overall quality after fine-tuning, it shows the largest cleaning gain of the three.

At the per-domain level, some cases sharpen this story. For Gemma-4 Conversation, unfiltered fine-tuning actually drops BLEU below the 3-shot base (31.61 vs. 32.54), while cleaned fine-tuning delivers a clear gain (34.74): fine-tuning on noisy data can be net-negative for that domain. On BLEU, cleaning wins every domain for every model except TranslateGemma@Other miscellaneous (−0.09); on chrF++, cleaning wins every domain except three isolated cells (NLLB@Narration, Gemma-4@Comments, and TranslateGemma@Other miscellaneous), consistent with a slightly wider surface-form vocabulary in the unfiltered pool, helping character n-gram matching on domains dominated by rare terms.

As shown in our pipeline results (Table 1), even datasets considered “cleaned” upstream retain substantial noise and require further downstream filtering; we leave deeper investigation to future work. Taken together, the ablation supports the paper’s data-quality thesis: for the English–Belarusian pair, systematic filtering of parallel data is a first-order lever on translation quality rather than a secondary tweak.

5 Related Work

Multilingual translation. While high-resource languages benefit from large parallel corpora suitable for training and evaluating translation models, many low-resource languages long lacked sufficient support. Yet, openly released encoder-decoder systems have been progressively expanding the language coverage: OPUS-MT provided per-pair models trained on the OPUS collection (Tiedemann, 2012; Tiedemann and Thottingal, 2020); M2M-100 introduced many-to-many translation across 100 languages (Fan et al., 2021); NLLB scaled this recipe to more than 200 languages (NLLB Team et al., 2022); MADLAD-400 extended coverage past 400 (Kudugunta et al., 2023); and the recently-announced Omnilingual MT targets 1 600 languages (Alastruey et al., 2026).

LLM-based translation. In parallel with the development of dedicated encoder-decoder systems, the scaling of LLMs has opened an alternative path to translation. Off-the-shelf language models were found to already translate well between high-

resource languages using few-shot learning (Garcia et al., 2023), but fall behind dedicated systems when translating low-resource languages (Robinson et al., 2023). However, fine-tuning the LLMs can close this gap – with approaches like the Advanced Language Model-based translator (ALMA; Xu et al., 2024). Finetuned from Llama 2 (Touvron et al., 2023), ALMA outperformed the 54B-parameter NLLB model despite its much smaller size of 7B parameters. Its training follows a two-stage recipe: continued pretraining on monolingual data in the target languages, followed by fine-tuning on a modest amount of high-quality bilingual corpus. The authors argue that excessive training on parallel data washes out the LLM’s pre-existing knowledge, although a subsequent study (Iyer et al., 2024) shows that this finding does not hold for low-resource languages with limited pre-training coverage. Subsequent work has further refined this recipe: Tower mixes translation with related tasks during fine-tuning, further improving translation quality (Alves et al., 2024; Rei et al., 2024). Its successor, Tower+, scales data quality, quantity, and model size and additionally employs reinforcement learning to refine the generated outputs (Rei et al., 2026). TranslateGemma applies a similar recipe to the open Gemma 3 models, providing one of our fine-tuning bases (Finkelstein et al., 2026).

WMT General MT Task. WMT has been focusing on machine translation of low-resource languages for many years: in 2022, the shared task included Livonian and Yakut languages (Kocmi et al., 2022), and the range of low-resource languages has further expanded in the following iterations. In 2023, translation systems based on LLMs started to be competitive in WMT (Kocmi et al., 2023), and then in 2024, finetuned LLM systems started outperforming the standard encoder-decoder approaches (Kocmi et al., 2024). This trend aligns with our findings that the best performance on English–Belarusian translation is achieved with a finetuned LLM.

Data quality and filtering. Much of the parallel data behind multilingual MT is web-mined with cross-lingual sentence embeddings: LASER-based mining (Artetxe and Schwenk, 2019) produced corpora such as CCAIaligned, CCMATRIX or WikiMATRIX (El-Kishky et al., 2020; Schwenk et al., 2021b,a), and NLLB mined its low-resource bi-text with distilled LASER 3 encoders (NLLB Team

et al., 2022). We are using a later improvement of LASER called BLASER in this study (Communication et al., 2023). These heuristically-aligned corpora are known for systematic quality issues, especially for low-resource languages (Kreutzer et al., 2022). Proper filtering of these corpora has thus become an essential step in training NMT systems, supported by tools such as Zipporah (Xu and Koehn, 2017), Bicleaner and Bifixer, its neural successor (Ramírez-Sánchez et al., 2020; Zaragoza-Bernabeu et al., 2022), as well as the modular OpusFilter and OpusCleaner toolkits (Aulamo et al., 2020; Bogoychev et al., 2023). These pipelines combine many approaches but remain largely language-agnostic; our work follows the same overall design while adding Belarusian-specific orthographic normalisation and spellchecking that generic tools cannot provide.

6 Conclusion

In this work, we presented REMOVA, a Belarusian-specific data-cleaning pipeline, and fine-tuned translation models for the English–Belarusian language pair across both the encoder-decoder and decoder-only paradigms. The pipeline combines language identification, Unicode normalisation, spelling correction with Wikificator and LanguageTool, detection of misaligned parallel segments, and filtering or replacement of *Taraškievica* and obvious non-Belarusian forms; applying it exposed substantial noise even in datasets we initially considered higher-quality. On the cleaned corpus, full fine-tuning of NLLB-200-3.3B, LoRA fine-tuning of Gemma-4-12B-IT, and LoRA fine-tuning of TranslateGemma-12B-IT produced comparable performance on the BOUQuET test set, with TranslateGemma leading on every reported metric (32.63 BLEU, 55.36 chrF++, 90.03 COMET) and Gemma-4 reaching parity with the fully fine-tuned dedicated MT baseline without any translation-specific pretraining objective. A matched ablation in which each model was re-fine-tuned on unfiltered data from the same sources showed substantial BLEU gains from filtering for all three models, with the decoder-only LLMs (Gemma-4 +3.62, TranslateGemma +4.38) gaining roughly two times more than the dedicated encoder-decoder MT system NLLB (+1.99); chrF++ and COMET follow the same architectural split. Taken together, these results support the hypothesis that for the

English–Belarusian pair, data quality is a first-order lever on translation quality.

Several directions remain for future work. Scaling back-translation beyond the current 2,300 curated HPLT segments (using a larger monolingual pool, additional back-translation models, and pivoting through related higher-resource pairs such as Belarusian–Russian) could broaden training coverage beyond OPUS-derived sources. Diversifying the stylistic range of training data would likely close the per-domain gap on formal and reflective content. Finally, human evaluation and statistical-significance testing would give the model rankings and ablation deltas reported here a firmer footing than metric-based comparison alone.

Limitations

We did not conduct a thorough human evaluation, which would provide the strongest signal of translation quality and help analyse mistakes. Metrics such as BLEU underestimate quality for morphologically rich languages like Belarusian and may correlate weakly with human judgements. COMET was trained on human judgements dominated by high-resource language pairs and has not been meta-evaluated on Belarusian; absolute COMET values reported here should be treated as internally consistent within this paper rather than directly comparable to published COMET numbers on other language pairs. We did not test the statistical significance of our results, which would be especially valuable for the smaller ablation gaps and the tight ordering between models.

Some larger or potentially more capable models for Belarusian translation were not evaluated. We did not benchmark against commercial systems such as Yandex Translate or Google Translate, which cover Belarusian; comparison with a strong commercial baseline would help calibrate our absolute scores against production-quality MT and is left for future work. The few-shot examples used for base-LLM evaluation may have affected performance and should be varied in future evaluations to test prompt sensitivity.

Compute constraints motivated the LoRA-only setup for the LLM fine-tuning and the 4B parameter cap on the initial base-LLM screen; both choices may have affected the absolute scores and the cross-model comparisons. The hyperparameter search was limited, and we cannot guarantee the chosen settings were optimal.

Use of AI assistants

During the preparation of this work, the authors used Claude (Anthropic, `claude-opus-4-7`, `claude-opus-4-8`, `claude-opus-5-0`) and GPT (OpenAI, `gpt-5.6-sol`) to assist with writing the paper, specifically for refining wording and proof-reading. All AI-assisted output was reviewed and edited by the authors, who take full responsibility for the correctness, design decisions, and content of the submitted work. No sensitive or personal data was entered into the tools, and all data handling complied with the University of Oslo's data regulations.

References

- Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, Artyom Kozhevnikov, Christophe Ropers, Eduardo Sánchez, Charles-Eric Saint-James, Ioannis Tsiamas, Xiang "Tony" Cao, Chierh Cheng, Joe Chuang, Paul-Ambroise Duquenne, Mark Duppenhaler, Nate Ekberg, Cynthia Gao, Pere Lluís Huguet Cabot, João Maria Janeiro, Jean Maillard, and 12 others. 2026. [Omnilingual mt: Machine translation for 1,600 languages](#). *Preprint*, arXiv:2603.16309.
- Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. [Tower: An open multilingual large language model for translation-related tasks](#). *Preprint*, arXiv:2402.17733.
- Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Mark Duppenhaler, Nathaniel Paul Ekberg, Cynthia Gao, Daniel Edward Licht, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, and Shireen Yates. 2025. [BOUQuET : dataset, benchmark and open initiative for universal quality evaluation in translation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27515–27535, Suzhou, China. Association for Computational Linguistics.
- Maksim Aparovich, Volha Harytskaya, Vladislav Poritski, Oksana Volchek, and Pavel Smrz. 2025. [BelarusianGLUE: Towards a natural language understanding benchmark for Belarusian](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–527, Vienna, Austria. Association for Computational Linguistics.
- Mikel Artetxe and Holger Schwenk. 2019. [Margin-based parallel corpus mining with multilingual sentence embeddings](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.
- Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. [OpusFilter: A configurable parallel corpus filtering toolbox](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156. Association for Computational Linguistics.
- Belarusian Wikipedia contributors. 2026. [Vikipiedyja:Hadžety/Vikifikatar \[Wikipedia:Gadgets/Wikificator\]](#). Accessed: 2026-05-20.
- Nikolay Bogoychev, Jelmer van der Linde, Graeme Nail, Barry Haddow, Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Lukas Weymann, Tudor Nicolae Mateiu, Jindřich Helcl, and Mikko Aulamo. 2023. [Opuscleaner and opustrainer, open source toolkits for training machine translation and large language models](#). *Preprint*, arXiv:2311.14838.
- Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 2025. [An expanded massive multilingual dataset for high-performance language technologies \(HPLT\)](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.
- Seamless Communication, 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, and 49 others. 2023. [Seamlessm4t: Massively multilingual & multimodal machine translation](#). *Preprint*, arXiv:2308.11596.
- DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, and 300 others. 2026. [Deepseek-v4: Towards highly efficient million-token context intelligence](#). *Preprint*, arXiv:2606.19348.
- Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. [CCAligned: A massive collection of cross-lingual web-document pairs](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Man-deep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).
- Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. [TranslateGemma technical report](#). *Preprint*, arXiv:2601.09012.
- Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.
- Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026. [Gemma 4 technical report](#). *Preprint*, arXiv:2607.02770.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Aaron Grattafiori, 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, and 542 others. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Yuras Hetsevich, Jauheniya Zianouka, David Latyshevich, Mikita Suprunchuk, and Valer Varanovich. 2022. [Clarín knowledge centre for belarusian text and speech processing \(k-blp\)](#). In *Selected Papers from the CLARIN Annual Conference 2021*, volume 189 of *Linköping Electronic Conference Proceedings*, pages 46–55, Linköping. Linköping University Electronic Press.
- Vivek Iyer, Bhavitvya Malik, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. [Quality or quantity? on data scale and diversity in adapting large language models for low-resource translation](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1393–1409, Miami, Florida, USA. Association for Computational Linguistics.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. [Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. [Findings of the 2023 conference on machine translation \(WMT23\): LLMs are here but not quite there yet](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.
- Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. [Findings of the 2022 conference on machine translation \(WMT22\)](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Uladzimir Koshchanka. 2018. [Nacyjannaŭny korpus bielaruskaj movy ŭ kanteksie korpusnaj linhvistyki slavianskich krain \[The National Corpus of the Belarusian Language in the Context of Corpus Linguistics of Slavic Countries\]](#). In *XVI International Congress of Slavists*, Belgrade.
- Uladzimir Koshchanka and Viačaslaŭ Martysiuk. 2025. [N-corpus albaruthĖnicum, quo vadis?](#)
- Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Alahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. [Quality at a glance: An audit of web-crawled multilingual datasets](#). *Transactions of the Association for Computational Linguistics*, 10:50–72.

- Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. *Madlad-400: A multilingual and document-level large audited dataset*. *Preprint*, arXiv:2309.04662.
- Sanjeev Kumar, Preethi Jyothi, and Pushpak Bhat-tacharyya. 2026. *Evaluating extremely low-resource machine translation: A comparative study of chrF++ and bleu metrics*. *Preprint*, arXiv:2602.17425.
- LanguageTool. 2026. LanguageTool: Free AI Grammar Checker. <https://languagetool.org/>. Accessed: 2026-05-20.
- Sviatlana Liasovich. 2024. *Abahulniennie dosviedu bielaruskich dasliedčykaŭ u halinie korpusnaj linhvistyki: sučasny stan, zadačy, mietadyčny patencyjal linhvistyčnych karpusoŭ* [summary of the experience of belarusian researchers in the field of corpus linguistics: modern state, tasks, methodological potential of linguistic corpora]. In *Bielaruskaja mova i movaznaŭstva: stan, vyklyki, praktyčnyja zachady: materyjaly navukova-praktyčnaje kanferencyi, Bielastok, 28–29 červienia 2024 h.*, pages 49–58, Bielastok. Ekapres.
- Pierre Lison and Jörg Tiedemann. 2016. *OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles*. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).
- National Corpus of the Belarusian Language. 2026. *Pravierka pravapisu* [Spell checking]. <https://bnkorporus.info/spell.html>. Accessed: 2026-05-20.
- Tatsiana Nazaranka. 2025. Navigating language contact in the digital age. *Didaktik slawischer Sprachen*, pages 17–32.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, 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, and 20 others. 2022. *No language left behind: Scaling human-centered machine translation*. *Preprint*, arXiv:2207.04672.
- 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.
- Maja Popović. 2017. *chrF++: words helping character n-grams*. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. *Bifixer and bicleaner: two open-source tools to clean your parallel data*. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.
- Tacciana Ramza. 2018. *Taraškievica i narkamaŭka: ci jość hramatyčnaje procistajannie?* [taraškievica and narkamaŭka: is there a grammatical opposition?]. *Rozprawy Komisji Językowej Łódzkiego Towarzystwa Naukowego*, 66:409–432.
- Ricardo Rei, Nuno M Guerreiro, José Pombal, João Alves, Amin Farajian, Pedro Teixeira, and Andre Martins. 2026. *TOWER+: Bridging generality and translation specialization in multilingual LLMs*. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29614–29635, San Diego, California, United States. Association for Computational Linguistics.
- Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. *Tower v2: Unbabel-IST 2024 submission for the general MT shared task*. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. *COMET: A neural framework for MT evaluation*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. *ChatGPT MT: Competitive for high- (but not low-) resource languages*. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.
- Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. *WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia*. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.
- Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. *CCMatrix: Mining billions of high-quality*

- parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.
- Jörg Tiedemann. 2012. [Parallel data, tools and interfaces in OPUS](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).
- Jörg Tiedemann. 2020. [The tatoeba translation challenge – realistic data sets for low resource and multilingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.
- Jörg Tiedemann and Santhosh Thottingal. 2020. [OPUS-MT – building open translation services for the world](#). In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- WMT26 General Machine Translation Task Organizers. 2026. Shared Task: General Machine Translation. <https://www2.statmt.org/wmt26/translation-task.html>. EMNLP 2026 Eleventh Conference on Machine Translation; last updated 2026-05-06; accessed 2026-05-20.
- Curt Woolhiser. 2013. New speakers of Belarusian: Metalinguistic discourse, social identity, and language use. In David M. Bethea and Christina Y. Bethin, editors, *American Contributions to the 15th International Congress of Slavists, Minsk, August 2013*, pages 63–115. Slavica Publishers, Bloomington, IN.
- Xenova. [The tokenizer playground](#). Hugging Face Space.
- Hainan Xu and Philipp Koehn. 2017. [Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950, Copenhagen, Denmark. Association for Computational Linguistics.
- Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. [A paradigm shift in machine translation: Boosting translation performance of large language models](#). *Preprint*, arXiv:2309.11674.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. [Bicleaner AI: Bicleaner goes neural](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France. European Language Resources Association.
- Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. [Aya model: An instruction finetuned open-access multilingual language model](#). *Preprint*, arXiv:2402.07827.

A Belarusian tokenisation across tokenisers

To get insights on Belarusian cyrillic text tokenisation, we constructed the evaluation sample of 1 013 384 characters from the Belarusian Wikipedia articles listed by *Special:LongPages*, processing them in descending page-length order, and made a benchmarking on several tokenisers. Articles were retrieved as plain text through the MediaWiki API. We retained only paragraphs classified as Belarusian whose alphabetic characters belonged exclusively to the Belarusian alphabet, and concatenated complete article extracts until the sample exceeded one million characters. The resulting text was passed unchanged to each tokeniser.

For the Benchmark we used several tokenisers, available at The Tokeniser Playground ([Xenova](#)) and the tokeniser from CLARIN project, specifically made for Belarusian language ([Hetseovich et al., 2022](#)).

Tokeniser	Token count	Token/character
Xenova/gpt-4	663445	0,655
Xenova/claude-tokeniser	684587	0,676
Xenova/gemma-tokeniser	448819	0,443
Xenova/llama-tokeniser	519320	0,512
CLARIN Tokeniser	326903	0,323

Table 6: Token counts for the same Belarusian text sample of 1 013 384 characters and 263 707 words from Belarusian wikipedia *Special:LongPages* at the moment of 2 August 2026.