The Kyrgyz Seed Dataset Submission to the WMT25 Open Language Data Initiative Shared Task

Murat Jumashev
jumasheff@gmail.com

Alina Tillabaeva
alinatillabaeva42@gmail.com

Aida Kasieva
aida.kasieva@manas.edu.kg

Turgunbek Omurkanov
omurkanov@gmail.com

Akylai Musaeva
2101.10001@manas.edu.kg

Meerim Emil kyzy
kumaodoru43@gmail.com

Gulaiym Chagataeva
chagataevag.it@gmail.com

Jonathan Washington
jonathan.washington@swarthmore.edu

Abstract

We present a Kyrgyz language seed dataset as part of our contribution to the WMT25 Open Language Data Initiative (OLDI) shared task. This paper details the process of collecting and curating English–Kyrgyz translations, highlighting the main challenges encountered in translating into a morphologically rich, low-resource language. We demonstrate the quality of the dataset through fine-tuning experiments, showing consistent improvements in machine translation performance across multiple models. Comparisons with bilingual and MNMT Kyrgyz-English baselines reveal that, for some models, our dataset enables performance surpassing pretrained baselines in both English–Kyrgyz and Kyrgyz–English translation directions. These results validate the dataset's utility and suggest that it can serve as a valuable resource for the Kyrgyz MT community and other related low-resource languages.

1 Introduction

While machine translation has advanced significantly, progress remains uneven, with morphologically rich, low-resource languages facing substantial obstacles (Goyal et al., 2022). This disparity is particularly acute for the Turkic language family, where agglutinative structures pose unique challenges to standard MT architectures. Kyrgyz (кыргыз тили), a Turkic language with approximately 5.5 million speakers, exemplifies this situation. Despite a growing number of large, automatically-mined parallel datasets (Team et al., 2022), there is a significant shortage of high-quality, human-curated resources essential for building robust and reliable translation systems.

This paper details our contribution to the Open Language Data Initiative (OLDI) shared task: the creation of a high-quality, human-validated English-Kyrgyz dataset of 6,193 sentence pairs. The source text, drawn from diverse scientific domains on Wikipedia, provides rich terminological and syntactic complexity that inherently challenges MT systems. For a low-resource language like Kyrgyz, this complexity exposed a more fundamental obstacle: the absence of standardized scientific vocabulary. This makes high-quality human translation not just a matter of review, but of active linguistic curation. Our translation process, therefore, required developing novel strategies to handle significant terminological gaps and neologisms. This involved our team of native speakers making crucial terminological choices. For instance, we prioritized the native Kyrgyz word "дене" for scientific compounds like "celestial body" ("космостук дене") and "antibody" ("антидене") over the somewhat established Russian calque "тело", a decision that reflects modern usage and enhances both accuracy and naturalness.

Our main contributions are as follows:

1. We contribute the first high-quality, human-validated English-Kyrgyz seed dataset to the OLDI initiative.

2. We provide a detailed analysis of the key linguistic challenges in English-to-Kyrgyz translation, particularly regarding terminological adaptation, and document the effective strategies our team employed.

3. We conduct fine-tuning experiments with four major NMT models (mT5, mBART, M2M100, and NLLB-200) to empirically demonstrate that our high-quality dataset provides consistent performance gains.

Our results confirm that even a modest amount of high-quality parallel data is critical for advancing

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 1088–1102

MT performance for structurally divergent language pairs, particularly when involving a low-resource language.[1]

2 Related Work

Despite Kyrgyz being classified as a low-resource language, recent years have witnessed notable progress in the development of machine translation (MT) systems for this language (Alekseev and Turatali, 2024). The very first machine translation systems for Kyrgyz were rule-based, representing foundational efforts in the field. A key contribution in this area is the open-source finite-state morphological transducer for Kyrgyz developed by Washington et al. (2012). This transducer, developed within the Apertium platform, was a critical component for a prototype Turkish-Kyrgyz machine translation system and laid the groundwork for further language pairs, including an in-progress Kazakh-Kyrgyz system. Alkım and Çebi (2019) proposed a rule-based approach for multilingual translation among four Turkic languages, including Kyrgyz.

Another line of work focusing on Turkic languages is based on neural machine translation (NMT). Mirzakhalov et al. (2021a) trained bilingual models for 22 Turkic languages, including Kyrgyz. In addition to developing baseline systems, this study also introduced a large-scale parallel dataset containing translation pairs for these languages, thereby providing valuable resources for advancing research in Turkic language MT.

Later, the authors released a multi-way multilingual model neural MT model (MNMT) for these languages, showing that the multilingual model outperforms almost all bilingual baselines (Mirzakhalov et al., 2021b).

Fine-tuning multilingual models for low-resource languages has shown considerable promise (Maillard et al., 2023). Notably, Kyrgyz has been incorporated into several multilingual NMT models, such as mT5 (Xue et al., 2021), leading to improved translation quality. A particularly significant contribution to the field is the multilingual NLLB-200 model, trained on 200 languages including Kyrgyz (Team et al., 2022). The primary objective of this work was to provide extensive coverage of low-resource languages within a unified framework.

Another line of research aimed at improving machine translation quality focuses on enhancing tokenization methods. For example, the study by Tukeyev et al. (2020) proposes a tokenization approach based on the Complete Set of Endings (CSE), which reduces vocabulary size and, as demonstrated on Kazakh–English translation tasks, yields better generation quality compared to Byte Pair Encoding (BPE) segmentation. Similarly, Abduali et al. (2025) investigate the Kyrgyz–Kazakh language pair and develop a morphological tokenizers based on the relational segmentation model. The scientific contribution of this article is the creation of the morphological tokenizer for Kyrgyz and Kazakh, as well as fine-tuning experiments with this dataset of the neural model T5-small.

3 Kyrgyz-English Parallel Datasets

Another important line of work in improving machine translation involves the creation of multilingual parallel datasets. Below is a brief overview of open parallel datasets that include the Kyrgyz language.

As part of the TurkLang-7 project (Khusainov and Minsafina, 2021), a corpus of parallel sentences was compiled for Russian and seven Turkic languages. For Kyrgyz, the collection includes 426,190 parallel sentence pairs; however, the final version of the dataset has not been made publicly available.

The NLLB v1 dataset[2] is a large English-Kyrgyz parallel corpus containing 21,360,637 sentence pairs, produced during the development of the NLLB-200 model (Team et al., 2022), where Kyrgyz was included among the languages for which bilingual translation pairs were automatically collected. However, the quality of this dataset is limited, as demonstrated by our manual analysis of the public OPUS sample, which comprises 49 examples[3]. The fully annotated sample that forms the basis for this analysis, including examples of misalignment and various error types, is provided in Appendix A. Our analysis revealed that only 59.18% of the sentences labeled as Kyrgyz were actually written in Kyrgyz. Furthermore, of this reduced Kyrgyz subset, a mere 55.17% were deemed to be accurate and fluent translations. This indicates that only about 32.65% of the original sample represents a high-quality, usable translation pair.

Kyrgyz language was also incorporated into

[1] https://github.com/kyrgyz-nlp/oldi-dataset-experiments

[2] https://opus.nlpl.eu/NLLB/en&ky/v1/NLLB

[3] https://opus.nlpl.eu/sample/en&ky/NLLB&v1/sample

a set of 14 low-resource languages for which the GoURMET project team compiled parallel datasets (van der Kreeft et al., 2022). In total, 14,498 Kyrgyz-English parallel sentences and 23,017 Kyrgyz-Russian parallel sentences were assembled. These sentences were obtained through machine translation followed by editorial validation and control.

The Open Language Data Initiative (OLDI) is a collaborative project that enables language communities, researchers, and developers to contribute to foundational datasets essential for machine learning and natural language processing. At present, OLDI[4] maintains two key datasets: OLDI-Seed and FLORES+.

The OLDI-Seed dataset contains 6,193 English sentences paired with translations into approximately 40 low-resource languages. The English source sentences were drawn from a diverse range of Wikipedia articles covering fields such as biology, astronomy, the arts, history, mathematics, etc. We use the English corpus as a source dataset for Kyrgyz translations.

FLORES+ is an evaluation benchmark consisting of two subsets — a test set of 1012 sentences and a validation set of 997 sentences — each professionally translated into 200 languages by expert linguists.

The X-WMT benchmark (Mirzakhalov et al., 2021b) is a test set designed to evaluate machine translation quality for Turkic languages. It is based on the professionally translated English–Russian corpus from the WMT 2020 News Translation Task. The original news sentences were subsequently translated into eight Turkic languages. For the English–Kyrgyz language pair, the benchmark comprises 500 sentences.

4 Language Description

Kyrgyz (also known as Kirgiz or Kirghiz) is a Turkic language of the Kipchak and/or the South Siberian branch (ISO 639-2: kir, Glottocode: kirg1245). It is spoken primarily in Kyrgyzstan, where it holds official status due to it being the national language. However, it does spread further to Central Asian countries, specifically in Gorno-Badakhshan Autonomous Region of Tajikistan and it is also considered to be a minority language in the Kizilsu Kyrgyz Autonomous Prefecture in Xinjiang, China. And another regional variant of the Kyrgyz language, referred to as Pamiri Kyrgyz, is spoken in northeastern Afghanistan and northern Pakistan.

From a phonetic and phonological standpoint, the Turkic languages most closely resembling Kyrgyz are the southern dialects of Altay, although Kyrgyz also shares significant similarities with Kazakh (Washington et al., 2012). Additionally, the southern varieties of Kyrgyz exhibit distinctive features that align with Uzbek, which are not present in any other Kyrgyz dialects. Approximately, 5.5 million people primarily in Kyrgyzstan consider Kyrgyz their native language.

5 Dataset Translation

5.1 Translation Workflow

Our translation team consisted of six native Kyrgyz speakers, of which two were experienced English to Kyrgyz translators and two were students from the Kyrgyz-English Language Program at Kyrgyz-Turkish Manas University. We followed the OLDI contribution guidelines (Open Language Data Initiative, 2025). The translation workflow was structured into four sequential stages: (1) machine translation using Aitil[5], a Gemma3-based MT service fine-tuned by Ulut Soft LLC (2) manual correction by individual translators, (3) team terminology unification, and (4) consistency review.

At the initial stage, we used the Aitil translation service, which was kindly provided by Ulan Bayaliev and Ulut Soft LLC. At the second stage, translators worked individually to post-edit their assigned batches of machine-translated sentences. This human review process covered all 6,193 sentences, with 99.16% of them being modified. The post-edits addressed a range of common machine translation issues; for instance, our work frequently involved correcting literal translations, improving lexical and terminological choices, and resolving stylistic inconsistencies. To illustrate the nature of these corrections, Table 1 presents several examples.

Following the individual post-editing, we implemented a targeted consistency review process. Our main priority was to ensure high-quality terminological consistency across the entire dataset. To achieve this, translators flagged domain-specific or ambiguous terms for group discussion. Once the team reached a consensus on the translation, the decision was implemented systematically: the

[4]https://oldi.org/

[5]https://translate.mamtil.gov.kg/

Table 1: Examples of machine translations and human post-edits

ID	Sentence
5715	**English:** ...if anyone does something that truly is bad, *it must be unwillingly* or out of ignorance; consequently, all virtue is knowledge. **MT by Aitil:** ...эгер кимдир бирөө чындап эле жаман нерсе жасаса, анда ал муну билбестиктен же *аргасыздан жасашы керек*; демек, бардык жакшылык – бул билим. **Human Post-Edit:** ...эгер кимдир бирөө чындап эле жаман нерсе жасаса, анда ал муну *аргасыздан же билбестиктен улам жасайт*; демек, ар кандай жакшылык – бул билим. **Comment: Syntactic/Stylistic Error:** The MT's output "аргасыздан жасашы керек" is a literal and awkward translation of the English modal structure "it must be". The word order was also unnatural and was corrected in the human edit.
5721	**English:** Although rule by a wise man would be preferable to *rule by law*, the wise cannot help but be judged by the *unwise*... **MT by Aitil:** Акылман адамдын башкаруусу *мыйзамдын башкаруусунан* артык болсо да, акылмандарды *акылсыздар* соттойт... **Human Post-Edit:** Акылмандын башкаруусу *мыйзам үстөмдүгүнөн* артык болсо да, акылмандар сөзсүз *наадандардын* сынына кабылгандыктан... **Comment: Lexical/Terminological Error:** "мыйзам үстөмдүгү" is the correct term for "rule by law". The MT's choice of "акылсыздар" (dumb) for "unwise" was inaccurate; "наадандар" (ignorant/unwise) is more appropriate.
5733	**English:** While the objective of the Pyrrhonists was the attainment of ataraxia, after Arcesilaus the Academic skeptics did not hold up ataraxia as the *central objective*. **MT by Aitil:** Пирончулардын максаты атараксияга жетүү болсо да, Арцелайдан кийин Академиялык скептиктер атараксияны *борбордук максат* катары карманган эмес. **Human Post-Edit:** Пиррончулардын максаты атараксияга жетүү болсо, Аркесилайдан кийин академиялык скептиктер атараксияны *негизги максат* катары көрсөтүшкөн эмес. **Comment: Literal Translation:** "борбордук максат" is a literal translation of "central objective". The more natural term is "негизги максат" (main/primary objective). An extra word ("да") was also removed.

translator who had flagged the term would then perform a search across the entire dataset to update all its occurrences with the agreed-upon translation, ensuring uniformity.

This approach ensured consistency for key concepts. However, we acknowledge two limitations in our overall methodology. First, a comprehensive, sentence-by-sentence peer review was not conducted due to time constraints. Second, the translation workload was unevenly distributed, with one translator contributing the majority of the post-edits:

- Murat: 4910 sentences (79.28%)
- Gulaiym: 614 sentences (9.91%)
- Meerim: 284 sentences (4.58%)
- Elza: 242 sentences (3.91%)
- Akylai: 94 sentences (1.52%)
- Begayim: 49 sentences (0.79%)

5.2 Addressing Problematic Translations

During translation process we turned to dictionaries, both general and technical ones, specific for the given areas, such as, Yudakhin's Russian to Kyrgyz, Kyrgyz to Russian dictionaries (Yudakhin, 1957), (Yudakhin, 1965), Abdiev's English to Kyrgyz, Kyrgyz to English dictionary (Abdiev and Sydykova, 2015) and other dictionaries that are available online at el-sozduk.kg[6]. We also used a collection of Kazakh to Russian and Russian to Kazakh dictionary available at sozdik.kz to verify that international terms (for example, "antibody" which was searched for in Russian: "антитело").

Kyrgyz has certain features that make translation between Kyrgyz and English a challenging task. One of them is the word order. Unlike English, which uses SVO order, Kyrgyz is a SOV type lan-

[6] https://el-sozduk.kg

Table 2: Example of sentence fragmentation

ID	Sentence
4140	**English:** Perhaps the foremost mathematician of the 19th century was the German mathematician Carl Friedrich Gauss, who made numerous contributions to fields such as algebra, analysis, differential geometry, matrix theory, number theory, and statistics.
	Kyrgyz (less effective): Кыязы, 19-кылымдын эң залкар математиги немис математиги Карл Фридрих Гаусс болгон, жана ал алгебра, анализ, дифференциалдык геометрия, матрицалар теориясы, сандар теориясы жана статистика сыяктуу тармактарга көптөгөн салым кошкон.
	Kyrgyz (final): 19-кылымдын эң залкар математиги, кыязы, немис окумуштуусу Карл Фридрих Гаусс болгон. Ал алгебра, анализ, дифференциалдык геометрия, матрицалар теориясы, сандар теориясы жана статистика сыяктуу тармактарга зор салым кошкон.

guage. Although the placement of the verb does not pose any issues, it becomes more difficult the more complex the sentence is. Syntactic functions (e.g., adverbial phrases) also have to be displaced, which makes translation more complicated, as it becomes harder to keep the sentence eligible for the readers without losing the natural flow of the language. (Sankaravelayuthan, 2019). Our translators have also encountered this particular problem, that is dealt with sentence fragmentation. Another feature is that Kyrgyz is agglutinative. It forms words through the sequential addition of morphemes while preserving their original spelling and pronunciation. This linguistic structure allows for the creation of an extensive range of word forms, while English, being an analytical language, relies heavily on such features as auxiliary words, modal verbs and dependent clauses rather than inflection to express grammatical relationships (Kara, 2003).

While the challenges mentioned before were expected, our translators have encountered another two major problems: long, compound-complex sentences and terminological gaps and neologisms. Those problems significantly hindered the translation process. The first problem was dealt with sentence fragmentation, a syntactic transformation technique (Shermatova, 2010). This technique involves restructuring a single complex sentence from the source language into two or more sentences in the target language. The goal is to maintain a natural linguistic flow, and to preserve the original's clarity, emphasis, and stylistic integrity. An example of such fragmentation is listed in Table 2.

The less effective version contains two coordinated clauses with the conjunction "жана" ("and"), presents two critical problems: stylistical redundancy (the repetition of the word "математиги") and syntactic overload (while technically a single grammatical unit, the sentence is long and burdensome, combining two distinct ideas with a simple conjunction, which weakens its clarity and impact). This final resolves both issues. The redundant term is eliminated by using a more appropriate synonym ("окумуштуусу" - "scientist") in the first sentence and allowing the subject to be implied in the second. Most importantly, the structure is clear, emphatic, and stylistically natural in Kyrgyz. The main challenges were translational gaps and neologisms, especially in scientific and technical fields such as genetic engineering and astronomy. Lacking equivalents, we often used transliteration and calquing, typically via Russian—a legacy of the Soviet era when Russian dominated science and education. As many specialists are bilingual, adopting Russian-based forms speeds comprehension (Table 3).

The terms "knockout" and "extinction" are highly specific neologisms. With "knockout" creating a descriptive Kyrgyz phrase (e.g., "генди өчүрүү"–"gene deactivation") would be less precise and more cumbersome. Coining an entirely new Kyrgyz term would likely be unintelligible to specialists who are already familiar with the concept through international, often Russian-language, literature. Therefore, the most effective strategy is the transliteration of the international term (or calquing of Russian term) as "нокаут". This direct adoption is efficient and aligns with the established scientific vocabulary (Zaid et al., 2008).

For accuracy and consistency, we adopted the international term "экстинкция", sourced from an English–Russian astronomical dictionary (Murtazov, 2010), as it aligns with existing literature and is more recognizable to Kyrgyz readers. The problems with scientific terms would not have posed

Table 3: Examples of terminology mapping

ID	Sentence
3986	**English:** In a simple **knockout** a copy of the desired gene has been altered to make it non-functional **Kyrgyz:** Жөнөкөй **нокаутта** керек болгон гендин көчүрмөсү иштебей калышы үчүн өзгөртүлөт.
3506	**English:** In astronomy, **extinction** is the absorption and scattering of electromagnetic radiation by dust and gas between an emitting astronomical object and the observer **Kyrgyz:** Астрономияда **экстинкция** – бул нур чыгаруучу астрономиялык объект менен байкоочунун ортосундагы чаң жана газ тарабынан электромагниттик нурлануунун жутулушу жана чачырашы.

such substantial problems if not for the major linguistic flaw that calquing had created. In most of the cases calquing was justified, as it provided a more specific and in-depth understanding of the phenomenon, in some it resulted in a borrowed word replacing a perfectly suitable native synonym. The primary example is showcased in Table 4.

The case with "bodies" highlights this protrusion. The official dictionary (Eshbaev and Eshbaeva, 2023) suggests using the Russian calque "тело," whereas we observe that this term is now often replaced by the Kyrgyz word "дене." This modern usage is evident in real-world examples from publications such as BBC News Kyrgyz[7], which uses "антидене" for "antibody," and Nazar News[8], which uses "космостук дене" for "cosmic body." During translation, our team gave priority to the latter option, as it better reflects the language usage.

6 Translation Experiment

6.1 Model Training

To demonstrate the quality of the collected dataset, we fine-tuned several seq2seq machine translation models on the gathered 6,193 parallel sentences and evaluated their performance on the FLORES+ benchmark (Team et al., 2022; The Open Language Data Initiative, 2024). We also evaluate our model on the Turkic X-WMT benchmark to compare its performance with the bilingual and multilingual baselines proposed for Kyrgyz–English machine translation.

For fine-tuning, we utilized the Transformers library (Wolf et al., 2020). Across all experiments, we employed the AdamW optimizer with a learning rate of 0.0001 and trained for 6 epochs. The training was conducted on a single T4 GPU with 16 GB of memory.

For each model, we take the same base instance and fine-tune it twice independently — once for translating from English to Kyrgyz (en → ky) and once for translating from Kyrgyz to English (ky → en). The models selected for fine-tuning included:

mT5-base is a multilingual transformer model based on the T5 architecture. It is pretrained on a denoising objective across 101 languages, including Kyrgyz. While mT5 supports Kyrgyz at the pretraining stage, it requires fine-tuning on translation tasks to perform effective machine translation into and from Kyrgyz. (Xue et al., 2021)

mBART-large is a seq2seq transformer extending the BART model to multilingual settings. However, the base mBART-large model does not include Kyrgyz in its pretrained vocabulary, limiting its direct applicability to Kyrgyz translation without additional fine-tuning and vocabulary extension. (Tang et al., 2020)

M2M100 is a multilingual seq2seq model developed by Facebook supporting direct translation between 100 languages without relying on English as a pivot. Notably, Kyrgyz is not included in the 100 languages covered by M2M100, thus the model cannot translate Kyrgyz "out of the box." (Tang et al., 2020)

NLLB (No Language Left Behind) is a large-scale multilingual seq2seq model developed by Meta, designed to improve translation quality, particularly for low-resource languages. It supports over 200 languages, including Kyrgyz. NLLB can translate to and from Kyrgyz with high quality without requiring additional fine-tuning, making it well-suited for applications involving Kyrgyz language translation. (Team et al., 2022)

[7] https://www.bbc.com/kyrgyz/articles/c0lye5ngjwxo

[8] https://nazarnews.org/posts/zherdin-kagyilyishuusu-eki-asman-tiregen-asteroid-zherge-zhakyindadyi

Table 4: An example of an unnecessary calque

ID	Sentence
3550	**English:** Once it became clear that Earth was merely one planet amongst countless **bodies** in the universe, the theory of extraterrestrial life started to become a topic in the scientific community. **Kyrgyz:** Жер – ааламдагы сансыз асман **денелеринин** арасындагы катардагы эле бир планета экени айкын болгондон кийин, жерден тышкаркы жашоо теориясы илимий коомчулукта талкуулана баштаган.

6.2 Vocabulary expansion

For the models that were not pretrained on Kyrgyz (mBART and M2M100), we expanded the token vocabulary by training a SentencePiece model (Kudo and Richardson, 2018) on the Kloop corpus, a dataset of Kyrgyz news articles (kyrgyz-nlp, 2024). This allowed us to extend the vocabulary by 14,466 byte-pair encoding (BPE) tokens (Sennrich et al., 2016).

6.3 Metrics

To evaluate generation quality, we employ two variations of the ChrF metric: ChrF (Popović, 2015) and ChrF++ (Popović, 2017), both of which are well suited for morphologically rich languages. While prior work has shown that ChrF++ correlates more strongly with human judgments, we also report ChrF scores to enable direct comparison with the Bilingual and MNMT baselines (Mirzakhalov et al., 2021b). For this purpose, we use the implementation provided in the SacreBLEU toolkit[9], adopting the same parameter configuration as in the MWMT study for the computation of the ChrF metric to ensure consistency and comparability of results.

6.4 Experiment Results

The tables present the training results of our models, evaluated on two benchmarks: FLORES+ and X-WMT.

We compare the outputs of our fine-tuned models with its pretrained versions (**pretrained baselines**) on the FLORES+ benchmark to evaluate the impact of our dataset on translation quality. The results are reported separately for the two translation directions: English→Kyrgyz (Table 5) and Kyrgyz→English (Table 6). The results of the X-WMT benchmark are summarized in Table 7.

Fine-tuning on our dataset substantially improves generation quality for all models in both translation directions. (Table 5 and Table 6)

NLLB-200 achieved the highest performance in all the evaluated models. Although the models were already pre-trained on a large Kyrgyz corpus, the addition of even a relatively small dataset like ours can yield a noticeable improvement in translation performance.

When comparing translation directions, Kyrgyz–English translations (Table 6) outperforms English–Kyrgyz translations (Table 5) for all models except M2M100. This asymmetry reflects the stronger representation of English in the pretraining corpora, which enables more reliable decoding into English. The exception of M2M100 can be attributed to its non-English-centric pretraining design.

Manual validation of system outputs, however, revealed that automatic evaluation scores tend to overestimate real-world usability. Despite relatively high ChrF and ChrF++ scores, the translations still contained a considerable number of critical errors. In particular, M2M100 outputs suffered from frequent tokenization and word-concatenation issues, likely caused by suboptimal adaptation of the tokenizer. More generally, all systems struggled with morphological accuracy, especially the generation of correct suffixes. This challenge stems from the agglutinative nature of Kyrgyz, where the high variability of word endings makes exact surface realization difficult.

The performance of pretrained baselines largely depends on whether Kyrgyz was included in their pretraining data and the model's intended use. For example, although mT5's pretraining corpus contains Kyrgyz, it cannot perform machine translation "out of the box" and requires fine-tuning. Neither mBART nor M2M100 was pretrained on Kyrgyz; nevertheless, M2M100 achieves comparatively stronger baseline results, as we extended the model to Kyrgyz by introducing a new language token and initializing its embeddings from Kazakh, a closely related language. Finally, NLLB-200, which was trained on a large Kyrgyz corpus, achieves high

[9] https://github.com/mjpost/sacrebleu

Table 5: Performance comparison on the FLORES+ English-Kyrgyz translation benchmark.

	mT5	mBART	M2M100	NLLB-200
fine-tuning	ChrF: 30.57 ChrF++: 26.53	ChrF: 34.85 ChrF++: 30.82	ChrF: 40.77 ChrF++: 33.91	ChrF: 49.37 ChrF++: 44.62
pretrained baselines	ChrF: 0.20 ChrF++: 0.63	ChrF: 0.73 ChrF++: 1.76	ChrF: 10.88 ChrF++: 9.10	ChrF: 44.56 ChrF++: 40.21

Table 6: Performance comparison on the FLORES+ Kyrgyz–English translation benchmark.

	mT5	mBART	M2M100	NLLB-200
fine-tuning	ChrF: 36.82 ChrF++: 34.22	ChrF: 36.23 ChrF++: 33.88	ChrF: 38.77 ChrF++: 36.19	ChrF: 51.77 ChrF++: 49.34
pretrained baselines	ChrF: 1.84 ChrF++: 1.81	ChrF: 0.84 ChrF++: 1.22	ChrF: 13.29 ChrF++: 10.58	ChrF: 49.85 ChrF++: 47.48

Table 7: Performance comparison on the X-WMT dataset (ChrF). Values that exceed the performance of the Multilingual MNMT baseline are highlighted in bold. For the ChrF score, we did not reproduce the baselines ourselves, but used the results reported in the paper.

	mT5	mBART	M2M100	NLLB-200	Bilingual XWMT	Multilingual MNMT
en-ky	28.32	30.92	**36.41**	**47.61**	27	34
ky-en	34.69	33.68	36.10	**47.84**	29	39

performance without additional fine-tuning.

All of our trained models outperform the bilingual MWMT baselines (Table 7). Among them, the multilingual MNMT model is surpassed only by NLLB-200 in both translation directions, and by M2M100 in the English→Kyrgyz direction, demonstrating the competitive performance of our models across different translation setups.

7 Conclusion

This study, part of the OLDI initiative, contributed 6,193 English-Kyrgyz sentence pairs and showed that even modest, carefully curated data can improve neural machine translation for low-resource, morphologically complex languages. In this work, we addressed three key challenges: adapting to agglutinative morphology, restructuring complex subordination, and maintaining terminological consistency. Our preference for native Kyrgyz terms over Russian calques reflects contemporary usage and contributes to natural language evolution.

Fine-tuning mBART, M2M100, mT5, and NLLB-200 models on our dataset yielded consistent performance gains, validating the value of linguistically-informed data. However, the manual evaluation of model translations, highlight the need for further research to achieve production-quality NMT for Kyrgyz. Our collaborative workflow—combining MT, manual correction, and consistency review—offers a replicable methodology for other low-resource languages. Future work involves expanding the dataset and integrating community feedback. This research provides concrete resources and a methodological framework for community-driven language technology development, balancing technical advancement with cultural-linguistic authenticity.

8 Acknowledgments

We would like to thank Elza Mambetakunova and Begaiym Mamatova for helping us with the translations. We are also grateful to Ulan Bayaliev and Ulut Soft LLC for providing access to their Aitil translator, which served as a valuable resource for this work.

References

Taalaibek Abdiev and Lira Sydykova. 2015. *Anglische-Kyrgyzcha, Kyrgyzcha-anglische syozduk [English-Kyrgyz, Kyrgyz-English Dictionary]*. Avrasya Press, Bishkek.

Balzhan Abduali, Ualsher Tukeyev, Zhandos Zhumanov, and Nella Israilova. 2025. Study of Kyrgyz-Kazakh Neural Machine Translation. In *Proceedings of the 17th Asian Conference on Intelligent Information and Database Systems (ACIIDS 2025)*, pages 272–283, Kitakyushu, Japan. Springer.

Anton Alekseev and Timur Turatali. 2024. KyrgyzNLP: Challenges, progress, and future. *arXiv preprint*, arXiv:2411.05503.

Emel Alkım and Yalçın Çebi. 2019. Machine translation infrastructure for Turkic languages (MT-Turk). *The International Arab Journal of Information Technology*, 16(3).

A. A. Eshbaev and Ch. A. Eshbaeva. 2023. *Russian-Kyrgyz Explanatory Dictionary of Frequently Occurring Terms in Medicine*. Kyrgyz Encyclopedia and Terminology Center, Bishkek.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

D. S. Kara. 2003. *Kyrgyz*. Languages of the World/Materials. Lincom Europa, Munich.

Aidar Khusainov and Alina Minsafina. 2021. First results of the "TurkLang-7" project: Creating Russian-Turkic parallel corpora and MT systems. In *Proceedings of the International Workshop on Computational Models in Language and Speech (CMLS 2021)*. CEUR-WS.org. Held as part of the 10th International Conference on Analysis of Images, Social Networks and Texts (AIST 2021).

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

kyrgyz-nlp. 2024. Kloop corpus. `https://github.com/kyrgyz-nlp/kloop-corpus`.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Behzodbek Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. 2021a. A large-scale study of machine translation in the Turkic languages. *arXiv preprint*, arXiv:2109.04593.

Jamshidbek Mirzakhalov, Anoop Babu, Aigiz Kunafin, Ahsan Wahab, Bekhzodbek Moydinboyev, Sardana Ivanova, Mokhiyakhon Uzokova, Shaxnoza Pulatova, Duygu Ataman, Julia Kreutzer, Francis Tyers, Orhan Firat, John Licato, and Sriram Chellappan. 2021b. Evaluating multiway multilingual NMT in the Turkic languages. In *Proceedings of the Sixth Conference on Machine Translation*, pages 518–530, Online. Association for Computational Linguistics.

A. K. Murtazov. 2010. *Russian-English Astronomical Dictionary*. Ryazan State Pedagogical University, Ryazan.

Open Language Data Initiative. 2025. Contribution guidelines. Accessed: 2025-08-13.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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.

R. Sankaravelayuthan. 2019. Word order in translation. *Language in India*, 19(4):196–206.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Feruza S. Shermatova. 2010. *Features of the translation of syntactic stylistic devices from English to Kyrgyz*. Ph.D. thesis, Kyrgyz-Russian Slavic University, Bishkek.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

NLLB Team, 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, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

The Open Language Data Initiative. 2024. Flores+ dataset. https://huggingface.co/datasets/openlanguagedata/flores_plus. Accessed: 2025-09-04.

Ualsher Tukeyev, Aidana Karibayeva, and Zh. Zhumanov. 2020. Morphological segmentation method for Turkic language neural machine translation. *Cogent Engineering*, 7(1):1856500.

Peggy van der Kreeft, Sevi Sariisik, Wilker Aziz, Alexandra Birch, and Felipe Sánchez-Martínez. 2022. GoURMET – machine translation for low-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 225–226, Ghent, Belgium. European Association for Machine Translation.

Jonathan North Washington, Mirlan Ipasov, and Francis M. Tyers. 2012. A finite-state morphological transducer for Kyrgyz. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2244–2248, Istanbul, Turkey. European Language Resources Association (ELRA).

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

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

K. K. Yudakhin, editor. 1957. *Russko-kirgizskij slovar' [Russian-Kyrgyz Dictionary]*. State Publishing House of Foreign and National Dictionaries, Moscow.

K. K. Yudakhin. 1965. *Kirgizsko-russkij slovar' [Kyrgyz-Russian Dictionary]*. Sovetskaja Enciklopedija, Moscow.

A. Zaid, H.G. Hughes, E. Porcheddu, and F. Nicholas. 2008. *Glossary of biotechnology for food and agriculture (Russian Edition)*, volume 9 of *FAO Research and Technical Paper*. Food and Agriculture Organization of the United Nations (FAO), Rome. Translated into Russian by G. Kamarova, T. Gavrilenko, I. Anisimova, O. Antonova, O. Kuznetsova, and S. Kharitonov.

A Qualitative Analysis of the Official OPUS Sample for NLLB v1 (En-Ky)

Table 8: Sentence pairs drawn from the public sample of the NLLB v1 En-Ky corpus provided by the OPUS project. Each pair is manually evaluated for language identification (`Is Kyrgyz`) and translation accuracy (`Is Correct`). The `Comments` column details the types of errors found.

En	Ky	Is Kyrgyz	Is Correct	Comments
And if you call them to guidance, they do not listen.	Эгер аларды түз жолго чакырсаң угушпайт.	Yes	Yes	
(It is only) a provision in this world, then to Us is their return.	Бул дүйнөдө жыргал, андан соң алардын Бизге кайтуулары бар.	Yes	Yes	
Are you [more] satisfied with the life of this world, rather than the Hereafter?	Дүнё уларга, Охират бизга эканига рози эмасмисан?!	No	No	
If GOD wills, He can take away their hearing and their eyesight.	Эгер Алла каалаганда, алардын угуусун да, көрүүсүн да алып коймок эле.	Yes	Yes	
and you will be among those brought near."	Ва, албатта, менинг яқин кишиларимдан бўлурсизлар," деди."	No	No	
and We see it to be near.	Биз эса уни яқин деб билурмиз.	No	No	
But his people's only answer was, "Bring down upon us God's chastisement, if thou art a man of truth!"	Ошондо анын коомунун жообу: "Эгер чынчылдардан болсоң, Алланын азабын бизге келтирчи!" - дегенинен башка болбоду.	Yes	Yes	
That we might follow the magicians if they are the victorious?"[1]	Агар сеҳргарлар ғолиб бўлиб чиқсалар, эҳтимол бизлар ўшаларга эргашурмиз," дейилди.	No	No	
For the wrong-doers there will be no helpers.	золим кимсалар учун бирон ёрдамчи бўлмас!."[20]	No	No	
(And your abode is the Fire, and there is none to help you.)	Жойингиз жаҳаннамдир ва сизларга ҳеч ёрдамчилар йўқ.	No	No	
But worship Allah alone and be among the grateful" [Qur'an, 39:66].	Аллахка гана сыйын жана шүгүр кылуучулардан бол!"	Yes	Yes	

Table 8: (Continued)

En	Ky	Is Kyrgyz	Is Correct	Comments
Say, 'They are appointed periods of time for (general convenience of) people and for determining the time of Pilgrimage.	Айт: "Булар адамдарга убакытты жана ажылыкты белгилөө куралы."	Yes	Yes	
And remember the name of your Lord, morning and evening.	Жана эртели-кеч Роббиңдин атын зикир кыл!	Yes	Yes	
Indeed, there has come to you a bearer of glad tidings and a warner.	Ошентип, албетте, силерге куш кабар берүүчү жана коркутуучу келди.	Yes	No	Ошентип is extra
Call upon your helpers, other than Allah, to assist you, if you are true.	Жана эгер чынчыл болсоңор, Аллахтан башка күбөлөрүңөрдү (ишенген жардамчыларыңарды) чакыргыла.	Yes	No	has a comment in parentheses
Protect yourself and your family from the fire of hell."	Өзүңөрдү жана үй-бүлөңөрдү Тозоктун отунан коргогула!" - деп буюрулган.	Yes	No	- деп буюрулган. is extra
And that My punishment is a painful retribution.	менинг азобимдан азоблангани...	No	No	
- that Day, man will remember, but how [i.e.,	Адам ал күнү ойлонуп-эстейт, бирок (бул) эстөөдөн ага эмне пайда?	Yes	No	the English part is incomplete
Do they expect anything but the likes of the days of those who passed away before them?	Алар (капырлар) Кыямат Күнүнүн капилет келишин күтүп жатышабы?Анын белгилери келди.	Yes	No	has a comment in parentheses; Анын белгилери келди. is extra
Remember the name of your Lord morning and evening.	Жана эртели-кеч Роббиңдин атын зикир кыл!	Yes	Yes	
2:55 - Who were the sons of the servants of Solomon?	Эзра 2: 55 - Сулаймандын кулдарынын уулдары деген кимдер эле ?	Yes	Yes	
Why did you kill them, if you are telling the truth?"	Силер чынчыл болсоңор анда эмне үчүн аларды өлтүрдүңөр?!"	Yes	Yes	

Table 8: (Continued)

En	Ky	Is Kyrgyz	Is Correct	Comments
Indeed, Gehenna is your recompense, and the reward of those who follow you, an ample recompense.	Бас улардан ким сенга эргашса, у ҳолда, шак-шубҳасиз, жаҳаннам сизларга етарли жазо бўлур!	No	No	
WE have created them of that which they know.	Албетте, Биз аларды өздөрү билген нерседен жаратканбыз.	Yes	Yes	
in which they will remain forever, and will not find any guardian or helper.	Алар ал жерде (тозокто) дос жана жардамчы таппай, түбөлүккө калышат	Yes	No	has a comment in parentheses
Mankind were not but one nation (community), but they differed (later).	(Аввалда) одамлар фақат бир миллат (яъни, бир динда) бўлган эдилар.	No	No	
The Fire will be your refuge, and you will have no helpers.	Силердин жайыңар - от болот жана силерге эч кандай жардамчылар болбойт.”	Yes	Yes	
(The Day that Allah will not disgrace the Prophet (Muhammad) and those who believe with him.	Бир кундаки, унда Аллоҳ Набийни ва у билан бирга бўлган иймон келтирганларни шарманда қилмас.	No	No	
Thus God makes clear His Revelations to you, that you may be thankful.” (5:89).	Шүкүр кылууңар үчүн Алла силерге Өз аяттарын мына ушундай ачык-айкын баян кылат.	Yes	Yes	
And that My chastisement is the painful chastisement.	менинг азобимдан азоблангани...	No	No	
And what will make you know what Al-Qari’ah is?	(8) ”Сижжийн” қандай нарса эканини сенга нима билдирди?!	No	No	
(and it will be said to them,).This is the Fire you used to deny	52: 15 ”Силер жалганга чыгарган от - мына ушул.	Yes	No	the Kyrgyz part is incomplete
But surely now has come to you a bearer of glad tidings and a Warner.	Ошентип, албетте, силерге куш кабар берүүчү жана коркутуучу келди.	Yes	Yes	

Table 8: (Continued)

En	Ky	Is Kyrgyz	Is Correct	Comments
call your witnesses, besides Allah, if you are telling the truth.	Жана эгер чынчыл болсоңор, Аллахтан башка күбөлөрүңөрдү (ишенген жардамчыларыңарды) чакыргыла.	Yes	No	has a comment in parentheses
When a warner came to them, however, it only increased their aversion.	Бирок аларга бир эскертүүчү-коркутуучу келгенде, (бул алардын) жек көрүүлөрүнөн башкасын көбөйткөн жок.	Yes	No	has a comment in parentheses; mistranslation due to incorrect pluralization of the abstract noun жек көрүү (hatred).
But we see it (quite) near.”	Биз эса уни яқин деб билурмиз.	No	No	
Cretans and Arabs, we hear them in our own tongues speaking of the mighty deeds of God.”	криттиктер жана аравиялыктар, өз тилибизде алардын Кудайдын улуу иштери жөнүндө айтып жаткандарын угуп жатабыз.”	Yes	Yes	
And mention the name of your lord (in prayer) morning and evening (Al-Quran 48:9)	Жана эртели-кеч Роббиңдин атын зикир кыл!	Yes	Yes	
Rather, every one of them wishes that he should be given unrolled (divine) scriptures	Балки, улардан ҳар бир киши ўзига очилган саҳифалар берилишини ирода қилар.	No	No	
did you kill them, if you are true in your claim?”	Агар ростгуй булсаларингиз, нимага уларни улдирдингиз?” деб айт.	No	No	
Say: ”You can only expect for us one of two good things; while we wait for you to be afflicted by God with a retribution from Him, or by our hands.	Биз силерге Алладан жаза келерин же биз аркылуу силерге бир азап жиберерин күтүп жатабыз.Анда, силер күткүлө, силер менен чогуу биз да күтөбүз” - деп айт.	Yes	No	completely misaligned fragment
We are only waiting for the order of Allah the Almighty.	Биз анын акысын бир гана Аллаху тааладан күтөбүз,” - деди.	Yes	No	completely misaligned fragment

Table 8: (Continued)

En	Ky	Is Kyrgyz	Is Correct	Comments
Your abode is the fire, and you shall not have any helpers.”	Жойингиз жаҳаннамдир ва сизларга ҳеч ёрдамчилар йўқ.	No	No	
But Iblis (did it not), he refused to be with those who made obeisance.	Ичине Жин кирген адамдарга окшоп калышыптырго.	Yes	No	completely misaligned fragment
hath prepared for them a goodly recompense.	учун улуғ мукофот (яъни, жаннат) тайёрлаб қўйгандир.	No	No	
Then bring your book, if you are truthful.” (37:149-157).	Uzbek / Ozbekcha / Özbekçe Агар ростгўйлардан бўлсангиз, китобингизни келтиринг!	No	No	
When this was clearly shown to him he said: ”I know now that God is able to do all things.”	Ал ага (булар) ап-ачык белгилүү болгондон кийин (мындай) деди: ”Аллахтын бүт нерсеге күч жеткирүүчү экенин билем.”	Yes	No	has a comment in parentheses; awkward literal translation: The phrase ”able to do all things” (implying omnipotence) is rendered as күч жеткирүүчү, which literally means ”the one who delivers/supplies power.” A more idiomatic and accurate translation would be Аллахтын бүт нерсеге күчү жетет экенин билем.
For the wrong-doers there will be no helpers.	Энди золим кимсалар учун бирон ёрдамчи бўлмас!	No	No	
Those people, it has not been for them to enter them except fearing.	бу кимсалар учун ундай жойларга фақат қўрққан ҳолларида кириш жоиз эди-ку.	No	No	