Low-resource Buryat-Russian neural machine translation

Dari Baturova[1], Sarana Abidueva[2], Ivan Bondarenko[1], Dmitrii Lichko[3]

[1]Novosibirsk State University, Russia
[2]Saint Petersburg State University, Russia
[3]NUST MISiS, Russia

baturova.dari@gmail.com, abidueva.sarana02@gmail.com
i.bondarenko@g.nsu.ru, lichko2002@mail.ru

Abstract

This paper presents a study on the development of a neural machine translation (NMT) system for the Russian-Buryat language pair, focusing on addressing the challenges of low-resource translation. We also present a parallel corpus, constructed by processing existing texts and organizing the translation process, supplemented by data augmentation techniques to enhance model training.

We managed to achieve BLEU score of 20 and 35 for translation to Buryat and Russian respectively. Native speakers have evaluated the translations as acceptable.

Future directions include expanding and cleaning the dataset, improving model training techniques, and exploring dialectal variations within the Buryat language.

1 Introduction

The Buryat language is the national language of the Buryat people and is spoken in Russia, Mongolia, and China. It belongs to the Mongolic language group. However, due to its geographic distribution, the Buryat language has evolved differently in each country, influenced by the dominant languages and cultural contexts of the respective regions. In this article, we focus specifically on the variety of Buryat spoken in Russia, which is identified by the ISO code bxr.

Although Buryat is an official language of the Republic of Buryatia in the Russian Federation, the overwhelming majority of intellectual activity there is carried out in Russian and the number of young people speaking Buryat is rapidly declining. UNESCO included the Buryat language in the "Atlas of the world's languages in danger" (UNESCO, 2010).

As part and consequence of this problem, Buryat is underrepresented in computational linguistics, has limited available corpora and linguistic resources. That creates the main challenge of conducting a machine translation system for Buryat.

The Buryat language has undergone several transitions in its writing system throughout its history. Since 1939, it has been written using the Cyrillic alphabet. Before then, from 1930 to 1939, it utilized a Latin-based alphabet. Going further back, since the 18th century, the traditional Mongolian script served as the writing system for Buryat. This adds another complexity to the construction of machine translation, as some of the available literature was in Latin.

Furthermore, the preservation and revitalization of endangered languages through modern technologies have become critical goals in both linguistic research and cultural heritage preservation. In this context, creating robust machine translation systems not only aids communication but also contributes to the documentation and promotion of underrepresented languages.

2 Related Work

Low-resource machine translation has been an active area of research in recent years, driven by the need to support underrepresented languages.

Several large-scale initiatives have extended machine translation capabilities to hundreds of languages, including low-resource ones. Notable examples include the work of Bapna et al. (2022), the No Language Left Behind (NLLB) project (NLLB Team et al., 2022), and the efforts described in Fan et al. (2021). These projects demonstrate the potential of multilingual models to address challenges in low-resource settings.

Other studies focus on developing machine translation systems for individual low-resource languages by fine-tuning multilingual models. Examples include work on Erzya (Dale, 2022), Ngambay (Sakayo et al., 2023), Zarma (Keita et al., 2024), Karachay-Balkar (Berberov et al., 2024), and Aro-

Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters), pages 85–93

manian (Jerpelea et al., 2025). These efforts often rely on community-driven datasets and highlight the importance of adapting models to specific linguistic and cultural contexts.

In June 2024, Google Translate[1] added support for the Buryat language. According to company representatives, this was made possible by leveraging their language model PaLM 2. However, the dataset used for training has not been released as open source.

Research on Buryat in the context of natural language processing includes work by Konovalov and Tumunbayarova (2018), who explored word vector representations for Buryat using models such as pointwise mutual information (Church and Hanks, 1990), GloVe (Pennington et al., 2014), and Word2Vec (Mikolov et al., 2013), trained on data from the Buryat Wikipedia[2]. More recently, Shliazhko et al. (2024) introduced a multilingual variant of the GPT-3 large language model, trained on 61 languages, including Buryat, using corpora such as Wikipedia and the C4 dataset (Raffel et al., 2020).

3 Parallel Corpus Construction

Our parallel corpus was constructed through three main approaches: manual translation of Russian texts by hired translators, collaboration with local organizations in the Republic of Buryatia that have bilingual textual resources and web-based data collection.

The dataset is available as an open source[3].

3.1 Manual translation of texts

In order to create a quality Russian–Buryat parallel corpus from zero, we crafted a semi-automated system for the selection and preparation of the Russian source texts. The Taiga Corpus news segment[4] contains articles from various online media like Lenta.ru, Interfax, Komsomolskaya Pravda, N+1, Fontanka.ru and Arzamas. For our project, we utilized the text corpus that was news genre to ensure maximal variety, along with clearness of meaning, and a stable correspondence between the source and the target sentences.

To improve the ease of translation, we programmed fragments consisting of several sentences (up to five) into coherent passages instead of treating every sentence as isolated. Long paragraphs were split into smaller chunks, while ensuring they were semantically cohesive.

Text that did not suit requirements was discarded and the remaining content was processed using text embedding model `aiforever/sbert_large_mt_nlu_ru`[5] for the vectorized representations of the sentences. We then K-means clustered the data for initial selections to be more diverse and representative based on semantic similarity. The final Russian dataset contained 95,300 text fragments.

The fragments were sent to three professional translators who translated the text into Buryat. Currently, the corpus consists of 11,392 Russian–Buryat sentences that have been translated manually, with work still progress.

3.2 Collaborations with local organizations

In order to extend the corpus, we cooperated with some regional institutions that deal with Buryat language materials. These were:

1. The State Translation Service of the Republic of Buryatia, which contributed bilingual presidential decrees, government resolutions, and other legal acts of subordinate level decisions. They were at first in DOCX format and were converted by the means of some automation into a format of a parallel table – a step-by-step process.

2. The Buryat Research Center of the Siberian Branch of the Russian Academy of Sciences (BRC SB RAS) which enabled access to five parallel literary texts, but these texts had frequent mismatches at the level of sentence alignment due to translation losses or free rendering of the text. Reasonable estimates claim that only the texts which were structurally most homogeneous were chosen to be included in the corpus, other texts were set aside for possible later processing.

3.3 Web-based Data Collection

A significant portion of the Buryat-Russian parallel data was collected from the web.

[1] https://translate.google.com

[2] https://bxr.wikipedia.org/wiki/

[3] https://huggingface.co/datasets/buryat-translation/buryat_russian_parallel_corpus

[4] https://tatianashavrina.github.io/taiga_site/

[5] https://huggingface.co/ai-forever/sbert_large_mt_nlu_ru

ааараг 1. 1) редкий, необычный; **ааараг ушар** редкий случай; 2) *перен.* аморальный; негодный, никчёмный; **ааараг урагшагүй ябадалтай хүнүүд** никуда не годные люди (*о бездельниках, тунеядцах и т.д.*); **2.** редко; **алдуу хэхэнь ааараг хүн даа** человек-то он такой, что редко ошибается.

Figure 1: Example of the dictionary article

Several dictionaries are available for the Buryat language. For our purposes, we selected the Buryat-Russian dictionary by Shagdarov and Cheremisov (Shagdarov and Cheremisov, 2010) due to its extensive scope and detailed coverage, which surpasses that of other dictionaries. This dictionary contains 30,000 words, provides grammatical information and usage examples, characteristic of Buryat culture. We encountered several challenges during the data extraction process. First, the dictionary was available only as a PDF scan, which resulted in suboptimal optical character recognition (OCR) quality. We experimented with both ABBYY FineReader OCR and Tesseract OCR, but neither significantly improved the accuracy of the text extraction. Second, the extensive and complex nature of the information made it difficult to extract parallel data. The dictionary entries included multiple meanings separated by commas, semicolons, Arabic or Roman numerals, letters, and additional details enclosed in brackets. Buryat words were presented in bold, grammatical information in italics, and Russian translations in regular font. Example of the dictionary article presented in Figure 1. To parse this structured information, we relied on regular expressions. Third, some pages suffered from unrecognized fonts, which required alternative approaches. For these cases, we employed the large language model Claude Sonnet 3.5[6]. However, using a multimodal large language model to process the entire book was not feasible due to the high associated costs.

Religious literature is another common source of parallel data. For Buryat, the only available resource was the Bible[7]. We aligned the text using regular expressions based on the enumerated verses. However, the translations were not always precise. Certain content in the Buryat Bible was omitted, and in some instances, multiple verses were combined into a single sentence. These issues were also addressed using regular expressions to ensure proper alignment and extraction.

We identified several bilingual books, which were translations between Buryat and Russian. A key challenge was aligning sentence pairs from these texts, as differences in structure and translation styles introduced inconsistencies. To address this, we fine-tuned the LaBSE encoder (Feng et al., 2022) on a previously collected dataset using the methodology described in (Dale, 2022), allowing us to effectively extract parallel sentences.

We also explored the use of Wikipedia as a potential source of parallel data. However, the corresponding articles in Buryat and Russian were found to be significantly different. This discrepancy likely stems from the fact that much of the Buryat Wikipedia content was translated from Russian prior to 2015, while the Russian articles have undergone substantial changes since then. As a result, we were unable to extract high-quality sentence pairs from Wikipedia and it was excluded from the final dataset.

A buryat monolingual corpus[8] was created by collecting texts from books sourced from websites[9] [10] [11] and Internet news articles in Buryat[12]. The monocorpus is used to enhance tokenizer of translation models. To further expand the parallel corpus, a subset of the news articles was translated into Russian using the large language model Claude 3.5 Sonnet (20240620). At the time of creation, this model provided best translation quality, enabling us to significantly enrich the dataset and improve the overall performance of the translation system.

Finally, to improve the quality of the dataset, we filtered out poorly aligned sentences using a heuristic based on sentence length and cleaned up Russian borrowings by applying a heuristic based on Levenshtein distance, both methods following the approach outlined in Dale (2022). After this cleaning process, we obtained a final dataset of 33 thousand words and 94 thousand sentences. The detailed breakdown of amounts by source is provided in Table 1.

[6] https://www.anthropic.com/claude

[7] https://ibt.org.ru/buryatskiy/vsya-bibliya/elektronnaya-kniga

[8] https://huggingface.co/datasets/buryat-translation/buryat_monocorpus

[9] https://old.buryatika.ru/

[10] https://soyol.ru/culture/books/

[11] https://nomoihan.com/books/

[12] https://burunen.ru/bur/

Source	Amount
Dictionary phrases	45,169
Dictionary words	33,449
Book alignments by BRC SB RAS	12,893
News translated with Claude	11,380
The Bible	8,591
Organized manual translations	11,392
Book alignments by model	4,415
Tatoeba	808

Table 1: Data sources, total of 94 thousand sentences and 33 thousand words

4 General concept of neural network

4.1 Creation of a Russian-Mongolian Parallel Corpus

High-quality neural machine translation requires large amounts of parallel data for training. However, the Buryat language is severely underrepresented in digital resources and is considered a low resource language. In such cases, transfer learning techniques or model adaptation based on related languages are often employed to improve translation performance.

The closest high-resource cognate language to Buryat is Modern Mongolian. It is more widely represented in digital space and is commonly included in multilingual models. Pretraining on Mongolian can thus serve as a valuable step for enhancing Russian–Buryat translation.

In order to test this theory, we tried to find Russian Mongolian parallel corpora that was publicly accessible. The only relevant dataset was found in the OPUS collection, which is one of the largest repositories of multilingual corpora. The corpus contains 387,310 sentence pairs. However, many of these translations were found to be of insufficient quality or poorly aligned, making the dataset unsuitable for direct use.

This led us to the conclusion that generating Russian-Mongolian parallel data by translating Russian texts into Mongolian using pretrained multilingual models was a better option. To determine the most accurate model for this task, we evaluated several candidates that support both Russian and Mongolian:

1. `facebook/nllb-200-distilled-600M` (NLLB Team et al., 2022)
2. `facebook/nllb-200-1.3B` (NLLB Team et al., 2022)
3. `google/madlad400-3b-mt` (Kudugunta et al., 2023)

The steps outlined below were taken to determine the best model to use for creating a Russian-Mongolian parallel corpus:

1. Each candidate machine translation model was used to translate a shared set of Russian sentences into Mongolian.
2. The generated translations were compared to the corresponding Mongolian references in the OPUS corpus using the ChrF++ metric.
3. The model with the highest average ChrF++ score was selected as the most accurate for Russian–Mongolian translation (see Table 2). The same Russian source corpus used for the Russian–Buryat data — the Taiga corpus — served as the basis for the synthetic Russian–Mongolian dataset. In this case, text clustering was not applied, as a large volume of data was preferred for pretraining purposes.

The model `facebook/nllb-200-1.3B` was found to perform best and was used to translate a total of 90,548 Russian sentences from the Taiga corpus.

4.2 Model Selection for Russian–Buryat Machine Translation

Given the low-resource nature of the Buryat language, selecting an appropriate neural architecture is critical for achieving reasonable translation quality. When selecting the model architecture, we opted for encoder-decoder type, as the cross-attention mechanism enables the model to better capture dependencies within the input and incorporate contextual information during decoding — a crucial aspect in machine translation. Experimental results presented in Raffel et al. (2023) and Fu et al. (2023) have shown that encoder-decoder models consistently outperform decoder-only architectures in translation tasks.

The first model selected for training on the Russian–Buryat parallel corpus was Google's `mt5-large`. This model was chosen due to its strong performance on machine translation tasks and broad multilingual support, including related languages such as Mongolian, making it a suitable candidate for low-resource scenarios.

The second model, `nllb-200-distilled-600M` by Meta (formerly Facebook), was specifically designed for multilingual machine translation with a focus on

Model	Average ChrF++ Score
facebook/nllb-200-distilled-600M	26.4
facebook/nllb-200-1.3B	**27.8**
google/madlad400-3b-mt	10.8

Table 2: Comparison of machine translation models using the ChrF++ metric

low-resource languages. Its compact architecture and high translation efficiency, as demonstrated in the model comparison presented in Section 4.1, make it particularly well suited to the task at hand.

4.3 Final Training Procedure

Before initiating the main training process, it was necessary to update the tokenizer vocabulary by incorporating new tokens specific to the Buryat language, which is not included in the original models. For well-represented languages in the training data, it is typical for each word to correspond to 2–3 tokens on average. However, Buryat words are segmented into a significantly larger number of tokens, indicating insufficient vocabulary coverage (Figure 2).

To address this, we utilized a Buryat monolingual corpus to extend the tokenizer. We used a dedicated dataset, described in Section 3.3, and supplemented it with Buryat sentences extracted from the training data. A new SentencePiece tokenizer was trained on this combined corpus.

The missing tokens identified in the newly trained tokenizer were then added to the original vocabulary of the NLLB tokenizer. Corresponding embedding vectors were initialized and appended to the model's embedding layer, ensuring that the model could represent and learn these new units during training.

The roles of language tags are crucial to the NLLB tokenizer. These special tokens are added to the beginning of source and target sentences to explicitly indicate the language. For Russian–Buryat translation, we added the tag `bxr_Cyrl` to both the tokenizer and the model configuration.

Following this preparation, we proceeded with training the neural machine translation models. Training was performed in both directions (Russian–Buryat and Buryat–Russian), with the direction chosen randomly for each batch. Details of the training corpus, hyperparameter settings, and results are provided in Section 5.

5 Experiments

We now turn to the experimental setup. Multiple versions of the `mt5-large` and `nllb-200-distilled-600M` models were trained. Each version was trained on an incrementally larger dataset, as the Russian–Buryat parallel corpus was continuously updated with newly translated sentence pairs.

For both models, the following hyperparameters were used:

- Batch size: 16
- Maximum sequence length: 512
- Number of training steps: 60,000

To evaluate translation quality, we used the BLEU and ChrF++ metrics, which are widely adopted in machine translation research.

The `mt5-large` model was pre-trained on the Russian–Mongolian parallel corpus described in Section 4.1.

In contrast, no additional pretraining was applied to the NLLB model, as it demonstrated strong performance during the Russian–Mongolian model comparison and achieved results comparable to the `facebook/nllb-200-1.3B` model used to generate the synthetic corpus.

The training results of all model versions are presented in Table 3.

The initial version of the model, referred to as `Fine-tuned NLLB-v0`, was trained before the manual translation process had begun. As a result, this version did not include any of the high-quality human-translated data. This limitation affected the overall translation quality, but the model served as a useful baseline for evaluating the impact of incorporating manually translated content in later versions.

Starting from the first version, manually translated data was incorporated into the training set. Additionally, we refined the regular expressions used for mining data from the dictionary and introduced back-translated data generated by Claude. As expected, translation quality improved with

bxr	bxr words	bxr tokens
Зарим хүнүүдтэ һүлдэ сэсэн мэргэн үгэ хайрлада…	[Зарим, хүнүүдтэ, һүлдэ, сэсэн, мэргэн, үгэ, х…	[_Зарим, _хүн, үүд, тэ, _h, үл, дэ, _с, э, сэн…
һүзэглэгшэдэй бэе бэедээ дуратай байхые урмашу…	[һүзэглэгшэдэй, бэе, бэедээ, дуратай, байхые, …	[_h, үз, эг, лэг, ш, эд, эй, _бэ, е, _бэ, ед, …
бидэ гурбан эрэшүүлые зорюута буурал хүгшөөдэ…	[бидэ, гурбан, эрэшүүлые, зорюута, буурал, хүг…	[_бид, э, _гур, бан, _эр, эш, үү, лые, _зор, ю…

Figure 2: Example of Buryat token segmentation

	ru-bxr		bxr-ru	
model	**BLEU**	**ChrF++**	**BLEU**	**ChrF++**
Fine-tuned NLLB-v0	6.56	22.14	1.31	10.00
Fine-tuned NLLB-v1	18.74	46.17	32.20	53.37
Fine-tuned mT5-v1	12.49	39.39	14.47	37.28
Fine-tuned NLLB-v2 (last)	**20.61**	**48.68**	**35.43**	**56.21**
Google Translate	8.93	37.61	29.58	52.35
Claude 3.5 Sonnet 20240620	8.00	34.80	25.12	52.03

Table 3: Evaluation of our and Google Translate model on test-set

each iteration. At the first stage, based on the observed performance, we decided to continue using only the NLLB-based model for further development. Once additional manually translated data became available, we trained the second version of the NLLB model, which, at the time of writing, represents the latest iteration. This version achieved the best results for Russian–Buryat translation.

To assess the performance of our model `Fine-tuned NLLB-v2`, we compared against publicly available systems: Google Translate and Claude 3.5 Sonnet. As shown in Table 3, our model outperforms both baselines in both directions (Russian–Buryat and Buryat–Russian), achieving ablhigher scores in both BLEU and ChrF++.

Translation performance varies across text types and directions (Table 4). The NLLB-v2 model achieved higher scores on manual translations, likely because it is most familiar with this domain. In the case of Bible texts, Google Translate performs best in the Buryat-to-Russian direction—possibly due to similar phrasing in its training corpus—while NLLB is stronger in the opposite direction. Phrasebook examples result in the lowest scores overall, which could be explained by their short length, limited context, and the frequent presence of set expressions, all of which make them difficult to translate reliably. In literary and legal texts, NLLB-v2 and Claude show similar performance in the Buryat-to-Russian direction, though reasons remain unclear. It is possible that Claude was trained on similar data.

It is important to note, however, that both the training and test sets used in our experiments were derived from the same pool of source texts, although split and processed independently. While this setup allows or stable evaluation, it may introduce a slight bias in favor of our model due to potential domain similarity. Still, the consistent advantage in scores suggests that our model performs better for Russian–Buryat translation than Google Translate and Claude, particularly in the Russian-to-Buryat direction.

6 Online translator

To make our machine translation model accessible to the public, we released it online[13]. Figure 3 demonstrates the graphic user interface of the translator. To make the model suitable for usage in web, we made quantization of the model with ctranslate (Klein et al., 2020).

[13] `https://www.burtranslate.ru/`

		ru-bxr		bxr-ru	
Source Type	**Model**	**BLEU**	**ChrF++**	**BLEU**	**ChrF++**
Manual translations	Fine-tuned NLLB-v2	**21.88**	**52.15**	**38.10**	**60.20**
	Google Translate	8.56	39.18	23.69	52.31
	Claude 3.5 Sonnet 20240620	9.43	38.40	33.42	59.43
The Bible	Fine-tuned NLLB-v2	**20.49**	**47.56**	40.07	57.57
	Google Translate	10.00	36.83	**54.36**	**68.61**
	Claude 3.5 Sonnet 20240620	3.67	29.10	11.72	37.29
Phrasebooks	Fine-tuned NLLB-v2	**6.25**	**28.69**	**11.20**	30.89
	Google Translate	5.94	25.74	8.33	28.75
	Claude 3.5 Sonnet 20240620	4.43	25.82	8.43	**31.64**
Literature and regulations	Fine-tuned NLLB-v2	**16.83**	**39.86**	18.01	40.48
	Google Translate	9.18	36.12	13.20	37.45
	Claude 3.5 Sonnet 20240620	9.40	33.93	**24.94**	**49.64**

Table 4: Evaluation of our model and Google Translate on test-set by source types.

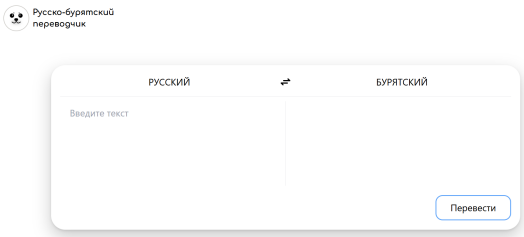

Figure 3: Graphic user interface of online translator

7 Human evaluation

We asked six native speakers of Buryat to participate in the evaluation of translations. Given that the average age of the participants was 57.8 years, we opted for a simplified rating scale consisting of two criteria: accuracy and fluency, both rated on a 5-point scale (with 5 indicating a perfect translation):

- Accuracy: Assesses how faithfully the translation preserves the meaning of the original sentence. Accuracy: Assesses how faithfully the translation preserves the meaning of the original sentence.
- Fluency: Evaluates the grammatical correctness and naturalness of the translation in the target language.

Each participant assessed 15 sentences in each translation direction (bxr-ru and ru-bxr), resulting in a total of 90 unique sentences evaluated manually. To ensure reliability, each sentence was reviewed by two different raters. The evaluated texts were randomly selected from the test corpus. The average scores are summarized in Table 5.

The manual evaluation suggests that the translations produced by the model are generally acceptable, particularly in terms of accuracy. However, lower fluency scores — especially in pessimistic cases — indicate that the output sometimes lacks grammatical correctness or natural phrasing. This highlights the need for further improvement.

8 Conclusion

In this work, we introduce a Buryat-Russian machine translation model, along with a parallel corpus of 127K sentence pairs and a monolingual Buryat corpus of 214K sentences. All resources are publicly released to support further research in low-resource language technologies.

Our model shows slightly better performance compared to Google Translate's Buryat-Russian system on our test dataset. Native speakers have evaluated the translations as acceptable for practical use.

We hope that this work will contribute to the development of computational linguistics for the Buryat language and provide a foundation for future research. By making these resources available, we aim to support efforts toward the preservation and promotion of Buryat in the digital domain.

Limitations

Machine translation systems for Buryat have great potential to support language learning and increase the availability of content in Buryat. However,

Table 5: Average manual evaluation scores for bxr-ru and ru-bxr

Metric	Group 1	Group 2	Group 3	Total Averages
ru-bxr				
Average Accuracy	4.13	4.06	3.19	3.79
Average Fluency	4.00	3.97	2.43	3.47
Pessimistic Accuracy	3.80	3.40	2.60	3.27
Pessimistic Fluency	3.67	3.27	1.73	2.89
bxr-ru				
Average Accuracy	3.34	3.63	3.06	3.34
Average Fluency	3.33	3.57	2.53	3.14
Pessimistic Accuracy	2.73	2.87	2.47	2.69
Pessimistic Fluency	2.73	2.87	1.40	2.33

these systems are not without significant limitations that need to be addressed.

A major concern is the accuracy of translations. Machine translation often makes mistakes, such as generating non-existent words, providing incorrect definitions, or producing grammatical errors. These inaccuracies can lead to misunderstandings and may even influence the language negatively if users unknowingly adopt incorrect forms. Additionally, the current model is still under development and cannot yet be fully trusted. Users are advised to double-check translations, especially in critical contexts, as over-reliance on automated systems can result in errors being propagated.

Another challenge is the lack of representation of Buryat dialects. Most models are trained on the literary standard of the language, leaving out the rich diversity of regional variations. This focus on a single dialect makes it harder for speakers of other dialects to benefit from the system and limits learners' exposure to the full range of linguistic expression within the Buryat language.

Cultural and contextual nuances also present difficulties. Machine translation struggles with idiomatic expressions, metaphors, and culturally specific references, which can lead to mistranslations or loss of meaning. For a language like Buryat, which carries deep cultural significance, this limitation can hinder effective communication.

Finally, the scarcity of high-quality training data further restricts the system's capabilities. Limited and imbalanced datasets can introduce biases and reduce performance, particularly in informal or specialized contexts. Addressing these challenges will require expanded and more diverse datasets, as well as ongoing refinement of the model.

While machine translation systems offer valuable support for Buryat, careful attention must be paid to these limitations to ensure their responsible and effective use.

Acknowledgments

The work is supported by the Mathematical Center in Akademgorodok under the agreement № 075-15-2025-349 with the Ministry of Science and Higher Education of the Russian Federation.

We would like to thank the translators who contributed to this project, as well as The State Translation Service of the Republic of Buryatia and The Buryat Research Center of the Siberian Branch of the Russian Academy of Sciences for providing parallel data. We are also grateful to David Dale for his great work on open-source low-resource translation.

Finally, we thank our relatives for their help in evaluating the models and for their support during the study.

References

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi N. Baljekar, Xavier García, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, and 5 others. 2022. Building machine translation systems for the next thousand languages. *ArXiv*, abs/2205.03983.

Ali B. Berberov, Bogdan S. Teunaev, and Liana B. Berberova. 2024. The first neural machine translation system for the karachay-balkar language. In *2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE)*, pages 1720–1723.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29.

David Dale. 2022. The first neural machine translation system for the Erzya language. In *Proceedings of the first workshop on NLP applications to field linguistics*, pages 45–53, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep 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).

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Zihao Fu, Wai Lam, Qian Yu, Anthony Man-Cho So, Shengding Hu, Zhiyuan Liu, and Nigel Collier. 2023. Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder. *Preprint*, arXiv:2304.04052.

Alexandru-Iulius Jerpelea, Alina Radoi, and Sergiu Nisioi. 2025. Dialectal and low resource machine translation for Aromanian. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7209–7228, Abu Dhabi, UAE. Association for Computational Linguistics.

Mamadou Keita, Elysabhete Ibrahim, Habibatou Alfari, and Christopher Homan. 2024. Feriji: A French-Zarma parallel corpus, glossary & translator. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–9, Bangkok, Thailand. Association for Computational Linguistics.

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The OpenNMT neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109, Virtual. Association for Machine Translation in the Americas.

VP Konovalov and ZB Tumunbayarova. 2018. Learning word embeddings for low resource languages: the case of buryat. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 331–341.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Toadoum Sari Sakayo, Angela Fan, and Lema Logamou Seknewna. 2023. Ngambay-French neural machine translation (sba-fr). In *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*, pages 39–47, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Lubsan Shagdarov and Konstantin Cheremisov. 2010. *Buryat-Russian dictionary*, volume 1-2. Republic typography, Ulan-Ude.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

UNESCO. 2010. *Atlas of the world's languages in danger*. UNESCO.