Enhancing Tuvan Language Resources through the FLORES Dataset

Ali Kuzhuget
Senior Member,
IEEE #100062617
iOS Developer
agisight@gmail.com

Airana Mongush
Machine Learning Founder
Algebras AI
aira@algebras.ai

Nachyn-E. Oorzhak
Data Scientist
moy.kot.tas.ool@gmail.com

Abstract

FLORES is a benchmark dataset designed for evaluating machine translation systems, particularly for low-resource languages. This paper, conducted as a part of Open Language Data Initiative (OLDI) shared task, presents our contribution to expanding the FLORES dataset with high-quality translations from Russian to Tuvan, an endangered Turkic language. Our approach combined the linguistic expertise of native speakers to ensure both accuracy and cultural relevance in the translations. This project represents a significant step forward in supporting Tuvan as a low-resource language in the realm of natural language processing (NLP) and machine translation (MT).

1 Introduction

Tuvan is a Turkic language, written using the Cyrillic alphabet and spoken by approximately 258,000 people (as of 2020), according to Ethnologue (2024). It is one of two official languages, along with Russian, of the Republic of Tuva, which is located in South Central Siberia, Russia. Despite its historical and cultural significance, Tuvan is classified as vulnerable by UNESCO, making it a critical target for preservation and technological integration. The FLORES (Goyal et al., 2022), spearheaded by Meta, aims to enhance machine translation systems by providing high-quality, controlled datasets for under-resourced languages for evaluation purposes. This paper, as a part of OLDI shared task (Initiative, 2024b), details our efforts to contribute to this dataset by providing translations from Russian to Tuvan.

2 Related work

It is essential to provide a brief overview of the FLORES (Goyal et al., 2022) dataset for those unfamiliar with this resource. The FLORES dataset, introduced by Goyal et al. (2022), is a benchmark for evaluating machine translation models on low-resource languages, which was translated to over 200 languages. It is comprised of two sets: the dev set contains 997 sentences and the devtest set includes 1012 sentences, that were sampled from Wikinews, Wikijunior and Wikivoyage. FLORES dataset is crucial in advancing NLP for languages like Tuvan by providing benchmark specifically designed to evaluate machine translation systems across a wide variety of languages. The NLLB project (NLLB Team et al., 2022) further exemplifies efforts to scale human-centered machine translation across diverse languages.

3 Language overview

3.1 Handling dialectal differences

As the Republic of Tuva is a federal subject of Russia, the majority of the population is bilingual, speaking both Russian and Tuvan. For that reason, Tuvan translators relied on the FLORES dataset in Russian when developing one in Tuvan. During the translation process, any variations in interpretation due to dialectal differences were resolved by defaulting to the Central dialect's interpretation. This approach ensured uniformity and consistency across the dataset, which is crucial for training machine translation models that need to generalize well across different contexts.

The language is taught in schools optionally, but one can get a higher education in the Tuvan language and Literature in one of the universities of the Republic. Although the number of youth that speaks the language fluently decreases, it is still widely used both in cities and rural areas.

3.2 Linguistic challenges

Tuvan is characterized by its complex phonological and grammatical structures, including vowel harmony and extensive use of suffixes. These features posed challenges in translation, particularly in ensuring that the meaning and tone of the original Russian texts were accurately conveyed in Tuvan. However, since our work was limited to written form, we did not address challenges related to vocal translation or spoken dialects.

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 593–599
November 15–16, 2024.

4 Data collection

4.1 Expertise of translators

Our translation team for the FLORES dataset was led by four native Tuvan speakers who are also proficient in Russian. The team included professional linguists and language enthusiasts with formal education in the Tuvan language. Although Tuvan language education is currently facultative, several of our translators had attended schools where Tuvan was the primary medium of instruction. This deep linguistic knowledge was crucial in ensuring that translations were not only accurate but also culturally relevant and sensitive. The following team of translators put their utmost effort to make the FLORES dataset available in Tuvan.

- Mongush Salim (Моңгуш Салим)
- Oorzhak Lyudmila (Ооржак Людмила)
- Ongai-ool Choduraa (Оңгай-оол Чодураа)
- Kuzhuget Ali (Кужугет Али)

4.2 Translation guidelines and training

Before beginning the translation tasks, all translators were provided with comprehensive guidelines, prepared in Russian, detailing the translation process. These guidelines, which are included in Appendix A, covered key aspects such as:

1. Maintaining the tone and style of the original text.
2. Handling idiomatic expressions and culturally specific references.
3. Ensuring pragmatic accuracy, including the correct use of pronouns and proper nouns.

Translators were instructed to adhere strictly to these guidelines in order to ensure a high level of consistency and quality across the entire dataset.

4.3 Managing the translation workflow

The translation process was managed using a Telegram group, where tasks were assigned, and progress was tracked using project management tools. This system allowed for effective coordination among the translators and ensured that the project stayed on schedule. It is worth pointing out that the workflow included multiple rounds of review and feedback among the translators to check one another and refine the translations further.

5 Experimental validation

5.1 Contribution to the evaluation of translation models

Our work makes a significant impact on the evaluation part of the Tuvan translation models by creating a reliable benchmark for this task. As for the current model of September 2024 developed by our team, it was trained on the existing Tuvan-Russian corpus, which consisted of approximately 200,000 pairs (Kuzhuget and Choigan, 2024) of translations sourced mainly from Wikipedia and other early Tuvan language projects (Kuzhuget et al., 2023). By contributing to the FLORES (Goyal et al., 2022) dataset, we have provided a more structured and high-quality resource, developed and checked by professionals, that is better suited for evaluating machine translation models.

The new dataset allowed us to run experiments to compare the quality of the existing translation models, available in Tuvan (Claude Sonnet 3.5, Google Translate v2 API, tyvan.ru). The results of these experiments are demonstrated on the Table 1.

Claude Sonnet 3.5 shows the highest performance overall, with BLEU scores of 35.65 and 33.04 for the Tyv-Rus translation in dev and devtest, respectively. The ChrF2++ scores are also the highest, reflecting good contextual understanding of this model. Google Translate v2 performs well, particularly with Tyv-Rus, achieving a BLEU score of around 28 in both datasets with ChrF2++ scores being also strong. The tyvan.ru model tends to have lower scores and lags behind the two above mentioned models.

Model	Dataset	Tyv-Rus		Rus-Tyv	
		BLEU	ChrF++	BLEU	ChrF++
tyvan.ru	dev	16.94	43.94	13.12	45.92
	devtest	16.41	43.09	13.35	46.11
Google Translate v2	dev	29.78	54.60	14.30	45.50
	devtest	27.16	52.87	15.58	46.18
Claude Sonnet 3.5	dev	35.65	59.65	16.67	48.45
	devtest	33.04	57.41	17.09	49.08

Table 1: Scores of Russian-Tuvan translation models on the FLORES dataset.

5.2 Manual evaluation of machine translation with the FLORES dataset

Objective

The objective of the manual evaluation on the FLORES dataset is to evaluate and compare the perceived translation adequacy of three translation services: Google Translate v2, Claude Sonnet 3.5,

and tyvan.ru. For that purpose we asked five Tuvan native speakers (annotators) to assess the quality of the translations for adequacy by giving scores on a scale of 1 to 5, the higher the better.

Data

31 sentences were translated from Russian to Tuvan using the above mentioned services. The data was taken from two sets of FLORES dataset:

- dev set: First 16 sentences from the dev set
- devtest set: Last 15 sentences from the devtest set

Annotators were given a table with the first column containing sentences in Russian and three other columns containing translated sentences to Tuvan by the services, they had no information of the service that provided the translation to make the evaluation unbiased.

Results

Google Translate v2

- Median score: ~ 4
- Distribution: Skewed toward higher scores (4-5), with some variability and occasional lower ratings.

Claude Sonnet 3.5

- Median score: ~ 4
- Distribution: Consistently high (3-5), with some sentences receiving lower scores (down to 2).

tyvan.ru

- Median score: $\sim 3 - 4$
- Distribution: Most variability, with a wider range of scores (1-5). This service had the most mixed feedback, with some high and many low ratings.

Key Insights

Google Translate v2 and Claude Sonnet 3.5 performed relatively well, with consistent high scores. tyvan.ru showed more variability in performance, reflecting either inconsistent translation quality or differences in how annotators perceived adequacy.

Visuals

Box plots (Figure 1) were used to compare the distribution of scores across services in both dev and devtest sets. A histogram (Figure 2) illustrated the overall distribution of scores for each service, with distinct colors for easy comparison (yellow for Google Translate v2, blue for Claude Sonnet 3.5, red for tyvan.ru). This experiment highlights differences in translation quality as perceived by human annotators, showing that while Google Translate v2 and Claude Sonnet 3.5 generally perform well, tyvan.ru's performance was less consistent across annotators.

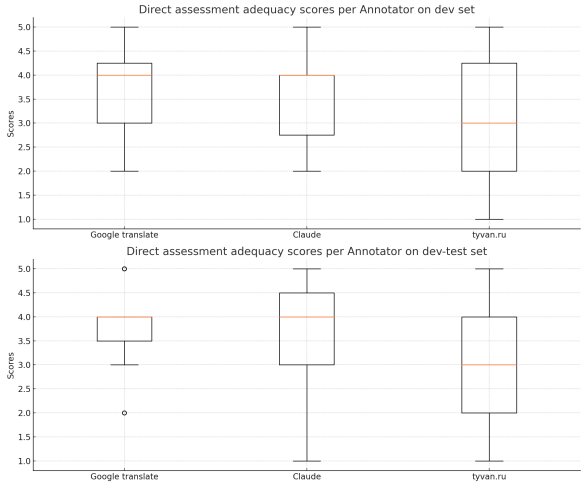

Figure 1: Direct assessment adequacy scores per Annotator

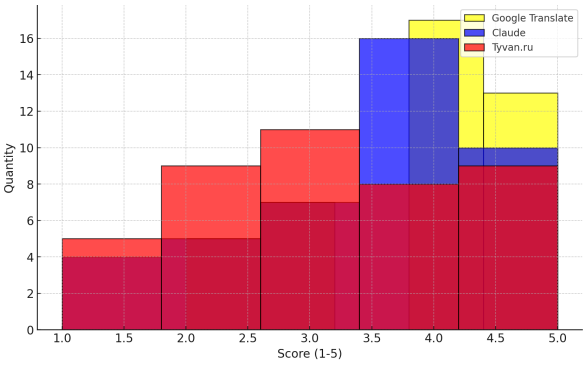

Figure 2: Perceived translation score distribution by Service

6 Data sample

As far as the the FLORES dataset in Tuvan is concerned, it has the following characteristics:

- dev set: 997 sentences, 18.26 average number of words in a sentence;
- devtest set: 1012 sentences, 18.44 average number of words in a sentence.

A table 2 showcases the examples of translated sentences from Russian to Tuvan. English translations were provided as examples so that English speakers could understand the general meaning of these sentences. But it is important to highlight that our translators were based only on the Russian

No.	English source	Russian translation	Tuvan translation
1	Tokyo will be the only city in Asia to have hosted the Summer Olympics twice.	Токио станет единственным городом Азии, который дважды принимал летние Олимпийские игры.	Токио Азияга чайгы олимпийжи оюннарны ийи катап эрттирген чаңгыс хоорай болуп арттып каар.
2	There were no reports of serious damage or casualties in Tonga, but there was a temporary power outage, which reportedly prevented the authorities from receiving the tsunami warning sent by the Pacific Tsunami Warning Center (PTWC).	Из Тонга не поступило сообщений о серьезных разрушениях или о пострадавших, но произошло временное отключение электроснабжения, что, по имеющимся данным, не позволило властям Тонга получить предупреждение о цунами, посланное Тихоокеанским центром предупреждения о цунами (PTWC).	Тонгадан шыңгыы үрегдээшкиннер азы когарааннар дугайында медээлер келбээн, ынчалза-даа электри хандырылгазы түр када хже берген, ол чүүл, амгы үеде бар медээлер-биле, Цунами дыңнадыр Ооожум-океанчы төптен (PTWC) цунами дугайында дыңнадыгны алыр арганы Тонганың чазаанга бербээн болуп турар.
3	The percentage of people with multidrug-resistant tuberculosis in the overall group of tuberculosis patients still appears low; 6000 out of 330,000 infected in South Africa.	Однако, процент людей с туберкулёзом с множественной лекарственной устойчивостью в целой группе больных туберкулезом все еще кажется низким; 6000 от общего числа 330 000 зараженных в ЮАР.	Туберкулез аарыг бөлүкте хөй санныг эмнер-биле эмнеттинмес туберкулезтуг кижилерниң хуузу ам-даа эвээш ышкаш сагындырар; амгы үеде ЮАР-да аарыг 330 000 кижиниң ниити санындан 6000 кижи.
4	In Japan, the first celebrations of cherry blossom viewing were arranged by the emperor only for himself and other members of the aristocracy.	В Японии первые празднования цветения сакуры устраивались императором только для себя и других членов аристократии при императорском дворе.	Японияга сакура частырының баштайгы байырлалдарын императорнуң чүгле бодунга болгаш императорнуң чанында өске-даа аристократчы кежигүннерге дээш эрттирип турган.
5	The airlines offering these services include: Air Canada, Delta Air Lines, Lufthansa - for flights departing from the USA or Canada, and WestJet.	Авиакомпании, предлагающие эти услуги включают: Air Canada, Delta Air Lines, Lufthansa — для рейсов, отправляющихся из США или Канады, и WestJet.	Ол ачы-дузаны чедирип турар авиакомпанияларже Air Canada, Delta Air Lines, Lufthansa – АКШ-тан азы Канададан чоруп турар рейстерге, база WestJet олар хамааржыр.

Table 2: Examples of translated sentences from English to Russian and Tuvan.

FLORES dataset and did not consider the English part during the development of the FLORES in Tuvan.

7 tyvan.ru and language preservation

The development of a Tuvan AI translator has been instrumental in preserving and revitalizing the Tuvan language, classified as endangered by UNESCO. Born out of a personal commitment to language preservation, the project began as a response to the lack of online translation resources for Tuvan. The initiative gained momentum with contributions from David Dale, who recognized the urgent need for language preservation during his work on the Erzya language (Dale, 2022), and Ali Kuzhuget, who has spent over a decade developing Tuvan language resources (Kuzhuget and Choigan, 2024), including dictionaries, keyboards, and translations of major platforms.

tyvan.ru serves as the online hub for these efforts, offering a range of Tuvan language tools, including the AI translator. Since its launch, the translator has been used by over 80,000 individuals, showcasing the growing interest in and need for Tuvan language resources. The project also extends its impact through the "One Code - Different Languages" volunteer initiative, which supports other vulnerable low-resource languages, such as Bashkir, Tatar, Chuvash, and Mari.

8 Limitations

The main challenge that occurs when dealing with low-resource languages like Tuvan is the small number of professionals, who could correctly translate to a low-resource language, abiding by all the grammar rules and nuances of the language. This results in a rather long translation process. Another project we are engaged in is the Seed (Maillard et al., 2023) project in Tuvan. One of the key limitations we faced in the second project is the lack of Tuvan speakers who are bilingual in any language other than Russian. Due to the historical context of Tuvans being part of the Soviet Union and Russia, Tuvan speakers typically only speak Russian in addition to Tuvan. Consequently, our extended research group is required to translate the Seed dataset from English to Russian first, and then from Russian to Tuvan. This two-step translation process introduces additional complexity and potential for translation inaccuracies, but it is a necessary approach given the linguistic resources available.

9 Conclusion

Our contribution to the FLORES dataset (Goyal et al., 2022) represents a significant step forward in supporting Tuvan as a low-resource language in the field of natural language processing. By focusing on the Central dialect and leveraging human expertise, we have created a high-quality resource that will aid in the development of more accurate and culturally sensitive machine translation systems.

In addition, we are currently in the process of translating the Seed (Maillard et al., 2023) dataset from English to Russian, and subsequently from Russian to Tuvan. This effort further enhances the resources available for Tuvan, contributing to the development of multilingual datasets and promoting the digital presence of the language.

This work not only enhances the digital presence of the Tuvan language but also contributes to its preservation and promotion.

10 Acknowledgments

We would like to express gratitude to the team of translators who implemented the FLORES dataset in Tuvan despite all limitations and difficulties. We also commend David Dale, author of the article on training a NLLB-200 model on Tuvan-Russian corpus Dale (2023), language enthusiast and a key figure in the Post-Soviet NLP community, whose insights and guidance were instrumental in refining our benchmarks and translation guidelines. David's commitment to advancing NLP for low-resource languages has greatly influenced our work, helping us achieve a higher standard of quality and relevance in our translations.

References

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.

David Dale. 2023. How to fine-tune a nllb-200 model for translating a new language, medium article. Accessed August 19, 2024.

Ethnologue. 2024. Ethnologue. Accessed August 19, 2024.

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.

Open Language Data Initiative. 2024a. Translation guidelines. Accessed August 19, 2024.

WMT24 Open Language Data Initiative. 2024b. The flores+ evaluation benchmark for multilingual machine translation. Accessed August 19, 2024.

Ali Kuzhuget and Ondar Choigan. 2024. Russian-tuvan parallel corpus, crowdsourcing organised by ali kuzhuget. Accessed August 19, 2024.

Ali Kuzhuget, Airana Mongush, and David Dale. 2023. The first Tyvan AI language project, tyvan.ru. Accessed August 19, 2024.

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.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

A Appendix. Translation guidelines

These guidelines were adapted and translated to Russian for the team of translators based on the OLDI translation guidelines (Initiative, 2024a).

Version 1.01
Author: Күжүгет А.А
Date: 2 марта 2024 г.

A.1 Важное примечание:

Ваши переводы будут использоваться для обучения или оценки движков машинного перевода. Поэтому этот проект требует человеческого перевода.

A.2 Общие рекомендации:

1. Контекст: Вы будете переводить предложения из разных источников. В некоторых случаях может быть предоставлена ссылка на исходный документ, чтобы дать вам больше контекста. Если она доступна, пожалуйста, обратитесь к ней.
2. Единицы измерения: Не переводите единицы измерения. Переводите их точно так, как указано в исходном содержании.
3. Сохранение тона: При переводе сохраняйте тон, используемый в исходном документе. Например, энциклопедический контент из источников вроде Википедии должен переводиться с использованием формального тона.
4. Плавность перевода: Предоставляйте плавные переводы, не отклоняясь слишком сильно от структуры исходного текста. Допускаются только необходимые изменения.
5. Точность: Не расширяйте или не заменяйте информацию по сравнению с тем, что присутствует в исходных документах. Не добавляйте никакой поясняющей или скобочной информации, определений и т.д.
6. Полнота перевода: Не игнорируйте любой значимый текст, который был в исходнике.
7. Выбор перевода: В случае нескольких возможных переводов, пожалуйста, выберите тот, который имеет наибольший смысл (например, для соответствия гендеру, культурной адаптации на целевом языке, уровня формальности и т.д.).

A.3 Именованные сущности:

Именованные сущности - это люди, места, организации и т.д., которые обычно упоминаются с использованием собственного имени. Этот раздел содержит рекомендации о том, как обращаться с именованными сущностями:

1. Общепринятые названия: Если в целевом языке существует общепринятое название для именованной сущности, используйте его.
2. Транслитерация: Если общепринятое название отсутствует, используйте транслитерацию оригинального термина, если это возможно. Если транслитерация не будет широко понята в контексте, вы можете сохранить оригинальный термин.

A.4 Идиоматические выражения:

Идиоматические выражения не должны переводиться дословно. Используйте эквивалентную идиому, если таковая существует. Если эквивалентная идиома отсутствует, используйте идиому схожего смысла. Если в целевом языке не существует похожих выражений, перефразируйте идиому так, чтобы значение было сохранено на целевом языке.

A.5 Неоднозначные местоимения:

Когда переводимое местоимение является неоднозначным (например, может быть интерпретировано как он/она или его/ее), выбирайте гендерно-нейтральные местоимения (такие как они), если таковые существуют на целевом языке. Однако, когда местоимение в исходном тексте четко обозначено по гендеру, вы должны следовать исходному материалу и сохранять гендерную маркировку.