FLORES+ translation and machine translation evaluation for the Erzya language

Isai Gordeev
École Polytechnique
Paris, France
isai.gordeev@polytechnique.edu

Sergey Kuldin
Astana, Kazakhstan
sergey@kuldin.com

David Dale
Paris, France

Abstract

This paper introduces a translation of the FLORES+ dataset into the endangered Erzya language, with the goal of evaluating machine translation between this language and any of the other 200 languages already included into FLORES+. This translation was carried out as a part of the Open Language Data shared task at WMT24. We also present a benchmark of existing translation models bases on this dataset and a new translation model that achieves the state-of-the-art quality of translation into Erzya from Russian and English.

1 Introduction

The Erzya language is the language of Erzya, one of the indigenous peoples of Russia. Despite its official status in the Republic of Mordovia (along with Russian and Moksha), the use of the Erzya language is limited mainly to everyday topics, and its representation in the digital space remains low. On the one hand, this situation contributes to the reduced status of language in society and inhibits its development and intergenerational transfer, and is one of the factors that make the language endangered (UNESCO, 2010). On the other hand, this makes it more difficult to develop natural language processing (NLP) technologies for Erzya, such as machine translation (MT), the availability of which could help increase the prestige of the language. Development of MT technologies for Erzya, in turn, is hampered by the lack of a generally recognized dataset for evaluating the translation quality.

In this article, we aim to close this gap by publishing the Erzya version of the FLORES+ dataset. This dataset, created as part of the No Language Left Behind project (NLLB Team et al., 2022) and later transferred to the community-run Open Language Data Initiative[1], became the de facto standard for evaluating MT quality for low-resource languages. FLORES+ consists of two thousand sentences sampled from English texts of Wikimedia projects and translated into more than 200 languages. The emergence of the Erzya version of FLORES+ will allow researchers to evaluate the quality of MT between any of these languages and Erzya. This translation has been submitted to the Open Language Data shared task at the WMT24 conference[2].

The quality of the new FLORES+ translation has been validated by independent annotators (for a sample from the dataset) and with a set of automatic metrics of accuracy and fluency which were themselves validated against human judgements.

In addition, we present a new neural model for translating between Erzya and other languages (primarily English and Russian), created by fine-tuning an NLLB-200 model. It achieves a BLEU score of 22% for translation from Erzya to Russian and 17% in the opposite direction, which implies the translation quality suitable for practical applications. Along with our new model, we evaluated the translation quality of several other models that also support the Erzya language. The model from Yankovskaya et al. (2023), also based on NLLB, achieved the highest scores for translation from Erzya into Russian, and a Claude model (Anthropic, 2024), into English, whereas our model achieved the highest scores for translation into Erzya.

In total, the contributions[3] of this article are:

1. We release and describe the first complete translation of the FLORES+ dataset into the Erzya language and validate its quality.
2. We evaluate the performance of available MT systems for Erzya on this dataset.
3. We present a model for translation between Erzya and several high-resourced languages, a state-of-the-art for translating into Erzya.

[1] https://oldi.org

[2] https://www2.statmt.org/wmt24/open-data.html

[3] Our code, data, and models will be made publicly available at https://github.com/slone-nlp/myv-nmt.

Proceedings of the Ninth Conference on Machine Translation, pages 614–623

2 Related work

2.1 The FLORES+ dataset

FLORES+ is the next stage in the evolution of the FLORES-200 dataset (NLLB Team et al., 2022), which, in turn, grew out of FLORES-101 (Goyal et al., 2022). The dataset is based on 3001 English sentences taken from three sources: Wikinews (international news), Wikijunior (non-fiction literature for children), and Wikivoyage (travel tips). It is divided into three roughly equal parts (dev, devtest, and test), of which the first two (2005 sentences) are included in FLORES+. The sentences were translated from English into 203 other languages by professional translators; among these languages are Russian and three Finno-Ugric ones (related to Erzya): Finnish, Estonian, and Hungarian.

A small subset of FLORES-200 (the first 250 sentences from the devtest subset, news domain) were translated by Yankovskaya et al. (2023) into 9 low-resource Finno-Ugric languages, including Erzya, and used to evaluate the quality of the machine translation system from this article.

As far as we know, no other multiway parallel datasets including the Erzya language have been published (one exception is the dataset of the Tatoeba project[4], however, it currently contains less than 100 Erzya sentences).

2.2 Erzya datasets

The available corpora of the Erzya language (especially parallel ones) are not numerous. Rueter and Tyers (2018) presented the Erzya corpus with morphosyntactic markup, including the translation of several hundred sentences into English and Finnish. Arkhangelskiy (2019) has compiled a web corpus of the Erzya language (available for download); there is also a corpus of the literary Erzya language, avaliable only for search[5]. Another corpus of the Erzya language, also searchable, is described by Rueter (2024a).

Medium-scale parallel datasets of Erzya with other languages have been considered only in recent papers on neural machine translation for Erzya: Dale (2022) and Yankovskaya et al. (2023).

2.3 Machine translation for Erzya

The Apertium platform implements a machine translation system between Erzya and related Moksha and Finnish languages[6], but this work gives an impression of being incomplete. In Dale (2022), one of the first machine translation systems for the Erzya language was created, based on a parallel Russian-Erzya corpus consisting of dictionaries and automatically aligned sentences from books and web publications (a total of 77K parallel pairs of sentences, words and phrases, as well as 333K sentences in Erzya without translation). Yankovskaya et al. (2023) collected parallel and monolingual datasets for 20 low-resource Finno-Ugric languages and trained a neural model for their translation, but the parallel part of their dataset remained unpublished. In both works, the new languages were added to the pre-trained multilingual translation models (mBART-50 (Tang et al., 2020) and NLLB-200 (NLLB Team et al., 2022), respectively) by adding new tokens to the model vocabulary and fine-tuning it with parallel and back-translated data.

The third known paper on neural machine translation for Erzya, Alnajjar et al. (2023), used the rule-based Apertium system to generate a synthetic Erzya-Moksha corpus, and also fine-tuned an NLLB-200 model on this data.

Finally, Mordovian State University announced the development of an Erzya-Russian MT system[7], but no publication on this topic has appeared yet.

2.4 Multilingual MT systems

In addition to MT systems explicitly trained with Erzya parallel data, some models might have learned this language from monolingual texts or parallel texts unintentionally found in web corpora. One could also hope that multilingual models can achieve some understanding of the Erzya language based on the grammar and vocabulary of other Finno-Ugric languages, as well as on the vocabulary borrowed by Erzya from Russian. Therefore, we are considering several public and proprietary systems that do not always explicitly include Erzya, but may still be suitable for its translation.

One of such systems is the NLLB-200 family of models, which is still the leader among open models in terms of the translation quality and language coverage trade-off. Their training data did not include the Erzya language (except perhaps a small number of web texts mistakenly classified as Russian), so the NLLB ability to translate Erzya is limited by the knowledge transferred from other languages. The

[4] https://tatoeba.org/
[5] https://erzya.web-corpora.net/
[6] https://github.com/apertium/apertium-myv-mdf and https://github.com/apertium/apertium-myv-fin
[7] See the announcement on msru.ru.

MADLAD-400 models (Kudugunta et al., 2024) were also not trained with Erzya parallel data, but used the monolingual web corpora collected in this paper (including Erzya) for unsupervised training and back-translation.

Finally, we consider three proprietary systems: Google Translate, which has recently added 110 new languages (Caswell, 2024) powered by the PALM-2 (Google, 2023) large language model (LLM), and the Claude (Anthropic, 2024) and GPT-4o (OpenAI, 2023) LLMs. All the three systems were trained with multilingual web data and with a large amount of parallel data (including, possibly, parallel texts for the Erzya language). Unfortunately, the technical reports for these systems give only very brief descriptions of their language coverage.

3 About the Erzya language

The Erzya language (myv) is one of the largest Finno-Ugric languages in the world and belongs to the Mordvinic branch of the Finno-Ugric group of the Uralic language family. Linguists distinguish five main dialects: Central, Western (Insar), northwestern (Alatyr), southeastern (Sura) and Shoksha (isolated in the northwestern regions). They differ mostly in their phonetic, and, to a lesser extent, morphological features. Our translation was based on the literary standard of the Erzya language (built primarily upon the Central dialect). Most modern Erzya is written using the Russian Cyrillic alphabet, although there are several Latin script proposals.[8]

Phonetically, the Erzya language contrasts palatalized and plain consonants and features vowel harmony. Grammatically, it is an agglutinative language with extensive systems of declension (including 12 noun cases, possesive suffixes and definiteness marking) and conjugation (including 7 moods and marking the person and number of subject and object). The word order is SVO, and postpositions are widely used. Lexically, most words have Finno-Ugric origins,[9] with a significant number of Russian and Turkic loanwords.

In the XX century, the number of Erzya was approaching one million people (according to the 1970 census, 1,263 thousand people, along with Moksha). The dispersed settlement of the people led to accelerated assimilation and the decrease in the number of native speakers. Thus, the Republic of Mordovia, where Erzya and Moksha are the titular nation, and their languages are co-official with Russian, hosts only about 30% of all Erzya, with the rest settled in Samara, Orenburg, Nizhny Novgorod, Penza, Saratov regions and other regions of Russia.

The status of the official language allows Erzya to function in the public space: there are newspapers and magazines, TV shows, theater, and popular music in this language. In addition, being a state language, Erzya is studied for 1-2 hours per week as an elective or mandatory lesson in many schools in the Republic of Mordovia, and in the settlements with a predominantly Erzya population, even as the first language.

Nevertheless, the Erzya language is poorly represented in the digital space. A few areas where it nevertheless functions are mentioned below:

- Several documentaries and feature films, music videos and video blogs; [10]
- The Wikipedia in Erzya with 7877 articles;[11]
- The Erzya interface of vk.com;[12]
- A few websites, thematic groups, and channels on social networks and messengers.

The problem of transferring the Erzya language to the younger generation is pressing: most children only understand, but almost do not speak their native language. We hope that translation of FLORES into Erzya will facilitate the development of machine translation for this language, which, in turn, could spur other technologies, such as speech synthesis, text and image generation models, contributing to the preservation, popularization and development of the language.

4 Translation of FLORES+

We translated FLORES+ from from Russian into the Erzya language. The translation was carried out by two native speaker volunteers who are also teachers of the language and writers (one with a doctoral degree in philology). The 250 translated sentences from Yankovskaya et al. (2023) were also included, but only after a thorough revision. All 2009 translations were revised by one of the native translators and a linguist with a profound expertise in the language. In addition, the translations were

[8] Table 2 features an example of a sentence in Erzya alongside with its Latin transliteration and translation.

[9] There are numerous, but not always easily recognisable cognates with other Uralic languages, such as Finnish and Estonian: for example, “ked / käsi / käsi / hand”, “koto / kuusi / kuus / six”, or “ĕräms / elää / elama / to live”.

[10] E.g., Azor and Кода эри эрзянь морось movies.

[11] https://myv.wikipedia.org

[12] A news article about the interface.

Neologisms	Examples in Erzya, Russian and English
Валдокаямо (luminosity, from валдо=light and каямо=output)	Весе **валдокаямось** ды чарамось сайневить тештень Россби числантъ муеманзо кис, конась сюлмазь плазмань потоконть марто. Совокупность **светимости** и вращения используется для определения числа Россби звезды, связанного с потоком плазмы. The **luminosity** and rotation are used together to determine a star's Rossby number, which is related to plasma flow.
Тевконёв (document, from тев=business and конёв=paper)	Ломантненень, конат арсить теемс сымень полавтомань операция омбо масторсо, эряви парсте ванкшномс, улезт сынст марто мекев самонень эрявикс **тевконёвост**. Люди, планирующие пройти операцию по смене пола за границей, должны убедиться, что у них при себе есть действительные **документы** для обратного пути. Voyagers planning sex reassignment surgery abroad must ensure they're carrying valid **documents** for the return trip.
Инедавол (hurricane, from ине=great and давол=storm)	Раськень **инедаволонь** кунщкакуронть коряс те шкас Джерри а канды кодамояк зыян мода лангс. Согласно Национальному **ураганному** центру, на данный момент Джерри не представляет никакой угрозы на суше. The National **Hurricane** Center (NHC) says that at this point Jerry poses no threat to land.
Озавтозетне (inmates, "the imprisoned")	Чокшне ланга 10:00 ды 11:00 шкантъ юткско пандонь шкантъ коряс **озавтозетне** тейсть кирвазтема вальмалост. Между 10:00 и 11:00 вечера по горному времени **заключенные** устроили пожар во дворе. Between 10:00-11:00 pm MDT, a fire was started by the **inmates** in the yard.
Кортницятне (negotiators, "talkers")	**Кортницятне** варчизь витемс тевенть, ансяк озавтозетнень вешемаст зть чарькодеве. **Переговорщики** попытались исправить ситуацию, но требования заключённых не ясны. **Negotiators** tried to rectify the situation, but the prisoners' demands are not clear.
Превмаксый (advisor, "intellect-giver")	1960 иетнестэ Бзежинский ульнесь Джон Ф. Кеннединь вакссо **превмаксыекс**, мейле Линдон Б.Джонсононь администрациясо. В 1960-х гг. Бжезинский занимал должность **советника** при Джоне Ф. Кеннеди, затем в администрации Линдона Б. Джонсона. Throughout 1960s, Brzezinski worked for John F. Kennedy as his **advisor** and then the Lyndon B. Johnson administration.

Table 1: Examples of lexical neologisms created according to the word-formation models of the Erzya language (the top 3) and semantic neologisms created by assigning a new meaning to already known words (the bottom 3).

scored with automatic quality metrics (Section 5), which helped identify several omissions and typos.

The successful translation of texts on topics that are uncommon for Erzya allows us to assess the capabilities of this language positively. However, we should note the difficulties in translating special terminology in various domains (such as science, politics, and sports). In such cases, we used lexical and semantic neologisms to solve the problem of insufficient vocabulary (see the examples in Table 1). For some sentences, to avoid distorting their meaning, we had to preserve the Russian terminology, (e.g. Table 2). In some cases, the neologisms are translations of one part of a complex word, for example, "пельсфинал/peĺsfinal" ("полуфинал" in Russian, "semi final" in English).

It would be difficult to directly evaluate how acceptable are these neologisms to the native speakers. But a human evaluation by two independent native speakers (Section 8) resulted in all sampled translations annotated as at least "acceptable", and the majority, as "good". This suggests that the new words are not perceived as serious problems to meaning preservation or fluency.

5 Automatic validation of data quality

To validate the quality of the newly translated Erzya FLORES dataset experimentally, we applied several automatic metrics of translation accuracy and fluency. To demonstrate the validity of the metrics themselves and to establish their baseline values, we needed human judgements of translation quality on some other dataset, and, fortunately, we had one.

Data. The baseline data consists of 1500 Erzya-Russian sentence pairs in the dev subset from Dale (2022), automatically aligned from various parallel documents. The sentence pairs were manually annotated by a proficient Erzya speaker for accuracy and fluency, with 0 standing for "unacceptable", 0.5 for "barely acceptable", and 1, for "good". The problems with meaning were mostly results of overly loose translations or incorrect alignment, whereas most of the fluency problems were caused by too literal translation from Russian (often by simply adding Erzya suffixes to the Russian words).

Our **simple metrics** are rel_sim (computed as the edit distance between the source and the target, divided by the maximum of the length of the two and then subtracted from 1) and len_ratio (the ratio

Examples in Erzya (Cyrillic), Erzya (transliterated to Latin), Russian and English
Докладсонть ули малав эрьва аспектэнь пшти критика малав неень исполнительной властень Ираконь коряс политикантъ коряс ды виев тердема сеске полавтомс улицяя курсонть.
Dokladsonť uli malav ěŕva aspektěń pšti kritika malav neeń ispolniteĺnoj vlasteń Irakoń koräs politikanť koräs dy viev terdema seske polavtoms ulicä kursonť.
В докладе содержится резкая критика почти каждого аспекта нынешней политики исполнительной власти в отношении Ирака и настоятельный призыв к незамедлительной смене курса.
The Report is highly critical of almost every aspect of the present policy of the Executive towards Iraq and it urges an immediate change of direction.

Table 2: An example of a translation with multiple loanwords from Russian (the loanwords are highlighted). Note that 4 out of these 6 words (aspektěń, kritika, politikanť, and kursonť), while being borrowed into Erzya from Russian, have ultimately Greek or Latin origins and are recognisable internationally.

of the character lengths of the source and target, the shortest of the two to the longest). We also include here the LID_rus metric: a probability, according to the GlotLID model (Kargaran et al., 2023), that the Erzya sentence is in fact Russian. We expect len_ratio to correlate with omissions; rel_sim and LID_rus are expected to correlate with fluency issues. The rest of the metrics, described below, target translation accuracy.

Dictionary-based metrics are computed by lemmatizing the words in a sentence pair and computing the proportion of words on the one side that has a counterpart on the other side that can be matched using a dictionary. WR_rus computes the proportion of Russian words whose translations can be found in the Erzya sentence, and WR_myv shows the opposite (they don't necessarily match because the Russian and Erzya sentences may have a different number of words).

Model-based metrics include two cosine similarities of sentence embeddings: LaBSE uses a LaBSE model (Feng et al., 2022) distilled for Erzya[13], and enc_sim uses the mean token embeddings from the encoder of the NLLB-based MT model described in Section 6. The latter model is also used for computing Ppl: the mean cross-entropy loss (perplexity) of the model for translating the Erzya sentence to Russian and in reverse. The Att metric is based on the encoder-decoder attention maps for this model: we average the cross-attentions to each encoder token over the layers and the heads, add up across the decoder tokens, and average across all the encoder tokens, except the first one (language code) and the last one (end of sentence), as they are expected to serve as "attention sinks". This metric is also averaged across two translation directions.

[13] https://huggingface.co/slone/LaBSE-en-ru-myv-v2

Correlations. To evaluate the quality of the metrics, we report their Spearman correlation with the quality annotation labels in the two last rows of Table 3. As we expected, rel_sim and LID_rus are predictive of fluency problems, and all dictionary- and model-based metrics correlate with accuracy.

Data comparison. The top two rows of Table 3 report the average values of our automatic metrics on our dev and devtest splits of FLORES. The next four rows report their values on the baseline data, depending on the presence or absence of fluency and meaning preservation problems. According to most metrics, the FLORES translations are similar or even better than the "good" (problem-free) subset of the baseline data. The only exception is WR_rus which for FLORES is slightly below the baseline; this might be explained by the difficulty of the FLORES domains, where many words are not yet covered by the Erzya dictionaries.

6 A new MT model for Erzya

Our preliminary exploration suggested that few existing models are capable of producing reliable Erzya sentences, so we tried training our own translation model, focused on translation into Erzya.

6.1 Datasets

"Natural" data. To train our model, we used the same monolingual Erzya and parallel Erzya-Russian datasets as Dale (2022). In addition, we collected and aligned some parallel news articles[14], and included several new translated books[15] and pairs of words and phrases from the Russian-Erzya dictionaries at Emerald (Rueter, 2024a,b) and Panlex (Kamholz et al., 2014). The books and articles have been aligned at the sentence level using the

[14] http://e-mordovia.ru

[15] A physics textbook, two books of Alexander Doronin, and a translation of The Little Prince. All copyright holders gave consent to use the texts as training data.

Dataset	rel_sim	len_ratio	LID_rus	WR_rus	WR_myv	LaBSE	enc_sim	Ppl	Att
FLORES dev	0.28	0.90	0.07	0.55	0.64	0.89	0.82	1.49	0.32
FLORES devtest	0.28	0.90	0.06	0.55	0.64	0.89	0.83	1.51	0.32
BL (good)	0.27	0.83	0.06	0.59	0.62	0.86	0.84	2.51	0.28
BL (fluency problems)	0.38	0.87	0.14	0.70	0.74	0.94	0.90	1.73	0.28
BL (meaning problems)	0.23	0.79	0.07	0.43	0.44	0.66	0.69	3.56	0.25
BL (both problems)	0.23	0.90	0.13	0.35	0.37	0.65	0.69	3.64	0.23
BL, corr. with meaning	0.23	0.01	-0.01	0.38	0.38	0.48	0.43	-0.39	0.30
BL, corr. with fluency	-0.30	-0.17	-0.35	-0.17	-0.19	-0.33	-0.26	0.26	0.05

Table 3: The average values of the automatic metrics on our FLORES translation (top 2 rows); their average values on the baseline data grouped by quality (next 4 rows), and their Spearman correlations with human quality labels on the baseline data (the last 2 rows).

algorithm from Dale (2022). We filtered out the pairs for which the Levenshtein distance was less than 20% of the text length, since their inclusion could lead to the model learning to copy the source words too often instead of translating them. We also dropped the pairs with more than 55% difference in length, as they were likely incorrect as translations. The volume of the cleaned Erzya-Russian dataset was 174460 pairs of sentences, phrases or words.[16]

Back-translation. To take advantage of the monolignual Erzya texts, we translated 200K Erzya sentences with the previous version of our model (trained only on "natural" data): 50% into Russian and 50% into 13 other languages previously represented in NLLB-200[17]. After filtering by string edit distances and length ratios, 176462 texts remained. To enhance the model's ability to translate from Erzya into languages other than Russian, we also translated 30K Russian sentences from the Erzya-Russian parallel dataset into the same 13 other languages using an NLLB-200-600M model.

During the training, the pairs of texts from one natural and two synthetic sources were randomly selected in the following proportion: 70% from the natural data (in both directions), 25% from the data translated from Erzya (only in the opposite direction, into Erzya), and 5% from the data translated from Russian to other languages (in both directions). We normalized 100% of the texts on the target side and 80% of the texts on the source side[18] using the algorithm from NLLB Team et al. (2022)[19].

6.2 Model training

When training the model, we followed an approach similar to Tars et al. (2022) and Dale (2022). As a basic model, we took NLLB-200 with 600 million parameters, enriched its dictionary with new tokens for the Erzya language, further trained embeddings for these tokens, and then further trained the entire model on parallel Russian-Erzya data, as well as on data obtained by reverse translation.

Vocabulary update. To better represent Erzya words in the model, we trained a new Sentencepiece tokenizer (Kudo and Richardson, 2018) on the Erzya side of the training dataset. Most of its tokens (6208 out of 8192) were missing from the NLLB vocabulary, so we added them there. The corresponding embeddings of each new token were initialized by the mean of the embeddings of the "old" tokens into which the new token could be decomposed. We also added a new language code to the tokenizer: myv_Cyrl.

Fine-tuning. Since the embeddings of the new tokens were initialized with a naive method, fine-tuning of the whole model with these parameters could introduce undesirable disturbances into other parameters. To avoid this, for the first 45K training steps, we updated only the embedding matrix, "freezing" the rest of the model parameters, and used an additional loss function (with a weight of 100): the mean squared error between the original and current embeddings of the "old" tokens. We used a single GPU, a batch of size 6, 4 gradient accumulation steps, and an Adafactor optimizer (Shazeer and Stern, 2018) with the learning rate that linearly warmed up from zero to 10^{-4} during the first 3000 steps and then stayed constant. After updating the token embeddings this way, we

[16]The collected parallel dataset (at least, its part that is free of copyright restrictions) will be made publicly available in our repository.

[17]Arabic, English, Estonian, Finnish, French, German, Kazakh, Mandarin, Mongolian, Spanish, Turkish, Ukrainian, and Uzbek. This choice was motivated by the international importance of the languages, by their connections to the post-Soviet region where most Erzya live, and by an attempt to represent diverse language families.

[18]We kept 20% of the source texts unnormalized to better prepare the model for potential downstream use cases when the input is not normalized.

[19]The normalization code was adopted from the SentenceSplitClean class in the Stopes repository.

continued training the whole model (with all the parameters unfrozen), for 220K more steps.

7 Comparing MT systems for Erzya

We used our FLORES+ translation (the dev subset) to evaluate the current quality of MT between Erzya on the one side and Russian and English on the other. We evaluated the translation quality with BLEU (Papineni et al., 2002).

7.1 The systems

Here we describe each of the systems that we tried including in our benchmark. For all models except the LLMs, we use beam search with the beam size of 5, and keep all other inference parameters at their default values, unless otherwise specified.

Ours. We used the model described in Section 6. For this model (and for NLLB), we normalized the input texts with the NLLB normalization algorithm.

SLONE (Dale, 2022). We use the myv-mul and mul-myv models from Dale (2022) to translate from and to Erzya, respectively. We use beam size of 5 and repetition penalty of 5, like in the original work (but we do not use reranking with the LID model).

SMUGRI (Yankovskaya et al., 2023). We used their latest model, also based on the NLLB-200 (with 1.3B parameters)[20], for translation in all 4 directions. We used the fairseq-interactive interface with beam size of 5, penalty of 100 for the unk token, and a maximum of 4 consecutive tokens repeating, and ran the model in the fp16 precision.

NLLB. We tried to use the NLLB-200-600M model to translate from Erzya into English and Russian. Since NLLB requires specifying the source language, and Erzya is not included in it, we indicated Estonian as the source language: it is genetically related to Erzya, and also, like Erzya, has experienced some lexical influence of Russian.

We used the MADLAD-400-3B-MT model (Kudugunta et al., 2024) in two configurations: as is (**MADLAD**), and fine-tuned with our Erzya-Russian dataset (without other languages) for 60K steps using a new Erzya token (**MADLAD-ft**). For both of them we used the HuggingFace package with beam size of 3.

[20]The paper mentions M2M-100 as a base model. But the model that we used, https://huggingface.co/tartuNLP/smugri3_14-finno-ugric-nmt, is, apparently, a newer version, and it has the language code formats and the size of NLLB-200-1.3B. According to the model card, it is currently powering https://translate.ut.ee.

System	ru-myv	en-myv	myv-ru	myv-en
NLLB	-	-	3.23	1.05
SLONE	8.11	4.99	15.12	11.70
SMUGRI	11.46	6.58	28.44	**21.12**
MADLAD	1.10	1.06	18.48	13.87
MADLAD-ft	15.50	-	25.50	-
Claude	14.13	7.09	**34.68**	20.18
GPT-4o	3.49	1.24	11.07	8.73
Ours	**17.09**	**7.35**	22.06	16.42

Table 4: BLEU scores for the evaluated MT systems on the dev subset of FLORES+.

Google Translate. We have tried several configurations of the Google Translate API: the general/nmt and general/translation-llm models in the Advanced v3 interface. Both models proved unable to translate from Erzya into Russian or English either when specifying related languages (Estonian, Finnish, Udmurt, Mari) as the source language, or when automatically detecting the language (most often it was defined as Udmurt, Mari, or Komi). In most cases, the models simply transliterated the Erzya text into the Latin alphabet, without trying to translate most of the words. Based on these results, we excluded Google Translate from the benchmark.

Claude (Anthropic, 2024) and **GPT-4**. We used the API of the Claude 3.5 Sonnet and GPT-4o models, respectively, to obtain the translations. For some texts, we obtained the necessary translations only after several iterations of inference without changing the prompt, because in the case of GPT, the prompt did not pass the jailbreak protection, and in the case of Claude, the prompt could produce empty translations, which were corrected upon retranslation. An example of our prompt is given in Appendix A.

7.2 Evaluation results

The results of MT evaluation with BLEU are in Table 4. As we had hoped, our model achieved the best quality of translation into Erzya from Russian and English. The Claude model won the first prize for translating from Erzya to Russian, and the model from Yankovskaya et al. (2023), from Erzya to English. In addition, the MADLAD model demonstrated promising results, with rather good understanding of Erzya even before fine-tuning.

8 Human validation of human and machine translation

While the BLEU scores reported above may help ranking translation systems, they cannot tell how good the translation is from the human point of

view. To shed some light on this, we engaged two new native Erzya speakers (not involved into the FLORES translation) to evaluate the translations of 30 randomly chosen FLORES sentences from Russian into Erzya. We showed them the human translations and machine translations by SLONE, Claude, and our new system, without displaying the system names and in a random order, to reduce the potential bias. The annotators were asked to rate each translation on a 1-5 point scale from Dale (2022), with the label 3 for "acceptable", 4 for "good", and 5 for "great" translations. Their full guidelines are given in Appendix B.

System	Annotator 1	Annotator 2	κ	ρ
Human	4.37(0.14)	4.33(0.12)	-0.03	-0.02
Claude	4.17(0.14)	4.1(0.13)	0.24	0.41
Our MT	3.9(0.19)	3.87(0.17)	0.29	0.54
SMUGRI	3.47(0.21)	3.57(0.21)	0.27	0.53

Table 5: Mean assigned scores (with standard errors in brackets), Cohen's kappa κ and Spearman correlation ρ of the two annotators' labels.

Table 5 reports the mean scores assigned by the annotators to each system, and their agreement scores. The inter-annotator agreement is fair for the MT systems, but there is none for human translations, indicating that more fine-grained annotation schemes might be needed in the future. Nevertheless, both annotators assign the highest (and similar) average scores to the human translations, reaffirming their quality.

Our MT system takes the third place in the human ranking, after the human translations and Claude. The reason for this low position is 3 "stupid" translation errors (out of the 30): two undertranslations and one cyclical hallucination. We hope that in the future, simple modifications of the decoding algorithm would help avoid such errors.

9 Conclusion

Although the *endangered* and *low-resourced* statuses for a language are by no means equivalent (Hämäläinen, 2021), they reinforce each other in a vicious circle. Lack of resources for a language lowers its prestige, which reduces the number of active speakers, which, in turn, disincentivises creation and maintenance of the language resources. As an example, the endangered Erzya language, with its few hundred thousand speakers and a modest Wikipedia community, did not make it to the FLORES-200 dataset and the NLLB-200 models — but if it did, it could have a positive impact at least on the Erzya Wikipedia.

Such projects as the Open Language Data Initiative shared task give one more chance to such languages. By releasing a FLORES+ translation into Erzya and using it to benchmark the few existing MT models that support it, we hope to help rolling the vicious circle in the opposite, virtuous direction. An Erzya version of FLORES might open a way to include this language into other NLP evaluation datasets, such as FLORES-based FLEURS (Conneau et al., 2023) for speech translation and Belebele (Bandarkar et al., 2024) for machine reading comprehension. And the presence of the language in such datasets, we hope, might motivate the researchers to include it in foundation models, which, in turn, might influence the development of practical applications, supporting the speakers of the language and increasing its status and chances of survival.

More directly, we are hoping that the emergency of the Erzya version of FLORES will facilitate improvements in machine translation research and applications for this language.

Some possible areas of future research based on the Erzya translation of FLORES+ include:

- Translating new MT training datasets, such as the Seed (Maillard et al., 2023)), into Erzya.
- Creating an automatic reference-free metric of translation quality for Erzya that would highly correlate with human judgments.
- Setting up a system of active learning that would help collect human translations for the sentences at which MT most likely fails.
- Creating an MT system suitable for translating content (such as books) into the Erzya language, minimizing the necessary revision and post-correction by human translators.

10 Acknowledgements

We express immense gratitude to our translators who accomplished the hardest work of this project and chose to stay anonymous. We are grateful to Jack Rueter for proofreading the translations, to Árpád Váldazs for providing the quality annotations used in Section 5, and to the volunteer annotators for helping us with the human evaluation in the last moment.

References

Khalid Alnajjar, Mika Hämäläinen, and Jack Rueter. 2023. Bootstrapping Moksha-Erzya neural machine translation from rule-based apertium. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 213–218, Tokyo, Japan. Association for Computational Linguistics.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Accessed: 2024-08-19.

Timofey Arkhangelskiy. 2019. Corpora of social media in minority Uralic languages. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 125–140, Tartu, Estonia. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Isaac Caswell. 2024. 110 new languages are coming to google translate. https://blog.google/products/translate/google-translate-new-languages-2024/. Accessed: 2024-08-19.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

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.

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.

Google. 2023. Palm 2 technical report. https://ai.google/static/documents/palm2techreport.pdf. Accessed: 2024-08-19.

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.

Mika Hämäläinen. 2021. Endangered languages are not low-resourced! *arXiv preprint arXiv:2103.09567*.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

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.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36.

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, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2207.04672*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia,

Pennsylvania, USA. Association for Computational Linguistics.

Jack Rueter. 2024a. On searchable mordvin corpora at the language bank of finland, emerald. *Journal of Data Mining & Digital Humanities*, (V. The contribution of corpora).

Jack Rueter and Francis Tyers. 2018. Towards an open-source universal-dependency treebank for Erzya. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 106–118, Helsinki, Finland. Association for Computational Linguistics.

Jack Michael Rueter. 2024b. erzya-bidix.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

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*.

Maali Tars, Taido Purason, and Andre Tättar. 2022. Teaching unseen low-resource languages to large translation models. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 375–380, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

UNESCO. 2010. Unesco atlas of the world's languages in danger (pdf).

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

A Example prompts

We used the following prompt format for translating with Claude: *"You are an AI assistant responsible for translating phrases between Russian and Erzya-Mordvin. Your task is to translate a given sentence from Erzya-Mordvin into Russian. Ensure that the translation remains specific to Erzya-Mordvin, avoiding confusion with Komi, Moksha, or other Finno-Ugric languages to prevent false cognates. Verify that each word in the original Erzya-Mordvin sentence has a corresponding translated word in Russian, maintaining the accuracy and completeness of the content. The final translated sentence should retain a similar word count without omitting any parts of the original text. Output only the final translated result in Russian.*
***{The source sentence}**. Output only the translated result."*. With GPT, we used a similar format, with the first paragraph fed as the system prompt, and the source sentence as the user prompt.

B Annotation guidelines

For human evaluation of human and machine translation in Section 8, we provided the annotators with a short guideline text in Russian. Below is its translation into English.

We ask you to rate the translations from Russian to Erzya on a scale of 1-5 points:

5 points - perfect translation (the meaning and style are fully preserved, the grammar and word choice are correct, the text looks natural);

4 points - good translation (the meaning is fully or almost completely preserved, the style and choice of words are acceptable for the target language);

3 points - acceptable translation (the core meaning is preserved; mistakes in word choice and grammar do not interfere with understanding; most of the text is fluent and in the target language);

2 points - poor translation (the text is mostly understandable and mostly in the target language, but there are serious errors in meaning preservation, grammar or word choice);

1 point - unsuitable translation (most of the text is in the wrong language, or nonsense, or has little in common with the original text).

If at least one word is incorrectly translated, the resulting score should not be 5; the choice between 1 and 4 is at your discretion.

If a word is an overly literary term or a neologism, but its meaning is clear, it does not lower the score. However, if the usage of an unusual word is unclear or it changes the original meaning, lower the score.