Dialectal and Low-Resource Machine Translation for Aromanian

Alexandru-Iulius Jerpelea[‡] **Alina Rădoi**[§] **Sergiu Nisioi**[⋆*]

[‡] Tudor Vianu National College of Computer Science, Bucharest
[§] West University of Timișoara
[⋆] Human Language Technologies Research Center, University of Bucharest
alex.jerpelea@gmail.com, alina.radoi@e-uvt.ro, sergiu.nisioi@unibuc.ro

Abstract

This paper presents the process of building a neural machine translation system with support for English, Romanian, and Aromanian - an endangered Eastern Romance language. The primary contribution of this research is twofold: (1) the creation of the most extensive Aromanian-Romanian parallel corpus to date, consisting of 79,000 sentence pairs, and (2) the development and comparative analysis of several machine translation models optimized for Aromanian. To accomplish this, we introduce a suite of auxiliary tools, including a language-agnostic sentence embedding model for text mining and automated evaluation, complemented by a diacritics conversion system for different writing standards. This research brings contributions to both computational linguistics and language preservation efforts by establishing essential resources for a historically under-resourced language. All datasets, trained models, and associated tools are public:[1] https://arotranslate.com

1 Introduction

Training good machine translation (MT) systems in a low-resource setting is a task far from being solved (Wang et al., 2021; Haddow et al., 2022). Current advances in Large Language Models (LLMs) and evaluation methodologies are centred on English or are massively multilingual, but do not engage with the particularities of low-resource languages. Of the 7000+ languages spoken in the world, only a small part is covered by current MT systems (Costa-jussà et al., 2022).

Aromanian (ISO 639-3 - rup) is an endangered Eastern Romance language (Moseley and Nicolas, 2010), which currently lacks large-scale corpora and electronic resources that can potentially contribute to the preservation of its cultural heritage. In this study, we build a dataset suitable for training MT systems in the low-resource setting between two related languages from the same dialect continuum: Romanian (ron) and Aromanian (ISO 639-3 - rup). Given the close similarity (Caragiu-Marioțeanu, 1975) between the two languages, we expand the existing pre-trained LLMs and machine translation models available for standard Romanian to its Eastern Romance sibling, in order to contribute to the preservation of this endangered language.[2]

The difficulties for such a task stem from the fact that: (1) there is little availability of monolingual or bilingual texts, (2) the writing system, vocabulary, and grammar have not been standardized in a widely accepted institutional manner, several types of spelling are currently being used (Caragiu-Marioțeanu, 1997; Cunia, 1997; Nevaci, 2008), and (3) Aromanian has several varieties / dialects that have been influenced by contact with Greek, Romanian, Turkish, Albanian, and South Slavic languages (Caragiu-Marioțeanu, 1975; Pascaru and Kahl, 2017). Furthermore, the language has historically been transmitted orally within families or small communities (Gica et al., 2009; Maiden, 2016), with public usage limited to a few towns in Albania and North Macedonia.

Our contribution is:

- the creation of a 79k multigenre Aromanian-Romanian sentence-aligned parallel corpus, augmented with machine-translated English sentences;
- a language-agnostic BERT Sentence Encoder (LaBSE) sentence encoder (Feng et al., 2022), fine-tuned for Aromanian support;
- a comparison of different No Language Left Behind (NLLB) (Costa-jussà et al., 2022)

*Corresponding author.
[1] https://huggingface.co/aronlp

[2] https://endangeredlanguages.com/lang/963, Accessed: 2024-09-16

Proceedings of the 31st International Conference on Computational Linguistics, pages 7209–7228
January 19–24, 2025.

models fine-tuned to translate in any direction between Aromanian, Romanian, and English;

- a range of instruction-tuned large language models, fine-tuned for Aromanian translation;

- a diacritics converter between two major orthographic variants of Aromanian.

2 Related Work

Low-resource machine translation has gained considerable traction in the past few years. Initiatives range from commercial ambitions to build "machine translation systems for the next thousand languages" (Bapna et al., 2022), the NLLB family of models that can translate between more than 200 languages (Costa-jussà et al., 2022), to approaches focusing on severely under-resourced languages (Parida et al., 2021; Sánchez-Martínez et al., 2024).

Varieties and closely related languages also receive special attention, from early work on dialectal and language varieties translations exploring statistical machine translation and rule-based systems (Zhang, 1998; Scannell, 2006; Otte and Tyers, 2011; Hassani, 2017) to more recent WMT Shared Tasks (Akhbardeh et al., 2021) covering translations from the same language families, such as Dravidian, Manding, and Romance languages. Although rule-based methods are still a strong baseline for certain language pairs (Sánchez-Martínez et al., 2024), neural approaches provide the state of the art in multiple tasks from Portuguese (Costa-jussà et al., 2018), Serbo-Croatian (Popović et al., 2020), Belorussian-Russian, to Arabic dialects (Karakanta et al., 2018; Kumar et al., 2021).

Aromanian remains an unexplored language in MT research. The electronic resources available for Aromanian consist primarily of multilingual word-aligned lists (Nisioi, 2014; Cristea and Dinu, 2020; Beniamine et al., 2020; Mititelu et al., 2021; Fourrier and Sagot, 2022), which are used to study its evolution, history, contact and relationship with other Romance languages. The recent work of Petrariu and Nisioi (2024) presents a multilingual parallel corpus of approximately 3k sentences covering Aromanian. While it is the largest published resource to date, this corpus only covers one genre, namely fairy tales, and the total number of sentence pairs renders it insufficient for training qualitative MT models, as concluded by Petrariu and Nisioi (2024).

To the best of our knowledge, no successful attempts have been made to integrate Aromanian into a translation system, and our work lays the first building blocks in this direction.

3 Collecting a Multigenre Dataset

The original texts collected in our dataset focus on the Romanian-Aromanian language pair and pertain to different genres such as news articles, literature, dictionaries, religious texts, music lyrics, and essays. Throughout the data collection process, we use a sentence embedder compatible with Aromanian to align sentences across languages. We fine-tune a language-agnostic BERT Sentence Encoder (Feng et al., 2022) using a similar approach to Dale (2022) (see Section 4).

The Bible has been translated by Dina Cuvata and was published in 2004. We use an online scan of the printed edition published by the Aromanian Library hosted by Dini Trandu.[3] Each page of the scan is digitized using Tesseract V5 (Smith, 2007). The engine does not support Aromanian, so the language parameter is set to Romanian. Although the OCR quality is mediocre, it is high enough for us to post-process the results. We then carefully check each verse and pair it with its match from the Romanian translated Bible. Each verse is sentence-split using regular expressions in both languages. If the number of sentences per verse matches, we pair the sentences and add them to our corpus. We obtain around 30.5k sentence pairs from this source.

The Divine Comedy has been translated by Dina Cuvata from Romanian.[4] By aligning it with the Romanian version, using multilingual sentence embeddings described in Section 4, we obtain 2.2k pairs.

Tao Te Ching has been partially translated from English by Mihali Prefti and was published by the Aromanian Library hosted by Dini Trandu.[5] Using automatic sentence alignment based on LaBSE (see Section 4), we obtain 260 sentences.

Lyrics Translate[6] is a platform where songs are translated into multiple languages. We scrape the

[3]`https://dinitrandu.com/wp-content/uploads/2022/04/Bibliea-limba-armaneasca.pdf`, Accessed: 2024-09-16

[4]`https://dinitrandu.com/wp-content/uploads/2018/08/Dante-Dina-Cuvata.pdf`, Accessed: 2024-09-16

[5]`https://dinitrandu.com/wp-content/uploads/2022/06/Cartea-a-Calillei.pdf`, Accessed: 2024-09-16

[6]`https://lyricstranslate.com`, Accessed: 2024-09-16

500+ songs in both Aromanian and Romanian and pair them verse by verse, obtaining 8.5k pairs.

The Multilingual Parallel Corpus of Aromanian (MPC-rup) consists of approximately 3k pairs of sentences from Aromanian fairy tales and short prose texts (Petrariu and Nisioi, 2024).

A list of parallel common phrases and idioms, consisting of 2.1k pairs, provided for this project by members of the Aromanian community.

The Avdhela Project[7] is a digital Aromanian library consisting of a collection of parallel poetry and prose texts. We use only the poems and pair them verse by verse, thus obtaining another 4k pairs for our corpus.

Radio Romania International is a Romanian public news radio station that includes editorials in many languages, among them Aromanian. Most radio broadcasts are published in text form on the official website[8] and Aromanian articles are usually translations of matching Romanian articles.[9] However, there are no backlinks to these pages, making it difficult to trace the corresponding articles. Thus, articles in both languages are scraped and matched by the images they contain. In cases where multiple matches are possible, we use our sentence alignment tool (described in Section 4) to align the titles in Aromanian with those in Romanian, pairing the titles with the greatest semantic similarity while allowing for unmatched titles. Then, for the sentence alignment of each pair of articles, we once again deploy the aforementioned alignment tool. This method yields 18.7k sentence pairs.

The Adventures of Tom Sawyer has been translated from Romanian and the digital version was donated for this project. A total of 3.1k sentences are obtained after automatic alignment.

A century of Aromanian poetry (Cândroveanu and Iorgoveanu, 1985) is a collection of Aromanian poetry translated into Romanian. We use a scanned digitized version and apply Open CV (Bradski, 2000) to identify text bounding boxes (Figure 1). We then apply OCR and the same post-processing steps used for the Bible. Finally, we align the resulting sentences for each page using the alignment tool described in Section 4, adding approximately 1.7k pairs to the corpus.

A collection of modern Aromanian poetry

[7] https://www.proiectavdela.ro Accessed: 2024-09-16

[8] https://www.rri.ro/ro_ar, Accessed: 2024-09-16

[9] A fact confirmed by several authors working for the radio station.

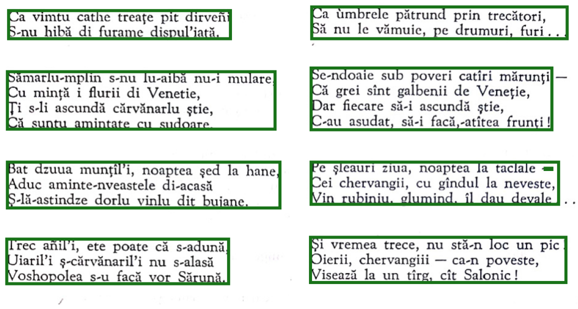

Figure 1: Detecting text bounding boxes in "A century of Aromanian poetry".

written by George Vrana was donated for this project with the author's agreement. Aligning the text verse by verse resulted in 2.1k pairs.

The Little Prince has been translated into Aromanian by Maria Bara and Thede Kahl. We extract 620 sentence pairs using automatic alignment.

Writings is a collection of texts consisting of news articles, personal stories, and essays. These are not available online and were donated for this project by Kira Mantsu, a prominent Aromanian writer. The 16 documents are originally in Aromanian and translated into Romanian by the author herself; the automatic alignment results in a total of 1.9k new pairs.

3.1 Dataset Split

Considering all the sources mentioned above, the dataset we use to build our models consists of 79k sentence pairs. A table with sample sentences from each source is available in Appendix A.

In preparation for the model training process, we split our dataset into *train*, *dev* and *test* splits. Sentences from "The Little Prince" and "Writings" are only included in the *dev* and *test* sets, but are not used for training. The reason behind this choice is that we want to evaluate the model's performance on novel data, and we believe that the literary genre is a difficult evaluation task.

For the *train* set, we extract in a stratified manner 95% random pairs from each of the above sources. The remaining 5% pairs from these sources make up the *dev* set. We also add the sentences from "The Little Prince" to the *dev* set to provide a difficult evaluation set of literary translations. The *test* set comprises only of the mixed-genre "Writings" documents to enforce evaluation on out-of-domain data; thus, the texts do not have a common authorial origin with the data on which our models are trained. Table 1 contains the exact number of

Split	Bible	Lyrics	MPC-rup	Phrases	Avdhela	Radio	Poetry	Tom Sawyer	Divine Comedy	Modern Poetry	Tao Te Ching	The Little Prince	Writings	Total
train	29004	8142	3208	2038	3852	17867	1663	2963	2133	2137	246	-	-	73253
dev	1531	431	169	116	182	880	102	171	110	115	17	623	-	4331
test	-	-	-	-	-	-	-	-	-	-	-	-	1397	1397

Table 1: Number of sentence pairs for each source. The *dev* set contains a random subsample of all the genres and texts from "The Little Prince". The *test* set contains out-of-domain mixed-genre texts.

sentence pairs extracted from each source.

3.2 Synthetic Data

To perform translations between Aromanian and English, we add synthetic English counterparts to all sentence pairs by translating Romanian sentences with the Google Translate API.

To evaluate the quality of automatic English translations, we employ cometkiwi-da model (Rei et al., 2020) for reference-free evaluations; therefore, English translations are evaluated only with regard to the original Romanian sentences. Throughout the corpus, we obtain a *system-score* of 0.765 (scores range from 0 to 1). The resulted score is an indicator that Romanian-to-English machine translations are of mediocre quality and that automatic translations to and from English are error-prone but still meaningful.

3.3 Orthography

It is important to note that there are numerous orthographic standards for Aromanian. The main spelling types in our corpus are *DIARO* (named after the Aromanian-Romanian Dictionary by Caragiu-Marioțeanu (1997), and *Cunia*, named after the author of the Cunia (2010) Dictionary. We also sparsely found Greek orthography, which we standardized using the Latin script.

The *Cunia* and *DIARO* spellings are easily convertible to each other using simple regular expressions, with the exception of the close central and mid central vowels. Both of these sounds are represented by the grapheme <ã> (U+00E3) within the *Cunia* standard. However, the *DIARO* standard follows the Romanian standardization (Petrariu and Nisioi, 2024).[10] using two different symbols <î> (U+00EE) and <â> (U+00E2) for close central and <ă> (U+0103) for mid central.

We release the corpus in both writing standards, but for training purposes the *Cunia* writing system is used. The main reason for this choice is the slightly lower 2.36 fertility rate of the tokenizers versus the *DIARO* orthography that reaches 2.52 tokens per word. We observed lower fertility scores across various models, including both NLLB-200 and pre-trained large language models. Since the pre-trained models are multilingual, it is likely that the tokenizers have seen byte-pairs similar to the *Cunia* orthography in its pretraining data, i.e., words from Albanian and romanised transliterations from South Slavic and Greek languages.

Combating the potential loss of details due to the merging of the mid central and close central vowels, we train an n-gram-based model to convert from *Cunia* to *DIARO*. Table 2 presents several statistics regarding our *Cunia*-converted datasets. Original Aromanian texts have the greatest lexical richness and shorter phrases. However, the Romanian human translations and the machine- translated texts in English have both lower lexical diversity (as estimated by the type-token ratio) and larger sentence lengths. In Appendix B, we also present these statistics for each text genre separately.

4 Aromanian Sentence Embeddings

To embed Aromanian sentences in the same latent space as Romanian and English sentences a language-agnostic BERT Sentence Encoder (Feng et al., 2022) is fine-tuned following the methodology described by Dale (2022).

[10] The Romanian standardization remains inconsistently adopted, particularly in online texts. The primary variation is the alternation between <â> (used mid-word) and <î> (used at the beginning and end of words) to represent the close central vowel. The introduction of the grapheme <â> by the Romanian Academy in 1993 as an *anticommunist* measure has been widely criticized by linguists for its lack of scientific and etymological justification (Dumistrăcel, 1993). Consequently, certain Romanian publications and publishers allow their authors to choose their preferred standard rather than enforcing this rule.

language	words	unique words	type-token ratio	words/sentence
Aromanian	1,189,000	175,000	0.15	15.31
Romanian	1,279,000	114,000	0.09	16.45
English	1,371,000	68,000	0.05	17.64

Table 2: Dataset statistics of texts converted to the *Cunia* standard. Words are extracted using a regex tokenizer.

First, the tokenizer's vocabulary is trained on a monolingual Aromanian corpus, using all Aromanian sentences prior to alignment. This is performed with the same WordPiece tokenizer from BERT (Devlin et al., 2019), yielding 4,400 new tokens. As a result, the tokens-per-word ratio of the LaBSE tokenizer decreases from 2.36 to 1.77.

Secondly, the model is fine-tuned on both Aromanian-Romanian and Aromanian-synthetic English parallel pairs, updating only the pooled output corresponding to Aromanian embeddings using contrastive loss.

At each training step, a batch of sentence pairs is randomly selected from one of the two pairs of languages. The dot product of the embeddings for all possible sentence pairs in the batch is computed, rewarding only the pairs that are correct matches. To prevent overfitting, a small margin of 0.3 is subtracted from the reward for matching pairs. Training stops after 150k steps, as both the loss and accuracy graphs (for matching translated sentences) flatten beyond this point (Figure 2). In a batch of random sentence pairs, the model pairs each Aromanian sentence with its corresponding translation with an accuracy of over 98%. Details on the training hyper-parameters are provided in Appendix C.

The model is used in two ways: (1) to calculate the BERTScore (Zhang et al., 2020) and evaluate trained machine translation models (see Section 6.1); and (2) to mine and align sentences in parallel documents. Similarly to Dale (2022), dynamic programming is used to select a sequence of sentence pairs that have the highest possible sum of similarity scores.

5 Machine Translation Models

5.1 GPT Baseline

Inspired by recent MT results (Hendy et al., 2023; Kocmi et al., 2024) where closed-source systems achieved state-of-the-art performance on machine translation, we employ the GPT-4o model, in hopes of leveraging its extensive multilingual pre-training (Achiam et al., 2024). Additionally, it is likely that

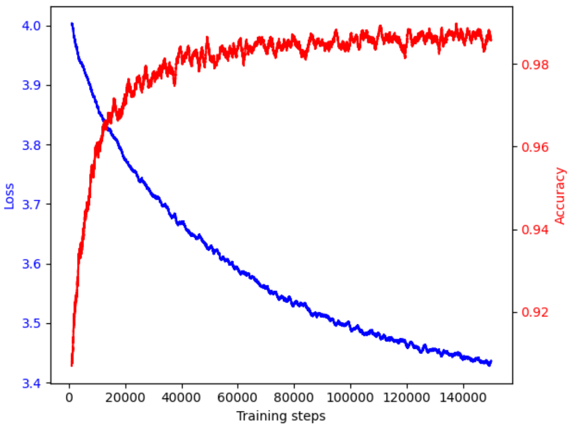

Figure 2: Average loss and accuracy during the LaBSE fine-tuning.

the model has been exposed to Aromanian in its pre-training stage. We use OpenAI's *gpt-4o* model in zero-shot mode to translate all pairs in the *test* split, which we later evaluate in Section 6.

5.2 NLLB Machine Translation

Seeking to benefit from transfer learning (Khiu et al., 2024), we fine-tune NLLB-200 (Costa-jussà et al., 2022), an encoder-decoder transformer architecture model that can translate between any of its 202 languages. The goal is to obtain a model that has the ability to translate between Aromanian, Romanian, and English. The last two are already supported by the NLLB.

The NLLB tokenizer uses language tags, i.e. special tokens are added to the source and target texts, which are employed by the model in the pre-training phase to identify the source and target languages. The special tokens for Romanian and English are <*ron_Latn*> and <*eng_Latn*>, respectively. We expand the tokenizer vocabulary with the <*rup_Latn*> token for Aromanian. Its meaning is supposedly constructed during the fine-tuning step. The embedding value of the newly added token is set to the mean of the <*ron_Latn*> and <*ell_Grek*> (Greek) embedding vectors. This is justified by the close relationship of Aromanian with standard Romanian and the large proportion

System prompt	You are a helpful assistant capable of accurately translating between Aromanian, Romanian, and English.
Instruction	Translate this sentence from Aromanian to English
Input	Shi-lji intrã Enoh tu vrearea-al Dumnidzã sh-ma multu nu s-aflã cã-l mutã Dumnidza.
Output	And God pleased Enoch, and then he was no more, because God moved him.

Table 3: Prompt format example for finetuning the LLaMA 3.1 Instruct model.

model	ron→rup	rup→ron	eng→rup	rup→eng	eng→ron	ron→eng
GPT-4o	19.0	37.2	17.6	36.7	**65.7**	**79.2**
Qwen2 7B Instruct	45.8	50.5	38.8	51.9	58.5	76.2
TowerInstruct v0.2	45.7	50.9	31.2	44.6	50.8	75.2
LLaMA 3.1 8B Instruct	46.3	51.4	38.8	52.6	60.8	77.2
RoLLaMA	46.0	50.6	39.7	52.2	60.2	77.3
NLLB 1.3B	47.8	**53.2**	43.6	**54.7**	63.2	76.3
NLLB 600M	**49.1**	52.7	**44.5**	53.8	62.2	75.5

Table 4: ChrF++ evaluation scores of different models on the *test* set. Zero-shot GPT-4o is unable to produce translations into Aromanian. NLLB models consistently outperform LLMs on translations from and to Aromanian. No significant differences can be observed between RoLLaMA and other models that have not been previously trained on Romanian instruction data. GPT-4o achieves the best scores on Romanian and English, although the English sentences are machine translated from Romanian.

of Greek influence and loanwords (Pascaru and Kahl, 2017; Dragomirescu, 2020).

We experiment with the 1.3B and 600M distilled models of the original Mixture-of-Experts 54B billion parameter model. Each model is trained for 100k steps, where each sentence pair in the batch is in a random direction between Aromanian, Romanian, and English (i.e. six possible directions). At every 10k steps, we save a checkpoint and evaluate the *dev* set. The NLLB-600M checkpoint at 90k training steps, and the NLLB-1.3B checkpoint at 70k training steps produce the highest evaluation scores on the *dev* split. The hyper-parameters used for training the two models are available in Appendix D.

5.3 LLMs for Machine Translation

LLMs have proven to be the state of the art for general machine translation, from zero-shot or few-shot translation (Kocmi et al., 2024) to different fine-tuning strategies(Alves et al., 2024).

We perform full fine-tuning on four base LLMs trained for instruction following:

- LLaMA 3.1 8B Instruct (Grattafiori et al., 2024) – a multilingual model from Meta AI in which Romanian is incidentally covered in the 15% of multilingual tokens used during pre-training
- Qwen2 7B Instruct (Yang et al., 2024) - a multilingual model from Alibaba Cloud covering as many as 30 languages, including 4 well-resourced Western and Italo-Romance languages: French, Spanish, Portuguese, Italian
- RoLLaMA 3 8B Instruct (Masala et al., 2024) – a LLaMA3-based model fine-tuned specifically to respond to tasks and instructions in Romanian, hoping to facilitate knowledge transfer to Aromanian
- TowerInstruct 7B v0.2 (Alves et al., 2024) – a LLaMA2-based model trained specifically to solve translation tasks (document-level translation, terminology-level translation, etc.) on 10 languages, including 4 well-resourced Western and Italo-Romance languages: French, Spanish, Portuguese, Italian

Except for the RoLLaMA model, which has been specifically fine-tuned for Romanian language tasks, all the other models support Romanian only incidentally.

Similarly to NLLB, we have the same objective of translating between Aromanian, Romanian, and English, so we train the LLMs with samples from all six possible directions. In the case of the four models, fine-tuning for more than 1 epoch leads to

Participant	Direction	Fluency		Style		Meaning	
		HT	MT	HT	MT	HT	MT
Subject 1	ron→rup	9.66	9.73	9.60	9.60	9.73	9.86
	rup→ron	9.46	9.66	9.46	9.53	9.46	9.33
Subject 2	ron→rup	9.66	9.66	9.33	9.60	9.66	9.73
	rup→ron	8.13	9.0	7.73	8.73	8.2	8.93

Table 5: Average direct assessment scores for machine translated texts (MT) and human translations (HT).

an increased loss on the *dev* split. Thus, we keep only the checkpoints fine-tuned for 1 epoch with sample packing (Krell et al., 2021). An example of a prompt for the LLaMA 3.1 model can be found in Table 3. More details about each model's training hyper-parameters and prompt format are available in Appendix E.

6 Evaluation

6.1 Automatic Evaluation

We evaluate all the models in all possible directions between Aromanian, Romanian, and English. We note here that the English references are machine-translated and the results involving English should be taken with reservation. However, we provide the evaluation scores here for completeness. The automatic metrics that we employ are BLEU (Post, 2018), ChrF++ (Popović, 2015), and BERTScore (Zhang et al., 2020). We report the ChrF++ scores for all the models in Table 4. All other results are available in Appendix F and Appendix G

Overall, for English-Romanian pairs, GPT-4o obtains the strongest scores in both directions, even when translating from noisy synthetic English into Romanian. Chances that the model has been previously exposed to the out-of-domain test set are very small, since the *Writings* are not available online. Given the synthetic nature of the English data, we cannot draw any conclusions with respect to translation quality into English.

More importantly, the scores for the Aromanian-Romanian language pairs obtained by any of the trained models are considerably higher than those obtained by GPT-4o. Among the trained models, the NLLB family consistently obtains higher scores than their fine-tuned instruction LLM counterparts, regardless of the automatic metric (see Table 4 and Appendix F).

When translating into Aromanian (→rup), the differences are noticeable – in every single case NLLB-type models perform better than LLM-based translation models. Still, when translating into a well-resourced language such as English or Romanian, the performance gap narrows, with the models differing by only a few points. In-domain evaluation on the *dev* set (in Appendix G) shows significantly higher scores than out-of-domain evaluation on *test* set, with $+14$ points in ChrF and double the BLEU scores. On the *dev* set, the differences between NLLB models and LLMs are less pronounced.

The evaluation scores do not indicate that Aromanian translation benefits more from using RoLLaMA, a model trained on Romanian instructions, or TowerInstruct v0.2, a model designed specifically for MT tasks.

All types of models struggle to generate high-quality Aromanian output, as evidenced by the overall scores. The differences range from 5 to 15 points, with translations into Aromanian scoring lower than translations from Aromanian. Appendix F provides additional metrics that align with our findings using ChrF++. We note that Costa-jussà et al. (2022) found that a $+1.0$ increase in ChrF++ is almost always noticeable by human evaluators.

Manual inspection of translations reveals that the models tend to translate sentences word-for-word, which misses word overlaps in BLEU-based metrics for out-of-domain evaluation on mixed-genre literary texts, news, and essays. More details on manual evaluation are presented in the next section.

6.2 Human Evaluation

Two human evaluators participated in this study. Both are native Aromanian speakers from Romania, fluent in both Romanian and Aromanian. They self-identify as speakers of the Aromanian Grămustean and Cipan dialects, classified as Type A in the typology of Caragiu-Marioțeanu (1997) and Pascaru and Kahl (2017), which is also the main Aromanian variety in the dataset.

We do not evaluate for English here. The evaluated sentences contain human and machine translations in equal proportions, and therefore each annotator evaluates a total of 80 samples. For MT, we use our NLLB 600M fine-tuned model.

Annotators are instructed to assign direct assessment scores (Graham et al., 2013) from 1 to 10 to evaluate the quality of each translation (1 being the lowest score, 10 the highest). This scale is culturally motivated by the fact that it is familiar to the study participants because it is used in Romanian schools at all levels of post-primary school.

They assign a score taking into account three categories: fluency, style, and meaning (logical sense). Fluency refers to how grammatically correct the sentence is. Style denotes how likely the speaker is to build the sentence they are evaluating in the exact same way. Meaning alludes to the logical sense of the sentence.

The evaluation was carried out bilingually for all language pairs. The results are in Table 5, showing that machine translation is rated slightly better than human translation. For each low grade, we ask annotators to provide comments with a description of errors. Lexical errors predominate: annotators identify words that do not exist or are not common in their language variety in both human and machine translated texts, e.g., usage of *bus* - instead of *aftuchină* (En. *bus*); usage of *sala di conțertu* (En. *concert hall*); *inițiativă* (En. *initiative*) are not recognized as valid Aromanian.

Evaluating machine translation systems for non-standard language varieties presents several challenges regarding representativeness and meaningfulness: 1) annotators may produce inconsistent results, as they may not be familiar with the full range of Aromanian varieties (Pascaru and Kahl, 2017); 2) machine translation systems generate neologisms derived from Romanian, even when established Aromanian terms exist, particularly when translating contemporary texts or news content; 3) human translations may receive lower ratings from speakers due to dialectal differences.

Therefore, the results of human evaluation for fine-grained quality assessment remain inconclusive. Nonetheless, the fact that 80% of machine translations into and from Aromanian received a perfect score from human annotators suggests that the MT system has certain strengths.

7 Online Translation System

To make the machine translation model accessible to the public and raise awareness of the endangered status of the Aromanian language, we have deployed the system online. The user interface is designed to be lightweight and includes accessibility features such as copying text content and switching between the source and target languages (Figure 3).

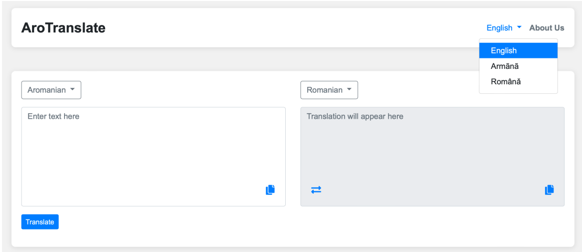

Figure 3: AroTranslate GUI.

7.1 Quantization

The fine-tuned NLLB-600M model demonstrates the highest performance and has the smallest number of parameters, making it suitable for deployment on a CPU. We quantize the model using the ctranslate2 (Klein et al., 2020) engine to 8-bit integer weights. Furthermore, the engine applies additional optimizations, including layer fusion, padding removal, batch reordering, and more.

Regarding the performance loss typically associated with quantization, we observe minimal differences in the automatic metrics used. As shown in Table 6, which analyzes the Aromanian-Romanian direction on the *test* split, the differences are measured at less than 1 BLEU point. Since the models are fine-tuned for sentence-level translations, the system splits user input into sentences and processes them in parallel. The system achieves a translation speed of approximately 65 tokens per second.

7.2 Orthographic Converter

The system generates translations using the *Cunia* orthography for Aromanian. Additionally, we also implement the option to convert the output to the linguistically-motivated *DIARO* standard. The only ambiguous letter in Cunia standard is <ã> because it can represent either the close central vowel [ɨ] (e.g. cãndu / when / /kɨnˈdw/) or the mid central vowel [ə] (e.g., [ə] / tricurã / passed by /triˈkuˈrə/; ãncã / yet / /ɨnˈkə/).

For this special case, we employ a two-fold solution: (1) word normalization, where the word is normalized and then matched with the most frequent replacement for that specific word form from

Metric	ron→rup BLEU	ron→rup ChrF++	rup→ron BLEU	rup→ron ChrF++
FP32	17.3	45.03	30.95	51.01
INT8 quantization	17.01	44.84	30.21	50.16

Table 6: Differences between the quantized (INT8) and floating point 32 (FP32) versions of the fine-tuned NLLB 600M model, measured on the *test* split.

a dictionary, and (2) a character 4-gram based solution where for each <ã> in each word we analyse the two neighbouring letters to the right and left and construct frequency masks for these 4-grams. During inference, after replacing all other diacritics, the system selects each 4-gram and chooses the most frequent of its two possible replacements. If neither approach succeeds, the default replacement is biased towards the mid central vowel [ə] based on phonological suggestions from native speakers. The accuracy of this solution is roughly 96%.

8 Conclusions

Building a dialectal machine translation system for a low-resource language comes with several challenges – from the lack of standardization and small online presence of the language to the diverse varieties specific to different countries.

Previous experiments show that modern neural machine translation systems are unable to learn from extremely small data sets of 3,000 pair of sentences, even for languages that are part of the same dialectal continuum such as Romanian and Aromanian (Petrariu and Nisioi, 2024). In our work, we provide the steps to construct a parallel corpus for Aromanian, which further requires training a multilingual sentence embedder. We show that a diverse parallel dataset of 79k examples is sufficient for fine-tuning pre-trained transformer models to do low-resource and dialectal machine translation. The preexisting knowledge of Romanian in both LLM transformers and sequence-to-sequence models enhances generalization and knowledge transfer capabilities. However, it also introduces bias and raises concerns about representativeness, as the generated Aromanian shows strong Romanian influence and similarities.

The introduction of English synthetic data in the training phase proves useful to extend the translation directions beyond Romanian. However, future work would require measuring the impact of translation errors in synthetic data.

Based on automatic evaluation, the smaller seq-to-seq models are better translators than LLM-based English-centric or multilingual fine-tuned models (that have not been pre-trained on Romanian). NLLB-based models outperform the LLM-based pre-trained models in terms of automatic metrics and have the advantage of being light-weight and easy to deploy on CPU using int8 quantization.

Regarding human evaluation, due to the relatively small size of the community and lack of access to a large pool of native speakers, we had difficulties conducting an extensive human evaluation. Furthermore, we identify several difficulties in conducting human evaluations on low-resource, unstandardized dialectal varieties, and future studies are mandatory to quantify the limitations of our MT system geographically and across language varieties.

With this work, we hope to bring Aromanian into the spotlight of machine translation and corpus-based linguistic research and contribute to preservation efforts for this endangered language.

9 Limitations

Aromanian is a very low-resource language and is not standardized, having several varieties. As a result, translation quality is experimental and often times sub-par to what a user might expect from a translation system for medium- to high-resource languages. This model is biased to favour some Aromanian varieties more than others. Translation errors may occur when certain information is missing or altered. There is also concern about the model's bias, as the Bible is a dominant source in our corpus. Last but not least, this work already mentions a limitation in terms of lack of availability to carry an in-depth speaker-centric human evaluation of the MT system. This is an issue that we are looking into in future work. We release the corpus under the Creative Commons Attribution-NonCommercial 4.0 International License [11].

[11] https://creativecommons.org/licenses/by-nc/4.0/

10 Ethical Considerations

We were able to collect and compile parallel translated texts from multiple sources. Note that for each text source we have received permission to use the text, verified appropriate licensing agreements, or confirmed fair-use applicability for academic research purposes. The Google Translate API calls cost around 180$ in credits and the OpenAI API calls cost around 10$. Model training was performed on cloud-based servers (A100 and RTX 4090 GPUs) rented by the hour, incurring a total cost of approximately 500$.

Acknowledgments

This work would not have been possible without the support of the Aromanian Community from Romania (CARo), who helped us with the data collection processes, system testing, and gave us invaluable advice. We specifically acknowledge Florentina Costea, Cristian Jiru, and Nicolae Todică. Regarding data collection, we are grateful to Kira Mantsu and George Vrana for all their help. The human evaluation process would not have been possible without the generous help of Georgiana Stavrositu, Sebastian Florea, and Zoia Dragomir. The machine translation system is hosted with the generous support of the University of Bucharest https://arotranslate.unibuc.ro.

This research is partially supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906 and by InstRead: Research Instruments for the Text Complexity, Simplification and Readability Assessment CNCS - UEFISCDI project number PN-IV-P2-2.1-TE-2023-2007.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, and others from OpenAI. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An Open Multilingual Large Language Model for Translation-Related Tasks. In *First Conference on Language Modeling*.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building Machine Translation Systems for the Next Thousand Languages. *Preprint*, arXiv:2205.03983.

Sacha Beniamine, Martin Maiden, and Erich R. Round. 2020. Opening the Romance Verbal Inflection Dataset 2.0: A CLDF lexicon. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 3027–3035. European Language Resources Association.

G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

H. Cândroveanu and K. Iorgoveanu. 1985. *Un veac de poezie aromână [A Century of Aromanian Poetry]*. Ed. Cartea Românească.

Matilda Caragiu-Marioțeanu. 1975. *Compendiu de dialectologie română [Compendium of Romanian Dialectology]*. Editura Științifică și Enciclopedică [Scientific and Encyclopedic Press].

Matilda Caragiu-Marioțeanu. 1997. *Dicționar aromân (macedo-vlah) DIARO. A-D. Comparativ (român literar - aromân) [Aromanian (Macedo-Vlach) Dictionary. Letters A-D. Comparative dictionary standard Romanian - Aromanian]*. Editura Enciclopedică [Encyclopedic Press].

Marta R. Costa-jussà, Marcos Zampieri, and Santanu Pal. 2018. A neural approach to language variety translation. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 275–282, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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. *Preprint*, arXiv:2207.04672.

Alina Maria Cristea and Liviu P. Dinu. 2020. Automatic Identification and Production of Related Words for Historical Linguistics. *Computational Linguistics*, 45(4):667–704.

Tiberiu Cunia. 1997. *On the Standardization of the Aromanian System of Writing*. Bitola Symposium.

Tiberiu Cunia. 2010. *Dictsiunar a limbãljei armãneascã [Dictionary of Aromanian Language]*. Editura Cartea Aromãnã [Aromanian Book Press].

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Adina Dragomirescu. 2020. *Balkan-Romance*. Oxford University Press.

Stelian Dumistrăcel. 1993. *Lupta în jurul literei â și demnitatea Academiei Romane*. Iași.

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.

Clémentine Fourrier and Benoît Sagot. 2022. Probing multilingual cognate prediction models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3786–3801, Dublin, Ireland. Association for Computational Linguistics.

Alexandru Gica et al. 2009. The Recent History of the Aromanians from Romania. *New Europe College Yearbook*, (09):173–200.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, and Aston Zhang et al. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Hossein Hassani. 2017. Kurdish interdialect machine translation. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial)*, pages 63–72.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv preprint arXiv:2302.09210*.

Alina Karakanta, Jon Dehdari, and Josef van Genabith. 2018. Neural machine translation for low-resource languages without parallel corpora. *Mach. Transl.*, 32(1-2):167–189.

Eric Khiu, Hasti Toossi, Jinyu Liu, Jiaxu Li, David Anugraha, Juan Flores, Leandro Roman, A Seza Doğruöz, and En-Shiun Lee. 2024. Predicting machine translation performance on low-resource languages: The role of domain similarity. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1474–1486.

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.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Mario Michael Krell, Matej Kosec, Sergio P Perez, and Andrew Fitzgibbon. 2021. Efficient sequence packing without cross-contamination: Accelerating large language models without impacting performance. *arXiv preprint arXiv:2107.02027*.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 110–121, Online. Association for Computational Linguistics.

Martin Maiden. 2016. Romanian, Istro-Romanian, Megleno-Romanian, and Aromanian. In *The Oxford guide to the Romance languages*, pages 91–125. Oxford University Press.

Mihai Masala, Denis Ilie-Ablachim, Alexandru Dima, Dragos Georgian Corlatescu, Miruna-Andreea Zavelca, Ovio Olaru, Simina-Maria Terian, Andrei Terian, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024. “vorbești românește?” a recipe to train powerful Romanian LLMs with English instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11632–11647, Miami, Florida, USA. Association for Computational Linguistics.

Verginica Barbu Mititelu, Mihaela Cristescu, and Manuela Nevaci. 2021. Un instrument modern de studiu al dialectului aromân: corpus adnotat morfosintactic [a modern study tool for the aromanian dialect: a morpho-syntactically adnotated corpus]. *Ex Oriente lux. In honorem Nicolae Saramandu*, pages 143–162.

Christopher Moseley and Alexandre Nicolas. 2010. *Atlas of the World's Languages in Danger*, 3rd ed. entirely revised, enlarged and updated edition. Memory of Peoples. UNESCO, Intangible Cultural Heritage Section, Paris.

Manuela Nevaci. 2008. Sisteme de scriere utilizate în limba publicațiilor aromânești actuale. *Revista „Diacronia”*, XXVII:29.

Sergiu Nisioi. 2014. On the syllabic structures of aromanian. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH@EACL 2014, April 26, 2014, Gothenburg, Sweden*, pages 110–118. The Association for Computer Linguistics.

Pim Otte and Francis Tyers. 2011. Rapid rule-based machine translation between dutch and afrikaans. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*.

Tache Papahagi. 1974. *Dictionarul dialectului aromân. General și etimologic. [The Dictionary of Aromanian Dialect. General and Etymological]*. Editura Academiei Române [Romanian Academy Press].

Shantipriya Parida, Subhadarshi Panda, Amulya Dash, Esaú Villatoro-Tello, A Seza Doğruöz, Rosa M Ortega-Mendoza, Amadeo Hernández, Yashvardhan Sharma, and Petr Motlicek. 2021. Open machine translation for low resource south american languages (americasnlp 2021 shared task contribution). In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 218–223.

Andreea Pascaru and Thede Kahl. 2017. Rrāmānj und armānj. wie einheitlich sind kultur, sprache und identität der aromunen? *Quo vadis Romania*, 50:51–77.

Iulia Petrariu and Sergiu Nisioi. 2024. A multilingual parallel corpus for aromanian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 832–838.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Maja Popović, Alberto Poncelas, Marija Brkic, and Andy Way. 2020. Neural machine translation for translating into Croatian and Serbian. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 102–113, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Felipe Sánchez-Martínez, Juan Antonio Perez-Ortiz, Aaron Galiano Jimenez, and Antoni Oliver. 2024. Findings of the WMT 2024 shared task translation into low-resource languages of Spain: Blending rule-based and neural systems. In *Proceedings of the Ninth Conference on Machine Translation*, pages 684–698, Miami, Florida, USA. Association for Computational Linguistics.

Kevin P Scannell. 2006. Machine translation for closely related language pairs. In *Proceedings of the Workshop Strategies for developing machine translation for minority languages*, pages 103–109. Citeseer.

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

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4636–4643. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xiaoheng Zhang. 1998. Dialect mt: a case study between cantonese and mandarin. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

A Aromanian-Romanian Sample Pairs

Source	Aromanian	Romanian
Bible	Și vasiljelu Ptolometi muri după trei dzali, shiashchirladzlji cari eara pit tsitãts lj-vatamara oaminjlji di-a loclui. Ti-agiutarea-a tsilor far di mindueari-arauda unã giudicata cu itsralji, a tinirlui om ti-unã cãnushteari shi bunã mindueari.	Și regele Ptolomeu a murit a treia zi, și ostașii care erau în cetăți au fost uciși de către locuitori. Pentru a prilejui celor fără gând rău o judecată isteață, omului tânăr cunoștință și bună cugetare.
Lyrics T.	Pisti giuguri trecu anarga, lãi niori Nicuchirã cu curunã:	Peste juguri trec încet, nori negri Ar trimite-o el acasă,
MPC-rup	Unã caprã mushcrã shi unã ghesucãnutã pãshteau di unã parte, shi chipurle lã asunau: cing-cing. Tsachilji giuca deavãrliga, arsãrea - tuts aleptsã, musheats, cu lãna lor ca mãtasea, di-lji yinea shi-al Tegã se-arsarã shi s-lu stringã-n bratsã picurarlu shi s-lu bashe, ahtare harauã lu-avea loatã!	O capră cu botul bălțat și alta roșcată - pășteau răzlețite și tot le sunau tălăngile de la gât: cing-cing-cling. Țapii se zbenguiau prin jur, țupăiau - frumoși toți, dragii, cu părul lor mătăsos, de-i venea lui Tegă, de bucurie, să-l ia în brațe pe cioban, să-l strângă la piept și să-l sărute, nestăpânit!
Phrases	nu shtiu si mpartu palji la doi yumari misuru stealili	fi prost hoinări
Avdhela	Di toamnã-ascuturatã Ma, s-ti caftu nji-easti fricã,	Se scutură de toamnă, Să te cer îmi este teamă,
Radio	Deputatlu USR Cristian Seidler spusi cã proiectul di nom ari un impact multu modestu ti hãrgili bugetare cu pensiile spetsiali.	Deputatul USR Cristian Seidler a afirmat că proiectul de lege are un impact foarte modest asupra cheltuielilor bugetare cu pensiile speciale.
Aromanian Poetry	Am un alt livend, di-Avdelă, alăvdatlu-atsel di NUShI, te da njeate cãt sh-la-aushi; Nã shcãmbã fãrã suflit i semn di tine, tută shi salta siminatã di tine, tu ubor...	În Avdela am un june tULLIU NUȘI - vorbă cerească versu-n veci o să-i trăiască ! O piatră doar aminte de tine mai aduce, și-o salcie sădită de mâna ta, cândva...
Tom Sawyer	Arucã funea! Mutri anvãrliga, canã... tu niheam di oarã u avea cartea tru mãnã.	Dă și parâma! Aruncă o privire de jur-împrejur, nu era nimeni... în clipa următoare, ținea cartea în mâini.
Divine Comedy	Shi-agãrsheashti-n frixea tsi-lji da dãgoarea tu doilji sherchi-mpriunats trupeashti	pârjolul să-și ia ceva, căci la copil ia seamă în cei doi șerpi împreunați
Modern poetry	mashi suflitlu nu ari moarti aclo iu suntu farurli apãryisiti.	doar sufletul este fără moarte acolo unde sunt farurile abandonate.
Tao Te Ching	Pots s-ti tradz ditu minduirea a ta Imnji după ea sh-nu-ari bitisitã.	Poți să te dai înapoi din propria ta minte urmează-l și nu are sfârșit.
The Little Prince	Gioaca cor cathi gioi cu featili di-n hoara. Elj avdu daima mashi alavdarli.	Se duc să joace, joia, cu fetele din sat. Vanitoșii nu aud niciodată decât laudele.
Writings	Li-alidzea ayonjia, unã dupu altã, ashi cum beai unã scafã cu apã! Bãnarã deadunu tu Vãryãrie.	Le citea repede, una după alta, așa cum ai bea un pahar cu apă! Trăiseră împreună în Bulgaria.

Table 7: Aromanian-Romanian sample pairs from various sources. The texts have been converted in the *Cunia* standard.

Our final data release also contains 27.3k dictionary pairs that we do not use to train the models. When training our initial machine translation models, we did not observe any improvements when using word-aligned dictionary entries. The entries are extracted from Papahagi's Aromanian dictionary (Papahagi, 1974).

Source	Aromanian	Romanian
Bible	Și vasil'elu Ptolometi muri dupâ trei dzali, și-așchirladzl'i cari eara pit țităț l'-vatamara oamińl'i di -a loclui. Ti-agiutarea-a ților far di mindueari-arauda unâ giudicata cu ițral'i, a tinirlui om ti-unâ cânușteari și bunâ mindueari.	Și regele Ptolomeu a murit a treia zi, și ostașii care erau în cetăți au fost uciși de către locuitori. Pentru a prilejui celor fără gând rău o judecată isteață, omului tânăr cunoștință și bună cugetare.
Lyrics T.	Pisti giuguri trecu anarga, lâi niori nicuchirâ cu curunâ:	Peste juguri trec încet, nori negri Ar trimite-o el acasă,
MPC-rup	Unâ caprâ mușcră și unâ ghesucănută pășteau di unâ parte, și chipurle lâ asunau: cing-cing. Țachil'i giuca deavârliga, arsărea-tuț aleptâ, mușeaț, cu lâna lor ca mătasea, di-l'i yinea și-al Tegâ se-arsară și s-lu stringă-n brațâ picurarlu și s-lu bașe, ahtare harauâ lu-avea loatâ!	O capră cu botul bălțat și alta roșcată - pășteau răzlețite și tot le sunau tălăngile de la gât: cing-cing-cling. Țapii se zbenguiau prin jur, țupăiau - frumoși toți, dragii, cu părul lor mătăsos, de-i venea lui Tegă, de bucurie, să-l ia în brațe pe cioban, să-l strângă la piept și să-l sărute, nestăpânit!
Phrases	nu știu si mpartu pal'i la doi yumari misuru stealili	fi prost hoinări
Avdhela	Di toamnă-ascuturată Ma, s-ti caftu ńi-easti fricâ,	Se scutură de toamnă, Să te cer îmi este teamă,
Radio	Deputatlu USR Cristian Seidler spusi că proiectul di nom ari un impact multu modestu ti hârgili bugetare cu pensiile spețiali.	Deputatul USR Cristian Seidler a afirmat că proiectul de lege are un impact foarte modest asupra cheltuielilor bugetare cu pensiile speciale.
Aromanian Poetry	Am un alt livend, di-Avdelă, alâvdatlu-ațel di NUȘI, te da ńeate cât ș-la-auși; nâ șcâmbâ fârâ suflit i semn di tine, tutâ și salta siminată di tine, tu ubor...	În Avdela am un june tULLIU NUȘI - vorbă cerească versu-n veci o să-i trăiască ! O piatră doar aminte de tine mai aduce, și-o salcie sădită de mâna ta, cândva...
Tom Sawyer	Arucâ funea! Mutri anvârliga, canâ... tu niheam di oarâ u avea cartea tru mânâ.	Dă și parâma! Aruncă o privire de jur-împrejur, nu era nimeni... în clipa următoare, ținea cartea în mâini.
Divine Comedy	Și-agârșeaști-n frixea ți-l'i da dăgoarea tu doil'i șerchi-mpriunaț trupeaști	pârjolul să-și ia ceva, căci la copil ia seamă în cei doi șerpi împreunați
Modern poetry	mași suflitlu nu ari moarti aclo iu suntu farurli apâryisiti.	doar sufletul este fără moarte acolo unde sunt farurile abandonate.
Tao Te Ching	Poț s-ti tradz ditu minduirea a ta Imńi dupâ ea ș-nu-ari bitisitâ.	Poți să te dai înapoi din propria ta minte urmează-l și nu are sfârșit.
The Little Prince	Gioaca cor cathi gioi cu featili di-n hoara. El' avdu daima mași alavdarli.	Se duc să joace, joia, cu fetele din sat. Vanitoșii nu aud niciodată decât laudele.
Writings	Li-alidzea ayońia, unâ dupu altâ, ași cum beai unâ scafâ cu apâ! Bânarâ deadunu tu vâryârie.	Le citea repede, una după alta, așa cum ai bea un pahar cu apă! Trăiseră împreună în Bulgaria.

Table 8: Aromanian-Romanian sample pairs from various sources, converted to an approximate form of the *DIARO* standard.

B Corpus Statistics

language	source	words	unique words	words/sentence	type-token ratio
Aromanian	Bible	493018	69780	16.15	0.14
	Lyrics T.	16039	7586	3.98	0.47
	MPC-rup	508090	74419	27.1	0.14
	Phrases	40941	9972	4.78	0.24
	Avdhela	7839	3567	3.48	0.45
	Radio	41036	12708	12.15	0.30
	Aromanian Poetry	36649	10433	11.69	0.28
	Tom Sawyer	23177	10325	13.13	0.44
	Divine Comedy	1269	747	4.83	0.58
	Modern poetry	11897	6336	5.3	0.53
	Tao Te Ching	5653	1886	2.62	0.33
	The little prince	3619	1625	5.81	0.44
	Writings Collection	17893	6646	12.81	0.37
Romanian	Bible	580288	42487	19.0	0.07
	Lyrics T.	16923	7022	4.2	0.41
	MPC-rup	494098	53549	26.36	0.10
	Phrases	43590	7484	5.08	0.17
	Avdhela	8186	3330	3.63	0.40
	Radio	47588	12291	14.09	0.25
	Aromanian Poetry	38719	11126	12.35	0.28
	Tom Sawyer	24665	9296	13.97	0.37
	Divine Comedy	1397	744	5.31	0.53
	Modern poetry	14146	6412	6.31	0.45
	Tao Te Ching	4877	1740	2.26	0.35
	The little prince	3977	1717	6.38	0.43
	Writings Collection	19328	6398	13.84	0.33
English	Bible	622430	28266	20.38	0.04
	Lyrics T.	20277	5067	5.03	0.24
	MPC-rup	518118	34576	27.64	0.06
	Phrases	49025	4950	5.72	0.10
	Avdhela	9466	2671	4.2	0.28
	Radio	53286	8348	15.78	0.15
	Aromanian Poetry	41069	7692	13.1	0.18
	Tom Sawyer	29411	6491	16.67	0.22
	Divine Comedy	1457	675	5.54	0.46
	Modern poetry	16959	4892	7.56	0.28
	Tao Te Ching	5089	1597	2.36	0.31
	The little prince	4258	1380	6.83	0.32
	Writings Collection	20998	4795	15.03	0.22

Table 9: Corpus statistics for each source, computed on the version converted to *Cunia* standard.

C LaBSE Training Hyper-paramateres

Training steps	150000
Batch size	8
Margin	0.3
Optimizer	Adafactor (Shazeer and Stern, 2018)
Learning rate	1e-5
Clip threshold	1.0

Table 10: LaBSE training hyperparameters, Appendix C.

D NLLB Training Hyper-parameters

Training steps	100000
Batch size	8
Optimizer	Adafactor (Shazeer and Stern, 2018)
Scheduler	constant
Learning rate	1e-4
Weight Decay	1e-3
Clip threshold	1.0
Maximum Sequence Length	512

Table 11: NLLB training hyperparameters for both the 600M and 1.3B distilled versions.

E LLM Training

E.1 Hyper-parameters

Effective batch size	32
Number of epochs	4
Learning rate	2e-5
LR scheduler	cosine
Warmup steps	100
Weight decay	0.001
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	1e-8
Maximum Sequence Length	512

Table 12: Training hyper-parameters for all the fine-tuned LLMs.

E.2 Prompt Format

We format the prompts of the LLaMA 3.1 8B Instruct and RoLLaMA 3 8B Instruct models using the standard LLaMA 3.1 Instruct prompt format (Grattafiori et al., 2024). For the Qwen and TowerInstruct models, we use their specific instruction prompt template (Yang et al., 2024; Alves et al., 2024).

System	<|start_header_id|>system<|end_header_id|> You are a helpful assistant capable of accurately translating between Aromanian, Romanian, and English. <|eot_id|>
User	<|start_header_id|>user<|end_header_id|> Translate this sentence from Romanian to Aromanian Adu-ți aminte că moartea nu zăbovește și hotărârea morții nu ți s-a arătat.<|eot_id|>
Model	<|start_header_id|>assistant<|end_header_id|> Adu-ts aminti cã moartea nu-amãnã shi-apofasea-a moartiljei nu tsã si-ari spusã.

Table 13: Prompt format for LLaMA 3.1 8B Instruct and RoLLaMA 3 8B Instruct. Note that this is an example, and the source and target language are not fixed, and do include *English* as well.

System	<|im_start|>system You are a helpful assistant capable of accurately translating between Aromanian, Romanian, and English.<|im_end|>
User	<|im_start|>user Translate this sentence from Romanian to Aromanian Adu-ți aminte că moartea nu zăbovește și hotărârea morții nu ți s-a arătat.<|im_end|>
Model	<|im_start|>assistant Adu-ts aminti cã moartea nu-amãnã shi-apofasea-a moartiljei nu tsã si-ari spusã.<|im_end|>

Table 14: Prompt format for Qwen2 7B Instruct. Note that this is an example, and the source and target language are not fixed, and do include *English* as well.

User	<|im_start|>user Translate this sentence from Romanian to Aromanian Adu-ți aminte că moartea nu zăbovește și hotărârea morții nu ți s-a arătat.<|im_end|>
Model	<|im_start|>assistant Adu-ts aminti cã moartea nu-amãnã shi-apofasea-a moartiljei nu tsã si-ari spusã.<|im_end|>

Table 15: Prompt format for TowerInstruct 7B v0.2. Note that this is an example, and the source and target language are not fixed, and do include *English* as well.

F Automatic Evaluation Results - *test* Set

model	ron→rup	rup→ron	eng→rup	rup→eng	eng→ron	ron→eng
GPT-4o	3.5	19.4	2.9	19.0	46.5	66.7
Qwen2 7B Instruct	15.2	28.8	10.4	32.4	35.4	63.0
TowerInstruct v0.2	16.3	29.5	7.3	23.7	27.1	62.1
LLaMA 3.1 8B Instruct	15.9	30.3	10.8	33.7	38.6	65.1
RoLLaMA	14.8	29.7	11.1	32.8	38.1	64.4
NLLB 1.3B	16.3	31.9	12.8	35.7	41.4	63.0
NLLB 600M	17.0	30.9	13.5	35.0	39.1	62.2

Table 16: BLEU evaluation scores on the *test* set align with the ChrF++ results from Table 4. NLLB models consistently outperform LLMs for Aromanian translations in both directions, with a notable discrepancy between translations into and from Aromanian.

model	ron→rup	rup→ron	eng→rup	rup→eng	eng→ron	ron→eng
GPT-4o	0.866	0.816	0.857	0.829	0.968	0.980
Qwen2 7B Instruct	0.905	0.896	0.880	0.903	0.955	0.976
TowerInstruct v0.2	0.904	0.894	0.840	0.870	0.923	0.973
LLaMA 3.1 8B Instruct	0.908	0.899	0.888	0.907	0.959	0.978
RoLLaMA	0.908	0.898	0.887	0.904	0.957	0.978
NLLB 1.3B	0.913	0.905	0.904	0.913	0.964	0.977
NLLB 600M	0.914	0.905	0.904	0.913	0.964	0.976

Table 17: BERTScore evaluation on the *test* set.

G Automatic Evaluation Results - *dev* Set

model	ron→rup	rup→ron	eng→rup	rup→eng	eng→ron	ron→eng
Qwen2 7B Instruct	31.9	50.7	23.9	51.7	49.8	67.8
TowerInstruct v0.2	34.5	52.6	26.4	53.1	52.1	70.5
LLaMA 3.1 8B Instruct	34.8	53.6	26.6	52.4	52.4	69.5
RoLLaMA	34.4	53.0	26.3	51.4	52.4	68.9
NLLB 1.3B	33.7	53.2	26.2	52.7	52.1	68.5
NLLB 600M	33.6	52.0	25.6	52.2	50.5	67.3

Table 18: BLEU evaluation scores on the *dev* set.

model	ron→rup	rup→ron	eng→rup	rup→eng	eng→ron	ron→eng
Qwen2 7B Instruct	61.9	68.1	55.6	67.6	70.1	80.3
TowerInstruct v0.2	63.3	69.1	56.9	68.8	71.9	82.1
LLaMA 3.1 8B Instruct	63.7	69.8	57.2	68.5	72.0	81.6
RoLLaMA	63.7	69.4	57.2	68.1	71.9	81.1
NLLB 1.3B	63.7	69.9	57.7	68.4	72.2	80.8
NLLB 600M	63.6	69.2	57.4	67.9	71.1	80.0

Table 19: ChrF++ evaluation scores on the *dev* set.

Compared to the out-of-domain *test* set, the evaluation scores on the *dev* set are considerably higher and the differences between LLMs and NLLB-type models is smaller.

ChrF signature is `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.4.3` and BLEU signature is `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.3`.