

Expanding Chechen Machine Translation Resources: SMOL, BOUQuET, and WMT24++ Extensions

Abu-Viskhan A. Umishov
abuviskhanumishov@gmail.com

Vladislav A. Grigorian
real.vladislav.grigorian@gmail.com

Abstract

We introduce a new fork of the SMOL training dataset along with forks of the BOUQuET and WMT24++ translation benchmarks for the Chechen language. We train a model on SMOL and calculate BLEU and chrF++ scores on the introduced benchmarks to evaluate how training on such a limited amount of data may contribute to the model’s Chechen-Russian translation capability. The introduced resources provide a unified evaluation framework for Chechen machine translation and allow measuring the transfer of improvements across different benchmark datasets. Our experiments establish baseline results for future research on Chechen translation and demonstrate the potential of utilizing small-scale parallel corpora for improving low-resource language models.

1 Introduction

In recent years, the Chechen language has been rapidly digitized: new translation models supporting Chechen are emerging, Chechen corpora are being published (Jones et al., 2023; Penedo et al., 2025), and the Chechen language is appearing in closed-source products such as Google Translate¹. Chechen is a Northeast Caucasian (Nakh) language spoken primarily in the Chechen Republic of the Russian Federation and by diaspora communities. Despite having more than one million speakers (Institute of Linguistics, Russian Academy of Sciences, 2024), it remains severely underrepresented in natural language processing resources compared to widely studied languages. Its rich morphology and limited availability of parallel corpora create additional challenges for machine translation.

However, parallel datasets for this language remain few compared to other European languages, and the Chechen language remains "low-resource". Furthermore, no generally accepted benchmark for

assessing translation quality has yet been published for this language.

In our work, we contribute to addressing these limitations. We prepare professional translations of the SMOL parallel training dataset (Caswell et al., 2025) along with the BOUQuET (Team et al., 2025) and WMT24++ (Deutsch et al., 2025) translation benchmarks for Chechen. We also fine-tune the MADLAD-400-3B-MT model (Kudugunta et al., 2023) on the SMOL smoldoc split and measure BLEU and chrF++ changes after training to evaluate how a limited amount of parallel data can improve a model’s Chechen-Russian translation capabilities.

The resulting Chechen versions of SMOL, BOUQuET, and WMT24++ are publicly released on their corresponding Hugging Face dataset pages. We hope that these resources will facilitate future research on Chechen machine translation and provide a common basis for evaluating progress in this low-resource language.

2 Related Work

To our knowledge, there are three widely recognized evaluation benchmark datasets for translation tasks: FLORES+ (Team et al., 2022), WMT24++ (Deutsch et al., 2025), and BOUQuET (Team et al., 2025). Among these benchmarks, we chose WMT24++ and BOUQuET for this work for several reasons: both datasets have Russian versions that could be used by our translators as references, they cover various text domains, and their texts were considered convenient for translation by our translators.

Since the survey provided in our previous paper (Umishov and Grigorian, 2025), new monolingual and parallel datasets for the Chechen language have emerged: (Penedo et al., 2025, 2026). We examined the FineTranslations² parallel dataset containing Chechen-English sentence pairs with human eval-

¹<https://blog.google/products/translate/google-translate-new-languages-2024/>

²<https://huggingface.co/datasets/HuggingFaceFW/finetranslations>

uation by Chechen native speakers and concluded that it cannot be used for training due to the poor quality of its parallel pairs. Another relevant resource is the SMOL dataset (Caswell et al., 2025), a professionally translated multilingual parallel training dataset designed for machine translation model development. Its Russian version³ curated by Ali Kuzhuget and Isaac Caswell made it a suitable source dataset for creating a Chechen translation, as it could provide translators with an additional reference during the translation process.

3 Datasets

We present Chechen versions of three multilingual datasets: SMOL, BOUQuET, and WMT24++. The Chechen versions of the datasets were prepared by Marem S. Magomaeva, a philologist specializing in the Chechen language and translation. The Russian versions of the original datasets were used as references during the translation process. A subset of the translated sentences was additionally inspected by one of the authors, Abu-Viskhan A. Umishov, a native speaker of Chechen, as a quality check.

The SMOL dataset (Caswell et al., 2025) was selected as the training corpus due to the availability of a Russian version, which was used as an additional reference during translation. We translated the smoldoc split into Chechen, creating a Chechen-Russian parallel training dataset. The resulting dataset contains 114 document pairs (863 sentence pairs) and was used for fine-tuning experiments.

During the preparation of the Chechen version of SMOL, we used only the Cyrillic Letter Palochka (U+04C0) character. At that time, the conventions for representing this letter in digital text were not fully established: major smartphone keyboard providers included only the uppercase form in their Chechen layouts, despite the grammatical use of both uppercase and lowercase forms. After the publication of the Chechen SMOL version, the convention of following the grammatical distinction between Cyrillic Letter Palochka and Cyrillic Small Letter Palochka (U+04CF) became preferable. The benchmark datasets prepared in this work follow the grammatically correct usage of both forms. The Chechen version of SMOL will be updated in a future release to follow the current orthographic

convention.

For evaluation, we translated the BOUQuET benchmark (Team et al., 2025). BOUQuET was selected because it covers diverse text domains and provides a Russian version that could be used by translators as a reference. The resulting Chechen version contains 504 sentence pairs in the dev split and 854 sentence pairs in the test split, which are used as validation and test benchmarks, respectively.

The WMT24++ benchmark (Deutsch et al., 2025) was translated into Chechen to provide an additional evaluation resource for measuring translation quality. Similarly to BOUQuET, WMT24++ was selected due to the availability of a Russian version, which could be used by translators as an additional reference, and its coverage of diverse text domains. The resulting Chechen version contains 998 sentence pairs.

The translation guidelines used for preparing all Chechen datasets, including orthographic and consistency requirements, are provided in Appendix A. After the translation was completed, we post-processed the datasets to normalize the orthographic representation of the Chechen Cyrillic Letter Palochka. During the translation process, the Digit One symbol was temporarily used by translators for convenience due to its visual similarity and easier availability on standard keyboards. Before releasing the datasets, all such occurrences were replaced with the appropriate Cyrillic Letter Palochka character.

For the SMOL dataset, we replaced Digit One with Cyrillic Letter Palochka; the corresponding code is provided in Appendix E. For BOUQuET and WMT24++, we used a two-stage procedure to correctly initialize Palochka characters. First, we replaced Digit One with either Cyrillic Letter Palochka or Cyrillic Small Letter Palochka based on its position in the Chechen sentence. The corresponding code is provided in Appendix F. We then manually checked the sentences for proper names beginning with Cyrillic Small Letter Palochka and replaced it with Cyrillic Letter Palochka where appropriate.

4 Experiments

We fine-tuned the MADLAD-400-3B-MT model (Kudugunta et al., 2023) on the Chechen version of the SMOL dataset using Low-Rank Adaptation (LoRA). For translation direction specification, we

³<https://huggingface.co/datasets/Agisight/google-smol-en-ru>

used the special tokens $\langle 2ce \rangle$ and $\langle 2ru \rangle$ for Russian-to-Chechen (RU $\rightarrow$ CE) and Chechen-to-Russian (CE $\rightarrow$ RU) translation, respectively. The LoRA configuration and training hyperparameters are summarized in Tables 2 and 3. We used a low LoRA rank ($r = 4$) with dropout regularization and applied adapters to all linear layers.

Due to the limited size of the training data, overfitting was a primary challenge during fine-tuning. To mitigate this issue, we used small microbatches during training. The increased stochasticity of gradient updates reduces the risk of memorizing individual training examples and improves generalization to unseen evaluation datasets. Additionally, we applied early stopping with a patience of four consecutive evaluations and a minimum improvement threshold of 0.001 based on validation loss. The best-performing checkpoint was selected according to the validation loss.

We used the PEFT library (Mangrulkar et al., 2022) for LoRA implementation and the Transformers framework (Wolf et al., 2020) for organizing the training process. The models were trained on one NVIDIA V100 GPU using BF16 mixed precision. We later found that V100 does not provide native BF16 hardware support, although the training completed successfully with BF16 enabled in the Transformers/PyTorch environment. For detailed LoRA and training configurations, see Appendix B and Appendix C.

The Chechen versions of BOUQuET and WMT24++ prepared in this work were used for evaluation. The inference parameters for the model we used are described in Appendix D. We report results for both translation directions using BLEU and chrF++ metrics, computed with SacreBLEU (Post, 2018). BLEU was computed using the default SacreBLEU configuration, while chrF++ was computed with `word_order=2`. The BOUQuET dev split was used as a validation set during training, while all test splits were reserved for final evaluation. For all splits, we used the dataset versions released in our [Hugging Face repository](#) without further preprocessing.

We report the fine-tuned model scores and the absolute improvements over the base MADLAD-400-3B-MT model.

5 Results

Table 1 presents the translation performance of the base MADLAD-400-3B-MT model and the model

Benchmark	Direction	BLEU	chrF++
BOUQuET	CE $\rightarrow$ RU	0.18 (+0.18)	13.41 (+13.19)
	RU $\rightarrow$ CE	0.02 (+0.02)	4.59 (+4.35)
WMT24++	CE $\rightarrow$ RU	0.23 (+0.23)	9.98 (+9.67)
	RU $\rightarrow$ CE	0.02 (+0.02)	5.66 (+5.40)

Table 1: Translation performance of MADLAD-400 fine-tuned on SMOL. Values in parentheses denote the improvement over the base model.

fine-tuned on the Chechen version of SMOL. Values in parentheses show the absolute improvement over the base model. Fine-tuning consistently improves both BLEU and chrF++ scores across all evaluated directions and benchmarks, demonstrating that even a relatively small amount of professionally translated data can substantially improve the model’s ability to process Chechen.

The largest improvements are observed for the Chechen-to-Russian direction. On the BOUQuET benchmark, fine-tuning increases chrF++ from 0.22 to 13.41 and BLEU from approximately 0 to 0.18, while on WMT24++ chrF++ increases by 9.67 points. The improvements in the reverse Russian-to-Chechen direction are smaller, with chrF++ gains of 4.35 and 5.40 points on BOUQuET and WMT24++, respectively. This asymmetry can be explained by the fact that MADLAD-400 already had stronger generation capabilities for the resource-rich Russian language, while acquiring the ability to process and translate Chechen required additional language-specific adaptation.

6 Conclusion

In this work, we introduced Chechen versions of the SMOL training dataset and the BOUQuET and WMT24++ evaluation benchmarks. We evaluated the impact of fine-tuning MADLAD-400-3B-MT on the prepared SMOL dataset and showed that even a limited amount of professionally translated parallel data can substantially improve Chechen translation performance.

The obtained results demonstrate consistent improvements over the base model across both evaluated benchmarks and translation directions, with the largest gains observed for Chechen-to-Russian translation. This indicates that language-specific adaptation can significantly improve the model’s ability to handle a low-resource language, while generating Chechen remains more challenging due to the limited prior exposure of the model to this language.

The released datasets provide additional resources for future research on Chechen machine translation and evaluation, helping to improve the coverage of low-resource languages in multilingual translation systems.

7 Acknowledgements

We thank Maren S. Magomaeva for preparing the Chechen translations and for her linguistic expertise.

We also thank David Dale for his advice on this article and for his broader efforts to support the Russian-speaking low-resource language community.

References

- Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Djibrila Diane, Solo Farabado Cissé, Koulako Moussa Doumbouya, Edoardo Ferrante, Alessandro Guasconi, Christopher Homan, Mamadou K. Keita, Sudhamoy DebBarma, Ali Kuzhuget, David Anugraha, and 7 others. 2025. [Smol: Professionally translated parallel data for 115 under-represented languages](#). *Preprint*, arXiv:2502.12301.
- Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. [Wmt24++: Expanding the language coverage of wmt24 to 55 languages dialects](#). *Preprint*, arXiv:2502.12404.
- Institute of Linguistics, Russian Academy of Sciences. 2024. Chechen language. <https://minlang.iling-ran.ru/lang/chechenskiy-yazyk>. Minority Languages of Russia project, accessed: 2026-08-07.
- Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. [GATITOS: Using a new multilingual lexicon for low-resource machine translation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.
- Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. [Madlad-400: A multilingual and document-level large audited dataset](#). *Preprint*, arXiv:2309.04662.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. [PEFT: State-of-the-art parameter-efficient fine-tuning methods](#). <https://github.com/huggingface/peft>.
- Guilherme Penedo, Hynek Kydlíček, Amir Hossein Kargaran, and Leandro von Werra. 2026. [Finetranslations](#). <https://huggingface.co/datasets/HuggingFaceFW/finetranslations>.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. [Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language](#). *Preprint*, arXiv:2506.20920.
- 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, Belgium, Brussels. Association for Computational Linguistics.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- The Omnilingual MT Team, Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Cynthia Gao, Jean Maillard, Alex Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenco, Albert Ventayol-Boada, and Shireen Yates. 2025. [Bouquet: dataset, benchmark and open initiative for universal quality evaluation in translation](#). *Preprint*, arXiv:2502.04314.
- Abu-Viskhan A. Umishov and Vladislav A. Grigorian. 2025. [The first open machine translation system for the chechen language](#). *Preprint*, arXiv:2507.12672.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Translation guidelines (English version)

1. The sentences must be translated accurately and sound natural in Chechen.
2. Consider the context of the other sentences in the paragraph when translating each sentence.
3. If you have difficulty with a sentence, skip it.
4. Use dictionaries if needed, but do not rely on Google Translate or any other machine translation systems.
5. Use 1 (Digit One) for Cyrillic Letter Palochka and Cyrillic Small Letter Palochka.
6. Maintain the capitalization of the original text. Example: "КОТ - ЦИЦИГ", "КОТ - ЦИЦИГ".
7. For Chechen words, use only Cyrillic characters (and 1 for both cases of Palochka). An example of what not to do: "парень - к1ант", here letter k is Latin.
8. Preserve the original untranslatable parts of the text, such as emojis, IDs and words that were not translated in the Russian reference, exactly.

B LoRA configuration parameters

Parameter	Value
task_type	"SEQ_2_SEQ_LM"
r	4
lora_alpha	8
lora_dropout	0.1
bias	"none"
target_modules	"all-linear"
use_rslora	True
layers_to_transform	None

Table 2: Parameters of the LoRA Configuration

C Training Configuration Parameters

Parameter	Value
learning_rate	1e-4
optim	"adamw_torch_fused"
per_device_train_batch_size	4
per_device_eval_batch_size	16
gradient_accumulation_steps	1
gradient_checkpointing	False
max_grad_norm	1.0
warmup_steps	200
weight_decay	0.01
lr_scheduler_type	"cosine"
num_train_epochs	16
fp16	False
bf16	True
bf16_full_eval	True
save_strategy	"steps"
save_steps	200
eval_strategy	"steps"
eval_steps	200
save_total_limit	3
predict_with_generate	True
generation_max_length	256
dataloader_pin_memory	True
load_best_model_at_end	True
metric_for_best_model	"eval_loss"
greater_is_better	False

Table 3: Training Configuration Parameters

D Inference parameters

Parameter	Value
num_beams	4
do_sample	False
max_length	256

Table 4: Inference Parameters, all those ones that were not mentioned in this table were set to default values

E SMOL Palochka character initialization

```
import re

PALOCHKA = "\u04c0\u04cf"
PALOCHKA_RE = re.compile(
    rf"(?<=[а-яА-ЯёЁ]{PALOCHKA})1|1(?=[а-яА-ЯёЁ]{PALOCHKA})"
)

def init_palochka(text: str) -> str:
    return PALOCHKA_RE.sub("\u04c0", text)
```

F BOUQuET and WMT24++ Palochka character initialization

```
import re

PALOCHKA = "\u04c0\u04cf"
SMALL_PALOCHKA_RE = re.compile(
    rf"(?<=[а-яА-ЯёЁ]{PALOCHKA})1|1(?=[а-яА-ЯёЁ]{PALOCHKA})"
)
CAPITAL_PALOCHKA_RE = re.compile(
    rf"^[1](?=[а-яА-ЯёЁ]{PALOCHKA})"
)

def init_small_palochka(text: str) -> str:
    return SMALL_PALOCHKA_RE.sub("\u04cf", text)

def init_capital_palochka(text: str) -> str:
    return CAPITAL_PALOCHKA_RE.sub("\u04c0", text)

def init_palochka(text: str) -> str:
    return init_small_palochka(init_capital_palochka(text))
```