The OSCAR PUBLISHING Services
American Journal of Philological Sciences

VOLUME Vol.06 Issue03 2026
PAGE NO. 13-17
DOI 10.37547/ajps/Volume06Issue03-03

Fine Tuning Strategies for Uzbek And Russian Neural Machine Translation

Avezov Sukhrob Sobirovich
PhD, Lecturer at the Department of Russian Language and Literature, Bukhara State University, Bukhara, Uzbekistan

Received: 22 January 2026; **Accepted:** 18 February 2026; **Published:** 11 March 2026

Abstract: The article examines which fine tuning strategy is most rational for Uzbek and Russian neural machine translation under low resource conditions and domain variation. The analysis compares six model lines, MarianMT, mBART 50, M2M100, NLLB 200, uzT5, and SeamlessM4T, through a structured set of linguistic diagnostics that includes agglutination, analytic verb forms, terminology, named entities, script variation, and discourse dependent ambiguity. The study is not presented as a large scale benchmark with new numerical scores. Instead, it offers an evidence based analytical comparison grounded in Russian language scholarship, open academic reports, and an author designed stress test for Uzbek and Russian. The main result is that no single architecture is optimal for every scenario. NLLB 200 and mBART 50 appear most promising as core backbones for supervised adaptation, MarianMT remains useful in narrow institutional domains, uzT5 is valuable as an Uzbek aware auxiliary model, and SeamlessM4T is strategically relevant for future speech to speech pipelines rather than as the default pure text solution.

Keywords: Uzbek language, Russian language, neural machine translation, fine tuning, low resource languages, multilingual transfer, parallel corpora, evaluation metrics, terminology consistency.

Introduction: Machine translation has returned, in a new technical form, to an old linguistic problem. The decisive variable is still predictability of data. Dyomochkina, Gruzdev, and Lukyanova formulate it with unusual precision, “the lower the entropy, the higher the predictability” [1]. In the Uzbek and Russian pair, this proposition remains highly relevant because the pair combines structural asymmetry, domain scarcity, and unstable terminology layers in educational, legal, and literary texts. Recent Russian language research has also shifted the center of gravity from abstract adequacy to workflow design. Kamshilova and Belyaeva write that “Special attention is paid to post-editing of MT products” [2], and this shift matters because Uzbek and Russian translation quality depends not only on the backbone model, but also on editing depth, corpus curation, and task format. A scientific abstract, a subtitle line, and a literary sentence stress the same architecture differently. So do short sentences and long syntactic periods. The problem is therefore architectural and philological at once.

For this language pair, corpus quality is not a secondary issue. Abdurakhmanova and Abzhalova remind us that “The parallel corpus is a bilingual corpus” [3], while Ergasheva, Khaitkulov, and Kuchimova stress the “the purpose and functions of the corpus and parallel corpus” [4]. Once Uzbek and Russian data are aligned at sentence level, the researcher gains access not simply to equivalents, but to recurrent asymmetries in tense, aspect, derivation, politeness, and lexical packing. These asymmetries become the real terrain of fine tuning.

METHODS

The study uses a structured analytical design. Its source base consists only of Russian language academic publications, conference papers, official research reports, and project documentation devoted to machine translation, corpus building, evaluation metrics, Uzbek morphology, and parameter efficient adaptation. The comparison is therefore literature

driven, yet it is organized around a practical research question, namely which model family gives the best return when one needs to fine tune for Uzbek and Russian in scientific, educational, official, and partly literary domains.

The model set includes MarianMT, mBART 50, M2M100, NLLB 200, uzT5, and SeamlessM4T. Lavrentyeva and Kogan explicitly list “mT5, mBART, M2M100 and SeamlessM4T” and also note that MarianMT is a popular framework for training and using MT models. NLLB is included because a current Russian project on low resource Turkic translation selected it as the core architecture for multilingual training and subsequent pair specific adaptation, including Russian and Uzbek. uzT5 is not a classical bilingual MT backbone, but it is highly relevant as an Uzbek aware auxiliary model for data cleaning, generation, compression, and repair.

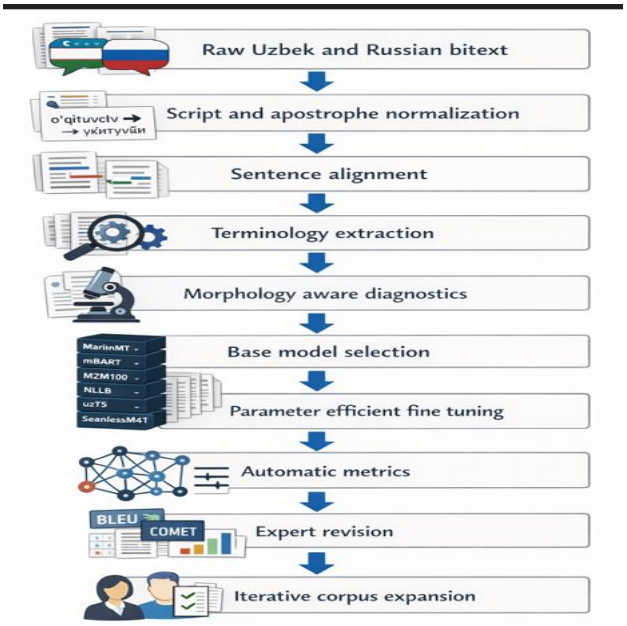

Table 1. Comparative Characteristics of MT Models for Uzbek and Russian

Model	Role in the comparison	Main advantage for Uzbek and Russian	Main limitation	Most rational fine tuning scenario
MarianMT	Compact bilingual or narrow multilingual baseline	Low inference cost, controllable domain adaptation	Weaker broad transfer when pair data are scarce	Institutional corpora, legal templates, educational prose
mBART 50	Multilingual encoder decoder backbone	Strong supervised adaptation, good fit for Russian centered setups	Rare Uzbek morphology may be smoothed	Curated medium size parallel corpus with PEFT
M2M100	Broad many to many baseline	Useful multilingual warm start, flexible directionality	Terminology drift, heavier decoding	General domain preadaptation before domain fine tuning
NLLB 200	Translation oriented multilingual backbone	Best transfer logic for low resource and related Turkic data	Requires aggressive corpus cleaning and normalization	Cross pair training plus pair specific adaptation
uzT5	Uzbek adapted text to text auxiliary model	Better Uzbek lexical and morphotactic sensitivity	Not a dedicated translation default	Synthetic data generation, normalization, terminology repair
SeamlessM4T	Multimodal translation family	Strong strategic value for ASR MT TTS pipelines	Excessive for lightweight pure text scenarios	Subtitle and speech translation ecosystems

Table 2. Diagnostic Phenomena for Uzbek and Russian MT Fine Tuning

Diagnostic phenomenon	Example	Main translation risk	Priority in fine tuning
Agglutination	kitoblaringizdan	Loss of possessive and ablative meaning	Very high
Analytic predicate	qaror qabul qilmoq	Literal segmentation instead of verbal unit	Very high
Evidential nuance	u kelgan ekan	Omission of evidential colouring	High
Scientific noun phrase	til birliklarining funksional semantik tavsifi	Terminological dispersion	Very high
Named entity	Buxoro davlat universiteti	Inconsistent transliteration or calque	High
Script variation	oʻqituvchi and укитувчи	Orthographic mismatch in training data	Very high
Russian agreement chain	сложные синтаксические конструкции научного текста	Morphological flattening on Uzbek side	High
Idiomatic compression	koʻngliga tugmoq	Overliteral rendering	Medium
Official formula	Mazkur shartnoma bir yil davomida amal qiladi	Template instability	Medium
Literary perception	U deraza oldida turib, manzaraga qaradi	Style loss and clause reordering	Medium

RESULTS

The compact baseline, MarianMT, is still valuable when the objective is narrow domain control rather than maximal cross lingual coverage. In official or repetitive prose, for instance in contract clauses such as Mazkur shartnoma bir yil davomida amal qiladi, a smaller pair oriented system can often be tuned to behave more consistently than a very large multilingual model. Its weakness emerges when the corpus is thin and lexical variety rises, especially in academic prose where one lemma appears in several derivational forms and where Uzbek nominal chains must be resolved into Russian syntactic groups. MarianMT therefore remains a rational starting point for institutional deployment, but not the best single backbone for a dissertation level Uzbek and Russian MT program.

mBART 50 occupies a more balanced position. A Russian report on prefix tuning states that adaptation was implemented “for the mBART model, since it supports the Russian language”, and the same report shows that prefix tuning is particularly useful when data are limited. For Uzbek and Russian this matters greatly, because a medium sized but well aligned corpus often exists only in fragments, textbooks, subtitles, small administrative corpora, or manually cleaned literary passages. mBART 50 becomes especially persuasive when the research design includes parameter efficient adaptation, terminology lists, and human review after each training cycle. It is not the cheapest option. It is, however, one of the most research friendly. M2M100 is broader and less controlled. Its value lies in multilingual pretraining and direction flexibility. This helps when one wants to inject related Turkic material, or when Uzbek and Russian data need support from neighbouring language directions during preadaptation. Yet this same breadth can become a liability in scientific writing, where a phrase such as til birliklarining funksional semantik tavsifi should not drift toward a semantically close but terminologically weaker Russian paraphrase. M2M100 is thus stronger as a warm start than as a final domain model.

NLLB 200 appears most promising where the central problem is low resource transfer rather than minimal inference cost. The RSF project on low resource translation reports that Russian and Uzbek corpora were assembled within a larger Turkic oriented program and that NLLB was selected as the main architecture for the project. The same document describes multilingual training followed by pair specific adaptation as a concrete implementation of transfer learning between related languages. For Uzbek and Russian this is a decisive argument. The pair benefits not only from bilingual data, but from structured contact with related agglutinative systems and from

data augmentation on top of unified morphologically informed corpora.

uzT5 requires a narrower reading. It should not be overstated as a direct substitute for a dedicated translation backbone. Still, the Uzbek language adaptation results are too important to ignore. Kushmuratov and Davronov report that “uzT5-base demonstrated the highest performance”. Their task is text generation rather than bilingual translation, yet the finding matters because it shows that Uzbek adapted encoder decoder models can outperform more generic baselines under limited resource conditions. In an Uzbek and Russian MT pipeline, uzT5 is especially attractive for pretranslation normalization, synthetic paraphrase generation, post correction, terminology repair, and augmentation of the Uzbek side of the corpus. It is auxiliary. But strategically auxiliary.

SeamlessM4T belongs to the forward looking layer of the comparison. Lavrentyeva and Kogan include it among current transformer based MT families, and the HSE report on speech models highlights SeamlessM4T as one of the systems chosen for comparative analysis in a multilingual setting. If the long term objective is not only text translation but also subtitle generation, dubbing, or a full speech to speech Uzbek and Russian pipeline, this family becomes highly relevant. For pure text MT, though, it is rarely the first economical choice. Its strongest argument is not simplicity, but ecosystem breadth.

DISCUSSION

The comparison suggests that fine tuning for Uzbek and Russian should be staged, not monolithic. One begins with orthographic normalization and alignment, then adds terminology control, then chooses a backbone, and only after that performs either full fine tuning or a parameter efficient variant. The logic is supported by two independent lines of Russian language research. First, corpus based work insists on aligned bilingual material as the real instrument of analysis and engineering. Second, recent MT work treats post editing and domain adjustment as central rather than peripheral procedures. Evaluation needs equal caution. Mitrenina and Mukhambetkalieva write that “Current translation quality assessment metrics produce distorted results”[5]. The warning is especially relevant for Uzbek and Russian because the pair often compresses meaning differently. BLEU may reward local overlap while ignoring evidential nuance, discourse naturalness, or the correct treatment of culturally loaded items. A dissertation level study should therefore combine BLEU or COMET style automatic assessment with targeted human scoring on morphology, terminology, idiomaticity, and named entities. Otherwise, improvements will remain partly illusory.

Parameter efficient adaptation is not a fashionable addition here. It is a methodological necessity. Kushmuratov and Davronov show strong Uzbek side gains under LoRA oriented adaptation, and the NSU report on prefix tuning argues that the method suits low data settings because only a much smaller parameter subset is optimized. This matters in practice. Uzbek and Russian corpora are rarely large, perfectly aligned, and genre balanced at the same time. A compact and repeated adaptation loop is therefore more realistic than one expensive full retraining cycle.

CONCLUSION

The comparative analysis shows that Uzbek and Russian MT does not benefit from a single universal recommendation. Different model families solve different bottlenecks. MarianMT is efficient in narrow domains. mBART 50 is highly suitable for supervised adaptation with limited but curated data. M2M100 remains a useful multilingual warm start. NLLB 200 offers the strongest low resource transfer logic, especially when related Turkic material is available. uzT5 is not the main bilingual engine, yet it is a strong auxiliary model on the Uzbek side. SeamlessM4T is best treated as an investment in multimodal futures rather than as the default text only baseline.

REFERENCES

1. Dyomochkina V. V., Gruzdev D. Y., Lukyanova E. V. Machine translation in hindsight //Научный результат. Вопросы теоретической и прикладной лингвистики. – 2024. – Т. 10. – №. 2. – С. 21-45.
2. Камшилова О. Н., Беляева Л. Н. Машинный перевод в эпоху цифровизации: новые практики, процедуры и ресурсы //Terra Linguistica. – 2023. – Т. 14. – №. 1. – С. 41-56.
3. Абдурахманова М. Т., Абжалова М. А. Параллельные корпуса на примере узбекского, русского и английского языков //Конвергентные технологии XXI: вариативность, комбинаторика, коммуникация. – 2022. – С. 74-81.
4. Эргашева Г., Хаиткулов З., Кучимова Н. Параллельный корпус в автоматизированном переводе (на примере коллокаций) // O‘zbekistonda xorijiy tillar. – 2023. № 5 (52). – С. 162-173.
5. Митренина О. В., Мухамбеткалиева А. Г. Как и какой перевод (не) оценивают компьютеры //Journal of applied linguistics and lexicography. – 2021. – Т. 3. – №. 2. – С. 77-84.

6. Nigmatova L., Avezov S. ПРИМЕНЕНИЕ МЕТОДОВ NLP В КОРПУСНЫХ ИССЛЕДОВАНИЯХ: ОСОБЕННОСТИ И ОГРАНИЧЕНИЯ //«УЗБЕКСКИЕ НАЦИОНАЛЬНЫЕ ОБРАЗОВАТЕЛЬНЫЕ ЗДАНИЯ ТЕОРЕТИЧЕСКОЕ И ПРАКТИЧЕСКОЕ СОЗДАНИЕ ВОПРОСЫ" Международная научно-практическая конференция. – 2023. – Т. 2. – №. 2.
7. Авезов С. Корпусная лингвистика: новые подходы к анализу языка и их приложения в обучении иностранным языкам //International Bulletin of Applied Science and Technology. – 2023. – Т. 3. – №. 7. – С. 177-181.