MACHINE TRANSLATION IN LOW-RESOURCE LANGUAGES: THE UZBEK CASE

Ismatova Yulduz Nu'mon qizi
PhD, Senior Lecturer, Samarkand State Institute of Foreign Languages
https://orcid.org/0000-0001-8698-6250

Abstract. This article examines the challenges of machine translation for low-resource languages, focusing on Uzbek as a representative case. The concept of resource richness is analysed through three dimensions: parallel corpora, monolingual corpora, and linguistic tools. The specific difficulties of Uzbek — its agglutinative morphology, the 1990s script transition, and insufficient representation in international neural models — are analysed. Experimental observations show that English–Uzbek neural translation produces systematic errors across five groups (lexical-semantic, phraseological, morphological-syntactic, cultural-realie, artistic-aesthetic), with a high proportion of 'hidden' errors. The article proposes strategies for improving Uzbek machine translation quality: corpus development, domain adaptation, community-driven data collection, and the systematic use of post-editing. The conclusion argues that low-resource status is a temporary condition that can be overcome through coordinated national and international efforts, and that post-editing currently serves as the essential quality guarantee for Uzbek literary translation.

Keywords: low-resource languages, machine translation, Uzbek language, neural models, parallel corpora, post-editing.

Machine translation (MT) has achieved remarkable quality for high-resource language pairs such as English–French, English–German, and English–Chinese. However, the majority of the world's languages — estimated at over 6,000 — remain low-resource in MT terms: they lack the large parallel corpora that neural systems require. Uzbek, with approximately 35 million speakers, is a typical case: despite its significant speaker population, it remains under-resourced in international MT systems. This article analyses the specific challenges of Uzbek MT, examines the errors produced by current neural systems, and proposes strategies for improvement. The research is based on comparative analysis of English–Uzbek MT outputs (Google Translate, DeepL), theoretical works on low-resource MT, and Uzbek computational linguistics literature (Abdurakhmonova, 2020; Po'latov, 2011; Olimjonov & Abidova, 2023).

A language's resource richness in MT is determined by three factors. First, parallel corpora: the availability of large amounts of text translated between the language and major world languages. English–French has billions of words of parallel text (European Parliament proceedings, UN documents, literary translations); English–Uzbek has only a fraction of this. Second, monolingual corpora: large collections of text in the language itself, used for language modelling. Uzbek has growing digital content, but it remains limited compared to English. Third, linguistic tools: morphological analysers, syntactic parsers, terminological databases. Uzbek has some tools developed by local researchers (Po'latov, 2011), but they are not integrated into international MT pipelines. The combination of these three factors determines the quality ceiling for MT: even the best neural architecture cannot overcome a lack of training data (Olimjonov & Abidova, 2023).

Uzbek presents several specific challenges beyond general low-resource status. First, agglutinative morphology: Uzbek words are formed by chains of suffixes (kitoblarida = book + plural + possessive +

© 2026Scottish International Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

locative). Neural models trained on limited Uzbek data struggle to generate these suffix chains correctly, especially in literary texts where morphological choices carry stylistic meaning (Article 8 of this series). Second, the script transition: Uzbek switched from Cyrillic to Latin script in the 1990s, creating a situation where historical texts exist in two scripts and modern digital content is unevenly distributed. This complicates corpus building and creates orthographic inconsistency. Third, typological distance from English: English is analytic, Uzbek is agglutinative; the morphological richness that Uzbek requires (verbal tense-person forms, case marking) must be generated from English source structures that do not explicitly encode them (Yendirboyeva, 2025). Fourth, domain imbalance: available Uzbek parallel text is dominated by official/legal documents, while literary, technical, and colloquial domains are under-represented. This produces MT systems that perform reasonably on formal text but poorly on literary dialogue and description.

Experimental analysis of English–Uzbek MT outputs reveals systematic errors across five groups (detailed in Article 4 of this series). Lexical-semantic errors (30–35%): the model selects the most frequent meaning of polysemous words, while literary contexts often require secondary meanings. Phraseological errors (15–20%): idioms are translated literally (to break the ice → *break the ice*), losing their communicative function. Morphological-syntactic errors (20–25%): complex verbal forms are simplified, case chains are broken, possessive markers are dropped. Cultural-realie errors (10–15%): English cultural items are transliterated without explanation or replaced with generic words. Artistic-aesthetic deficiencies (10–15%): authorial voice, character idiolects, rhythm, and subtext are flattened into a uniform style. A critical feature of these errors is that a high proportion (50–60%) are 'hidden': the output is syntactically well-formed and reads fluently, but is semantically or pragmatically wrong. This makes post-editing cognitively demanding and requires trained post-editors (Koponen, 2016; Bogdanova, 2024).

Error group	**Frequency**	**Visibility**	**Post-editing difficulty**
Lexical-semantic	30–35%	mostly hidden	medium
Phraseological	15–20%	open	medium
Morphological-syntactic	20–25%	open	low-medium
Cultural-realie	10–15%	semi-hidden	high
Artistic-aesthetic	10–15%	hidden	very high

Several strategies can improve Uzbek MT quality. First, corpus development: creating large, high-quality parallel corpora is the fundamental solution. This requires coordinated effort: national digitization programs, translation memory banks from government and international organisations, and community-driven collection of literary translations. Olimjonov and Abidova (2023) describe the creation of open parallel corpora for Uzbek — this work should be expanded. Second, domain adaptation: fine-tuning general models on domain-specific data (literary, technical, legal) improves performance in those domains. Adapters and parameter-efficient fine-tuning make this feasible even with limited data. Third, multilingual transfer: Uzbek can benefit from transfer learning from related Turkic languages (Turkish, Kazakh, Kyrgyz) — a multilingual model trained on Turkic data can transfer some knowledge to Uzbek. Fourth, orthographic normalisation: creating tools that convert Cyrillic Uzbek to Latin Uzbek would

© 2026Scottish International Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

unlock historical text for corpus building. Fifth, human-in-the-loop: systematic post-editing of MT outputs, with error annotation, creates valuable training data for model improvement — post-editing thus serves double duty: quality assurance today, training data tomorrow (Toral et al., 2018).

Until corpus and tool development reach critical mass, post-editing remains the essential quality guarantee for Uzbek literary translation. The previous articles in this series have developed the conceptual and methodological basis for this: the five-group error typology (Article 4), post-editing strategies for each group (Articles 5–8, 11), the four-stage post-editing model (Article 1), and the post-editor competence framework (Article 9). The key principle is that the post-editor's role is not temporary but constitutive: even as MT quality improves, the artistic-aesthetic layer of literary texts will remain the post-editor's domain, because it requires human aesthetic judgement that statistical systems cannot replicate (Moorkens et al., 2018). This principle is generalisable: it applies to all low-resource languages where MT provides a useful first draft but not a finished translation.

Uzbek is not alone in facing the combination of low-resource status and agglutinative morphology. Turkish, despite being relatively well-resourced, still shows systematic morphological errors in MT outputs; Finnish, Estonian, Hungarian, Basque, and Korean present similar challenges. Research on these languages has produced useful insights. First, morphological segmentation as preprocessing: splitting words into morphemes before MT and recombining after improves quality for agglutinative languages. Second, subword units (BPE, SentencePiece) with larger vocabularies help handle productive morphology. Third, explicit morphological features (case, number, person) can be provided to the model as additional input. Fourth, evaluation metrics designed for morphologically rich languages (avoiding penalising legitimate morphological variants) give more accurate quality estimates. Uzbek MT development can directly benefit from these lessons: morphological analysers exist (Po'latov, 2011), segmentation approaches are well-documented, and evaluation adaptations are feasible. The key missing element is coordination: bringing together linguists, computer scientists, and translators in a focused national effort.

Evaluating MT quality for low-resource languages presents specific challenges. Automatic metrics (BLEU, METEOR, TER) require reference translations; for Uzbek, high-quality reference translations are scarce, and the available ones may not represent the best translation practice. Human evaluation is more reliable but expensive and difficult to standardise. Toral et al. (2018) used post-editing effort as an indirect quality measure for literary translation — the logic being that a better MT output requires less post-editing effort. This approach is particularly suitable for low-resource languages: it directly measures what matters practically (how much human work is needed) rather than relying on imperfect automatic metrics. For Uzbek, we recommend developing a multi-metric evaluation framework: automatic metrics (with Uzbek-adapted tokenisation), post-editing effort measurement, and expert human evaluation on a standard test set of literary, technical, and official texts. Such a benchmark would enable systematic comparison of MT systems, track progress over time, and guide resource allocation.

MT development for low-resource languages has ethical dimensions that deserve attention. First, digital language equality: as MT becomes a gateway to information, languages poorly served by MT risk digital marginalisation — their speakers have less access to the world's knowledge. Second, cultural authenticity: MT systems trained predominantly on official documents may fail to represent the colloquial, poetic, and dialectal richness of Uzbek; over-reliance on such systems could flatten language use. Third, professional displacement: MT may reduce demand for human translators, but as this article

© 2026Scottish International Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

series argues, it also creates new roles (post-editors) — the transition must be managed through retraining programmes. Fourth, data sovereignty: corpus collection should respect intellectual property and community rights; literary translations are copyrighted works, and their use for MT training requires appropriate licensing. These considerations suggest that MT development for Uzbek should be guided not only by technical goals but also by language policy and cultural preservation principles.

The disparity in MT quality between high-resource and low-resource languages is one manifestation of the broader digital language divide. English-language content dominates the internet, digital tools, and AI systems; speakers of smaller languages face reduced access to digital services and information. Uzbekistan has recognised this challenge in its language policy: the state program on language development emphasises the expansion of Uzbek-language digital content and tools. Concrete steps include the digitisation of Uzbek literary classics, the development of national terminological databases, and support for local NLP research. However, MT development requires sustained, coordinated investment: corpus building is labour-intensive and yields results over years, not months. International cooperation can accelerate progress: partnerships with organisations that have corpus-building expertise (TILDE, ELRC, OPUS), participation in shared tasks for Turkic languages, and collaboration with the Uzbek diaspora for text collection. The digital language divide is not inevitable — it is the cumulative result of investment decisions, and those decisions can be changed. For Uzbek, the window of opportunity is now: the demand for translation is growing, MT technology is mature, and the path from low-resource to medium-resource status is well understood.

Uzbek MT faces the combined challenges of low-resource status, agglutinative morphology, script transition history, and typological distance from English. Current neural systems produce systematic errors across five groups, with a high proportion of hidden errors that make post-editing cognitively demanding. However, low-resource status is not a permanent condition: through corpus development, domain adaptation, multilingual transfer, and systematic post-editing with error annotation, Uzbek MT quality can be progressively improved. In the interim, post-editing remains the essential quality guarantee, and its methodology — developed in this article series — provides the practical framework. The Uzbek case illustrates a general principle: for low-resource languages, the combination of MT (for speed and volume) and trained post-editors (for quality and artistry) is the realistic path to meeting growing translation demands. Future research should focus on quantitative measurement of error reduction from corpus development, comparative evaluation of transfer learning approaches for Turkic languages, and the creation of a national Uzbek MT evaluation benchmark.

REFERENCES

1. Koponen M. Is machine translation post-editing worth the effort? A survey of research into post-editing and effort // The Journal of Specialised Translation. 2016. № 25. P. 131–148.
2. Moorkens J., Toral A., Castilho S., Way A. Translators' perceptions of literary post-editing using statistical and neural machine translation // Translation Spaces. 2018. Vol. 7, № 2. P. 240–262.
3. Toral A., Wieling M., Way A. Post-editing effort of a novel with statistical and neural machine translation // Frontiers in Digital Humanities. 2018. Vol. 5. Art. 9.
4. Богданова К. В. Постредактирование машинного перевода как студенческий проект (на материале художественного текста) // Исследования языка и современное гуманитарное знание. 2024. Т. 6, № 3. С. 172–183.

© 2026Scottish International Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

5. Abduraxmonova N. M. Mashina tarjimasing lingvistik ta'minoti. Toshkent: Lambert Academic Publishing, 2020.
6. Olimjonov A. F., Abidova Sh. B. Mashina tarjimasi uchun ochiq korpusli parallel korpus bazalari yaratish // Educational Research in Universal Sciences. 2023. № 2 (2). B. 44–50.
7. Po'latov A. Kompyuter lingvistikasi. Toshkent: O'zbekiston Milliy universiteti nashriyoti, 2011.
8. Yendirboyeva Z. Z. Mashina tarjimasi: til va texnologiya kesishmasidagi dolzarb masalalar // American Journal of Social Science. 2025. Vol. 3, № 9.

© 2026Scottish International Conference on Multidisciplinary Research and Innovation. All Rights Reserved.