# PARALLEL CORPORA AND NEURAL MODELS FOR THE UZBEK LANGUAGE: CURRENT STATE AND DEVELOPMENT PATHS

**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 role of parallel corpora in developing neural machine translation for the Uzbek language. The concept of a parallel corpus is analysed, existing resources for Uzbek are reviewed, and methodological principles of corpus construction are formulated: representativeness, alignment quality, metadata annotation, and licensing. The article analyses how corpus characteristics affect neural model behaviour — error types observed in English–Uzbek MT are correlated with corpus gaps. Development paths are proposed: national corpus programs, domain-specific corpora, community-driven collection, and post-editing data feedback loops. The conclusion argues that corpus construction is the single most effective investment for improving Uzbek MT quality, and that a coordinated national effort, combining academic, governmental, and community resources, is needed to move Uzbek from low-resource to medium-resource status.

**Keywords:** parallel corpora, neural machine translation, Uzbek language, corpus linguistics, alignment, language resources.

Neural machine translation is fundamentally a data-driven technology: the quality of a model is determined primarily by the quality and quantity of its training data. For Uzbek, as for most languages, the bottleneck is not algorithmic but data: parallel corpora (collections of texts with their translations) are scarce, unevenly distributed across domains, and often of insufficient quality. This article examines the state of parallel corpus resources for Uzbek, analyses how corpus characteristics shape MT behaviour, and proposes concrete development paths. The analysis draws on Uzbek computational linguistics research (Olimjonov & Abidova, 2023; Abdurakhmonova, 2020; Po'latov, 2011) and international corpus linguistics methodology.

A parallel corpus is a collection of texts in one language aligned with their translations in another language. Alignment can be at the document level (which documents correspond to each other), the paragraph level, or the sentence level — sentence alignment being the most useful for MT training. Parallel corpora differ by domain: legal (treaties, legislation), institutional (EU, UN proceedings — not applicable for Uzbek), literary (novels, stories), technical (manuals, documentation), web (crawled parallel content). For MT training, corpus quality matters more than raw size: a smaller, clean, well-aligned corpus outperforms a larger, noisy one. Key quality parameters: alignment accuracy (are sentence pairs truly translations of each other?), translation quality (are the human translations professional?), coverage (does the corpus represent the target domain?), and recency (does it reflect current language use?).

Existing parallel resources for Uzbek are limited but growing. The OPUS project provides some Uzbek parallel data, primarily from governmental and legal sources. The Jehovah's Witnesses corpus provides substantial parallel text (religious domain). Some international organisations (UN agencies) produce Uzbek versions of official documents. Local efforts include the open parallel corpora described by Olimjonov and Abidova (2023). Literary translations exist — Uzbek has a rich tradition of translating world literature — but these are scattered across printed editions, not digitised or aligned. The key gaps are: (1) literary domain — few digitised, aligned literary translations; (2) technical domain — manuals



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

and documentation rarely translated into Uzbek; (3) colloquial domain — almost no parallel data for spoken Uzbek; (4) quality — existing parallel data often comes from non-professional translation. The script transition adds complexity: older parallel texts may exist only in Cyrillic Uzbek, requiring transliteration before use.

The systematic errors observed in English–Uzbek MT (Article 4 of this series) can be directly correlated with corpus characteristics. Lexical-semantic errors (30–35%) reflect insufficient coverage of literary vocabulary in the training corpus: the model defaults to frequent, 'safe' meanings. Phraseological errors (15–20%) reflect the near-absence of idiom-aligned training pairs: the model has rarely seen idioms translated as units. Morphological-syntactic errors (20–25%) reflect agglutinative morphology being under-represented: complex suffix chains are rare in the corpus, so the model prefers simple forms. Cultural-realie errors (10–15%) reflect the dominance of official-domain training data: the model has seen few cultural items and treats them as ordinary vocabulary. Artistic-aesthetic deficiencies (10–15%) reflect the absence of stylistically annotated training data: the model has no signal for authorial voice. This correlation is actionable: it tells us exactly what data to collect to fix each error type. Collecting aligned literary translations with phraseological units and cultural items would directly reduce the first four error groups; the fifth (aesthetic) requires human post-editing, but even it benefits from training on post-edited data.

Building effective parallel corpora for Uzbek requires methodological rigour. Principle 1: representativeness — the corpus should cover the domains where MT will be used (literary, technical, official, web). A corpus of only official documents produces an MT system good at official style and poor at literature. Principle 2: alignment quality — automated alignment (e.g., using sentence-length and lexical similarity heuristics) must be manually verified for a sample; systematic misalignment corrupts training. Hamroyeva and Matyakubova (2023) describe alignment methods applicable to Uzbek. Principle 3: metadata annotation — each text should carry metadata (domain, source, translator qualifications, date, script) enabling targeted training and evaluation. Principle 4: licensing — corpus construction must respect copyright; literary translations are protected works, and their use requires licensing agreements with publishers and translators' unions. Principle 5: iterative growth — a corpus is never finished; it should grow continuously through feedback from MT use, with post-edited outputs contributing new training pairs (Toral et al., 2018).

We propose a coordinated national corpus program with five components. Component 1: digitisation of literary translations — Uzbek classics of world literature translation should be digitised, aligned, and licensed; this is the highest-value single investment for literary MT. Component 2: institutional parallel documents — government, UN agencies, and international organisations should release their Uzbek translations as open data. Component 3: web harvesting — parallel web content (news sites publishing in both languages, bilingual government sites) can be automatically harvested and filtered. Component 4: community-driven collection — a platform where translators and readers can contribute aligned texts, with quality review. Component 5: post-editing feedback loop — MT users (post-editors, translation agencies) contribute their post-edited outputs to the corpus, with error annotation; this creates a virtuous cycle where every use of MT improves it. Olimjonov and Abidova's (2023) work on open parallel corpora provides a foundation; the program scales it up to national level.

With a corpus in place, model development follows established paths. Baseline: fine-tune a strong multilingual model (e.g., NLLB, MADLAD) on Uzbek parallel data. The NLLB model already covers Uzbek, providing a starting point better than generic commercial systems. Domain adaptation: fine-tune separately on literary and technical sub-corpora to create domain-specialised models. Morphological


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

augmentation: use Uzbek morphological analysers (Po'latov, 2011) to generate additional training examples by applying valid morphological transformations to existing sentences — a form of data augmentation for agglutinative languages. Transfer learning from Turkic languages: multilingual fine-tuning on Turkish + Uzbek data leverages the shared Turkic features. Evaluation: maintain a held-out test set (literary, technical, official) for systematic model comparison. The infrastructure for these approaches exists; what is needed is the data and coordination.

For practical orientation, we summarise the main existing parallel resources relevant to Uzbek MT development:

| **Resource** | **Domain** | **Approx. size** | **Quality/access** |
|---|---|---|---|
| OPUS (Tanzil, JW300, etc.) | religious, legal | medium | open, variable alignment |
| UN documents | official | small-medium | open, professional translation |
| Government sites (bilingual) | official, informational | small | open, needs harvesting |
| Olimjonov & Abidova corpora | mixed | small-medium | open, research quality |
| Printed literary translations | literary | potentially large | not digitised, copyright |
| Post-editing outputs (potential) | mixed | growing | requires collection infrastructure |

The table shows the core paradox of Uzbek corpus work: the highest-value resource (literary translations) is the least accessible, while the most accessible resources (religious, official) are the least representative of general language use. Resolving this paradox requires targeted investment in digitisation and licensing of literary translations — the single highest-leverage action.

Corpus construction requires appropriate tools and formats. Alignment tools: Hunalign, Bleualign, and fast-align support sentence alignment; for Uzbek, alignment quality benefits from morphological normalisation before alignment (because suffix chains create spurious differences). Formats: the standard exchange formats are TMX (Translation Memory eXchange) for aligned segments and XLIFF for localisation workflows; plain-text bilingual files (one language per column) are the simplest for MT training. Infrastructure: a corpus repository with version control (tracking corrections and additions), search interface (enabling researchers to query aligned pairs), and API access (enabling MT developers to download training data). Quality assurance workflow: automated checks (length ratios, script consistency, encoding) followed by human review of flagged segments. Annotation tools: web-based interfaces (e.g., Okapi, MateCat) support collaborative annotation of domain, quality, and error types. The technological infrastructure for corpus construction is mature and largely open-source; the investment required is primarily organisational and human.

Community-driven corpus collection offers scale but requires robust quality control. Models from successful crowdsourcing projects (e.g., Common Voice for speech, Tatoeba for sentences) suggest principles applicable to Uzbek parallel corpus building. Principle 1: graded contribution — simple tasks (validating existing pairs) for casual contributors, complex tasks (translating new segments) for verified translators. Principle 2: consensus validation — each segment is validated by multiple contributors; disagreements escalate to experts. Principle 3: contributor reputation — track record of quality


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

contributions unlocks higher-privilege tasks. Principle 4: immediate feedback — contributors see the impact of their work (e.g., their validated pairs entering the corpus). Principle 5: quality sampling — experts regularly audit random samples of contributed data. Applied to Uzbek, a crowdsourcing platform could engage the large community of Uzbek translators, teachers, and students: teachers could contribute classroom translation exercises, professional translators could contribute their work (with licensing), and students could validate pairs as learning activity. This approach converts corpus building from a specialised technical task into a national educational and cultural project — with quality control ensuring research-grade output.

Parallel corpus construction is the single most effective investment for improving Uzbek MT quality. The systematic errors of current systems are directly traceable to corpus gaps: literary translations, idioms, cultural items, and morphological richness are under-represented. A coordinated national corpus program — combining digitisation of literary classics, institutional open data, web harvesting, community collection, and post-editing feedback — can move Uzbek from low-resource to medium-resource status within a few years. Until then, and even after, post-editing remains the essential quality guarantee for literary translation, with its methodology developed in this article series. The Uzbek case demonstrates a general principle: in the neural MT era, language equality is built not only on algorithms but above all on data — and data is created by coordinated human effort. Future research should focus on corpus quality metrics for agglutinative languages, copyright frameworks for literary corpus licensing, and quantitative measurement of corpus growth effects on MT quality.

**REFERENCES**

1. 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.
2. Vieira L. N. Post-editing of machine translation // The Routledge Handbook of Translation and Technology / ed. by M. O'Hagan. London: Routledge, 2019. P. 319–335.
3. Abduraxmonova N. M. Mashina tarjimasing lingvistik ta'minoti. Toshkent: Lambert Academic Publishing, 2020.
4. Hamroyeva Sh., Matyakubova N. Mashina tarjimasida matnni moslashtirish usullari // Uzbekistan: Til va madaniyat. 2023. № 2. B. 6–19.
5. 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.
6. Po'latov A. Kompyuter lingvistikasi. Toshkent: O'zbekiston Milliy universiteti nashriyoti, 2011.
7. Yendirboyeva Z. Z. Mashina tarjimasi: til va texnologiya kesishmasidagi dolzarb masalalar // American Journal of Social Science. 2025. Vol. 3, № 9.



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