IQRO JURNALI , IQRO JOURNAL, ИКРО ЖУРНАЛИ
ISSN:2181-4341| wordlyknowledge.uz/index.php/iqro
Volume 24. Issue 01. Oktabr, 2026

COMPUTATIONAL APPROACHES AND LINGUISTIC ANNOTATION STRATEGIES FOR CONSTRUCTING UZBEK–CHINESE PARALLEL CORPUS RESOURCES: DATA COLLECTION, SENTENCE ALIGNMENT, AND QUALITY EVALUATION

Xiaoqi You
National University of Uzbekistan (PhD)
E-mail: 726366681@qq.com
https://doi.org/10.5281/zenodo.23858500

Abstract: Uzbek–Chinese is a low-resource language pair for which large, curated parallel corpora remain scarce, despite growing academic, economic and cultural exchange between Uzbekistan and China. This article examines computational approaches and linguistic annotation strategies for constructing Uzbek–Chinese parallel corpus resources. It reviews data collection channels, preprocessing problems specific to the two languages, embedding-based sentence alignment, filtering methods and annotation layers, and it proposes a quality evaluation framework that combines manual auditing with automatic and downstream measures. The analysis draws on recent work on multilingual sentence embeddings, linear-time sentence alignment, audits of web-crawled multilingual data and multilingual evaluation benchmarks. The article argues that the quality of an Uzbek–Chinese corpus depends less on its size than on script normalization, segmentation consistency, alignment confidence and documented human evaluation. A practical pipeline is outlined, together with annotation guidelines and indicators that allow corpus quality to be reported transparently.

Keywords: parallel corpus, Uzbek, Chinese, sentence alignment, sentence embeddings, corpus annotation, quality evaluation, low-resource machine translation.

INTRODUCTION

Parallel corpora, that is, collections of texts paired with their translations at the sentence level, are the basic resource for machine translation, bilingual lexicography, contrastive linguistics and translation studies. For widely studied language pairs such resources are abundant, but for many pairs involving lower-resource languages they remain limited. Large multilingual translation projects have attempted to close this gap. The No Language Left Behind project developed a single translation model covering 200 languages and introduced the FLORES-200 evaluation benchmark to measure performance across them [1].

Benchmarks of this kind have made it possible to evaluate translation quality for languages that previously had no reliable test sets. The FLORES-101 benchmark consists of 3,001 sentences drawn from English Wikipedia and translated into 101 languages by professional translators through a controlled process, so that the translations are aligned across all languages [2]. Such multi-way aligned data provide a valuable reference point, but they are small and are designed for evaluation rather than for training or for linguistic research on a specific language pair.

The availability of large, automatically collected corpora does not guarantee their quality. A manual audit of 205 language-specific corpora from five major public datasets found that at least 15 corpora were completely erroneous, that a significant fraction of lower-resource corpora

contained less than 50 percent of sentences of acceptable quality, and that 82 corpora were mislabeled or used nonstandard language codes [3]. These findings are directly relevant to Uzbek–Chinese, because web-crawled data for lower-resource pairs are most exposed to such problems.

Uzbek and Chinese differ sharply in their linguistic structure. Uzbek is an agglutinative Turkic language that forms words through chains of suffixes, and it is written in both Latin and Cyrillic scripts. Chinese is an isolating language written in characters without spaces between words. These differences affect every stage of corpus construction, from sentence segmentation to alignment and annotation. The purpose of this article is to analyze computational approaches and annotation strategies for constructing Uzbek–Chinese parallel corpora and to propose a quality evaluation framework suitable for this pair. The study addresses four questions: which data sources and preprocessing steps are required; which alignment methods are most appropriate; which annotation layers add the greatest value; and how corpus quality can be evaluated and reported.

LITERATURE REVIEW AND METHODOLOGY

Open resources and tools form the infrastructure on which new corpora are built. The OPUS-MT project provides more than 1,000 pre-trained translation models trained on openly available parallel data, together with open-source applications that can run on ordinary hardware [4]. Such resources are useful for corpus construction in two ways: they can serve as baseline systems for downstream evaluation, and they can be used to generate auxiliary translations during alignment and filtering.

Sentence alignment is the central computational problem in parallel corpus construction. Vecalign, proposed by Thompson and Koehn, performs bilingual sentence alignment in time and space linear in the number of sentences and requires only bilingual sentence embeddings. On a German–French test set it outperformed the previous state of the art by 5 F1 points, it substantially outperformed the widely used Hunalign toolkit on medium- and low-resource language pairs, and it improved downstream translation by 1.7 BLEU for Sinhala–English and 1.6 BLEU for Nepali–English [5]. These results indicate that embedding-based alignment is particularly suitable for pairs with limited bilingual dictionaries.

The quality of embedding-based alignment depends on the multilingual sentence encoder. The Language-agnostic BERT Sentence Embedding model (LaBSE) combines masked and translation language modeling with a dual-encoder translation ranking objective. It achieved 83.7 percent bitext retrieval accuracy on the Tatoeba benchmark across 112 languages, compared with 65.5 percent for an earlier baseline, and it was used to mine parallel data from web text for training competitive translation systems [6]. Encoders of this kind provide the similarity scores that alignment algorithms and filtering procedures rely on.

Related work on neighboring language pairs offers transferable methods. A study on obtaining parallel sentences for low-resource language pairs with minimal supervision demonstrated its approach by harvesting Uyghur–Chinese parallel sentences and building a translation system from them [7]. Uyghur is, like Uzbek, a Turkic language, and the challenges of pairing an agglutinative language with Chinese are therefore comparable. Recent work on Southern Uzbek has also shown how small, carefully curated resources can be built: it released a 997-sentence FLORES+ development set, 39,994 parallel sentences collected from dictionary, literary and web sources, and a fine-tuned NLLB-200 model [8].

The methodology of the present study is analytical and design-oriented. It synthesizes the methods reported in the cited studies, relates them to the linguistic properties of Uzbek and Chinese, and formulates a pipeline with explicit quality checks. Numerical examples given in the discussion are illustrative and are labelled as such; they are intended to show how the proposed indicators are calculated rather than to report results of a completed corpus. Sources were selected from peer-reviewed proceedings, journals and documented preprints published in 2016 or later.

1. Data collection channels. Potential sources of Uzbek–Chinese parallel text include bilingual official and institutional websites, translated news, multilingual documents of international organizations, translated literary and educational works, and subtitles. Each channel differs in domain, register and translation directness. Official documents tend to be formal and closely translated, while news and subtitles are freer. Many texts may also have been translated indirectly through Russian or English, which introduces pivot effects that should be documented.

2. Legal and ethical aspects. Corpus construction requires attention to copyright and licensing. Openly licensed sources should be preferred, and the license of each document should be recorded in the corpus metadata. Personal information appearing in collected texts, such as names in news or subtitles, should be handled according to applicable data protection rules, and the corpus documentation should state how such content was treated.

3. Preprocessing of Uzbek. Uzbek preprocessing must address script variation. Texts may appear in Latin or Cyrillic script, and the Latin letters o‘ and g‘ are written with several different apostrophe-like characters in practice. Normalizing these characters to a single convention, and converting Cyrillic text to the chosen script with a documented transliteration table, is necessary to prevent the same word from appearing in several forms. Because Uzbek is agglutinative, morphological segmentation into stems and suffixes can further reduce data sparsity for both alignment and analysis.

4. Preprocessing of Chinese. Chinese text requires consistent handling of character variants, full-width and half-width punctuation and word segmentation. Sentence boundaries are generally marked by distinctive punctuation, but word boundaries are not marked by spaces, so a segmentation tool and a fixed segmentation standard should be chosen and documented. Inconsistent segmentation across the corpus would distort length-based statistics and token-level annotation.

5. Sentence segmentation and alignment. After normalization, documents are split into sentences and aligned. Embedding-based alignment with an algorithm such as Vecalign, using sentence embeddings from a multilingual encoder such as LaBSE, allows one-to-one as well as one-to-many and many-to-one alignments to be found efficiently [5], [6]. Many-to-one alignments are particularly relevant for Uzbek–Chinese, because a long Uzbek sentence with several subordinate clauses may correspond to two Chinese sentences, and vice versa.

RESULTS AND DISCUSSION

The analysis leads to a pipeline with six stages: collection, normalization, segmentation, alignment, filtering and annotation, followed by evaluation. The central principle of the pipeline is that each stage produces documented output and quality statistics, so that problems can be traced to the stage where they arose. This principle responds directly to the audit findings on web-crawled

corpora, which showed that many quality problems remain hidden when datasets are released without systematic inspection [3].

Filtering removes sentence pairs that are likely to be misaligned or noisy. Four filters are recommended. The first is language identification, which verifies that each side of a pair is in the expected language and script; the audit of multilingual datasets showed that mislabeled language codes are a common problem [3]. The second is a length-ratio filter, which removes pairs whose lengths are inconsistent with typical Uzbek–Chinese ratios. Because Chinese characters encode more information per character than Latin letters, this ratio should be estimated from a manually verified sample rather than borrowed from other language pairs. The third is an embedding-similarity threshold, which removes pairs whose semantic similarity is low [6]. The fourth is deduplication, which removes repeated pairs that would otherwise bias training and evaluation.

The setting of the similarity threshold illustrates the trade-off between corpus size and quality. A higher threshold produces a smaller but cleaner corpus, while a lower threshold retains more data at the cost of more noise. As an illustrative example, suppose that 100,000 candidate pairs are scored and that a manual check of 200 randomly selected pairs at each of several thresholds is performed. If the share of acceptable pairs rises from 70 percent at a low threshold to 92 percent at a higher one while the corpus shrinks from 100,000 to 60,000 pairs, the choice of threshold becomes a documented decision about the intended use of the corpus. These figures are hypothetical and are given only to show how the decision should be supported by evidence.

Linguistic annotation adds value beyond the raw sentence pairs. Three annotation layers are proposed. The first is metadata annotation, which records the source, domain, register, translation direction where known, license and date of each document. The second is morphological and part-of-speech annotation, which is particularly useful for Uzbek because suffix chains carry grammatical information that is expressed by separate words or word order in Chinese. The third is word or phrase alignment on a subset of the corpus, which supports bilingual lexicography and contrastive analysis of constructions such as case marking in Uzbek and prepositions or word order in Chinese.

Annotation guidelines are a condition of consistency. Annotators should work with a written manual that defines categories, gives examples and specifies how difficult cases are resolved. A portion of the data should be annotated independently by two annotators, and agreement should be measured with a chance-corrected statistic. Disagreements should be reviewed and used to refine the guidelines. Without this process, annotation quality cannot be compared across releases or across research groups.

Quality evaluation should combine three types of evidence. The first is manual auditing of a random sample of sentence pairs, using categories such as correct translation, partial translation, misalignment, wrong language and non-linguistic content. The audit methodology of Kreutzer and colleagues shows that even a modest manual sample can reveal serious problems in large datasets [3]. The second is automatic indicators, such as the distribution of similarity scores, length ratios and the share of pairs removed by each filter. The third is downstream evaluation, in which translation models trained or fine-tuned with the corpus are tested on an independent benchmark.

Downstream evaluation can use existing multilingual benchmarks and open models. FLORES-200 provides professionally translated, multi-way aligned test sentences that include

both Uzbek and Chinese, so that a Uzbek–Chinese test direction can be evaluated without creating a new test set [1], [2]. Open systems such as NLLB-200 and OPUS-MT models can serve as baselines against which models fine-tuned on the new corpus are compared [1], [4]. The work on Southern Uzbek shows that fine-tuning an NLLB-200 model on a curated corpus of about 40,000 sentence pairs is a feasible design for a lower-resource Uzbek variety [8].

The choice of automatic metrics should reflect the properties of the two languages. Word-based metrics are sensitive to segmentation decisions in Chinese and to the rich morphology of Uzbek, where a single word may correspond to a phrase in Chinese. Character-level metrics such as chrF are less affected by these differences and are therefore a useful complement to word-based scores. Reporting both types of metrics, together with the exact segmentation and normalization settings used for scoring, allows results to be compared across studies. Where resources permit, a small human evaluation of translation adequacy and fluency should accompany automatic scores, because automatic metrics cannot fully capture errors in case marking, aspect or word order.

Downstream gains, however, should not be the only criterion. A corpus may improve a translation metric while still containing misaligned or inappropriate content, and a corpus intended for linguistic research may be valuable even if its size is too small to improve a large translation model. For this reason, the evaluation report should present manual audit results, automatic statistics and downstream scores together, and should explain the intended uses of the corpus.

The treatment of pivot translation is an issue that deserves particular attention for Uzbek–Chinese. When documents were translated from Russian or English into both Uzbek and Chinese, the two sides are parallel to each other only indirectly. Such pairs can still be useful, but their translation relationship differs from direct translation, and they may contain systematic differences introduced by each translator. Recording the translation direction in the metadata, where it can be established, allows users to select subsets appropriate for their research.

Script and orthographic variation in Uzbek also affects evaluation. If test sets are written in Latin script while part of the training corpus was originally Cyrillic, inconsistent transliteration may lower scores for reasons unrelated to translation quality. A single normalization procedure applied to both training and test data is therefore required, and the procedure should be published with the corpus.

Methods developed for related Turkic–Chinese pairs can be adapted with relatively little change. The minimal-supervision approach applied to Uyghur–Chinese shows that parallel sentences can be harvested even when seed resources are limited [7]. For Uzbek–Chinese, a similar strategy could begin with a small manually aligned seed set, use it to calibrate the similarity threshold and the length-ratio filter, and then extend alignment to larger collections of comparable documents.

Named entities and terminology require a separate annotation strategy. Personal names, place names and organization names are transliterated between Uzbek and Chinese in ways that are often inconsistent across sources, and the same entity may appear in several spellings. Annotating named entities on both sides of a subset of the corpus, and linking variant spellings to a single identifier, improves the usefulness of the corpus for terminology extraction and for evaluating how translation systems handle names. Domain terminology, for example in legal, economic or technical texts, can be annotated in a similar way to support the creation of bilingual glossaries.

Storage format and versioning determine whether a corpus can be maintained over time. Sentence pairs can be stored in standard tab-separated or translation memory formats, with document-level metadata kept in a separate structured file. Each release should carry a version number and a change log describing added sources, modified filters and corrected errors. Versioning is particularly important when manual audits lead to the removal of problematic pairs, because users need to know which version of the corpus a published result is based on.

Comparable corpora offer an additional source of parallel sentences. Documents that describe the same events or topics in Uzbek and Chinese, without being direct translations of each other, may still contain sentences that are close translations. Multilingual sentence encoders make it possible to search such comparable collections for sentence pairs with high similarity scores [6]. Pairs extracted in this way should be marked separately in the metadata, because their reliability is lower than that of pairs from directly translated documents, and their acceptance rate should be estimated through a separate manual audit.

A transparent reporting format is the final element of the framework. The documentation of the corpus should include the size of the corpus by source and domain, the normalization and segmentation conventions, the alignment method and its parameters, the filtering thresholds and the share of data removed at each step, the results of the manual audit with confidence intervals where possible, the annotation guidelines and agreement statistics, and the downstream evaluation protocol. Such documentation makes the corpus reusable and allows its quality to be compared with that of other resources.

The study has several limitations. It proposes a design rather than reporting the results of a completed corpus, and the numerical examples are illustrative. The effectiveness of the proposed thresholds and filters for Uzbek–Chinese must be confirmed empirically on real data. In addition, the availability of openly licensed Uzbek–Chinese texts may limit the size of the corpus that can be released publicly. Future work should build a pilot corpus following the pipeline, report its audit results and evaluate its contribution to translation and linguistic research.

CONCLUSION

Constructing high-quality Uzbek–Chinese parallel corpora requires a combination of computational methods and careful linguistic decisions. Embedding-based sentence alignment and multilingual sentence encoders provide efficient tools for a language pair with limited bilingual resources, while script normalization, consistent segmentation and documented annotation guidelines determine whether the resulting corpus can be trusted. Audits of multilingual datasets show that size alone is not a measure of quality, and multilingual benchmarks make it possible to evaluate the downstream value of new data. Based on the analysis, the following recommendations are proposed:

first, Uzbek text should be normalized to a single script and apostrophe convention, and Chinese text should be segmented according to a fixed, documented standard;

second, sentence alignment should be performed with an embedding-based method that supports one-to-many and many-to-one alignments;

third, filtering thresholds should be calibrated on manually verified samples specific to the Uzbek–Chinese pair;

fourth, annotation should include metadata, morphological information for Uzbek and word alignment on a representative subset, supported by written guidelines and agreement measurement;

fifth, every release should be accompanied by a quality report that combines manual audit results, automatic statistics and downstream evaluation on an independent benchmark.

Following these principles will allow Uzbek–Chinese corpus resources to support both machine translation and linguistic research, and will make the quality of such resources visible to the researchers who use them.

REFERENCES

1. NLLB Team, Costa-jussà M.R. et al. No Language Left Behind: Scaling Human-Centered Machine Translation. – arXiv preprint, 2022. – arXiv:2207.04672. – URL: https://arxiv.org/abs/2207.04672 (accessed: 01.11.2026).
2. Goyal N. et al. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation // Transactions of the Association for Computational Linguistics. – 2022. – Vol. 10. – P. 522–538. – DOI: 10.1162/tacl_a_00474 (accessed: 01.11.2026).
3. Kreutzer J. et al. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets // Transactions of the Association for Computational Linguistics. – 2022. – Vol. 10. – P. 50–72. – DOI: 10.1162/tacl_a_00447 (accessed: 01.11.2026).
4. Tiedemann J., Thottingal S. OPUS-MT – Building open translation services for the World // Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. – Lisboa, 2020. – P. 479–480. – URL: https://aclanthology.org/2020.eamt-1.61 (accessed: 01.11.2026).
5. Thompson B., Koehn P. Vecalign: Improved Sentence Alignment in Linear Time and Space // Proceedings of EMNLP-IJCNLP 2019. – Hong Kong, 2019. – P. 1342–1348. – DOI: 10.18653/v1/D19-1136 (accessed: 01.11.2026).
6. Feng F., Yang Y., Cer D., Arivazhagan N., Wang W. Language-agnostic BERT Sentence Embedding // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – Dublin, 2022. – P. 878–891. – DOI: 10.18653/v1/2022.acl-long.62 (accessed: 01.11.2026).
7. [Author(s) not verified.] Obtaining Parallel Sentences in Low-Resource Language Pairs with Minimal Supervision. – Hindawi, 2022. – URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9365574/ (accessed: 01.11.2026).
8. [Author(s) not verified.] Filling the Gap for Uzbek: Creating Translation Resources for Southern Uzbek // Proceedings of the Tenth Conference on Machine Translation (WMT 2025). – 2025. – URL: https://aclanthology.org/2025.wmt-1.83/ (accessed: 01.11.2026).