PRINCIPLES OF DEVELOPMENT OF PARALLEL CORPUS-BASED LINGUISTIC RESOURCES FOR UZBEK AND CHINESE LANGUAGES

Xiaoqi You
PhD, National University of Uzbekistan
e-mail: 726366681@qq.com
tel.: +998503304175
https://doi.org/10.5281/zenodo.23858642

Abstract: The article substantiates the principles of developing parallel corpus-based linguistic resources for the Uzbek–Chinese language pair, which remains one of the least resourced combinations in Central Asian computational linguistics despite the rapid growth of political, economic and educational ties between Uzbekistan and China. On the basis of a comparative analysis of recent Uzbek, Chinese and multilingual corpus projects, the study identifies the typological, technological and organizational factors that determine the quality of a bilingual resource: the agglutinative structure and two-script tradition of Uzbek, the absence of word boundaries and the classifier system in Chinese, the scarcity of authentic translations, and the risks of synthetic data. Seven principles are proposed, namely representativeness and balance, authenticity and directionality, multi-level alignment, interoperable annotation, quality assurance, legal and ethical transparency, and openness and sustainability, together with a six-stage development model that integrates human expertise with embedding-based alignment tools.

Keywords: parallel corpus, Uzbek language, Chinese language, sentence alignment, corpus annotation, low-resource languages, LaBSE, machine translation, linguistic resources, corpus design.

O‘ZBEK VA XITOY TILLARI UCHUN PARALLEL KORPUSGA ASOSLANGAN LINGVISTIK RESURSLARNI ISHLAB CHIQISH TAMOYILLARI

Annotatsiya: Maqolada O‘zbekiston va Xitoy o‘rtasidagi siyosiy, iqtisodiy va ta’limiy aloqalar jadal rivojlanayotganiga qaramay, Markaziy Osiyo kompyuter lingvistikasida eng kam resursga ega til juftliklaridan biri bo‘lib qolayotgan o‘zbek–xitoy tillari uchun parallel korpusga asoslangan lingvistik resurslarni ishlab chiqish tamoyillari asoslab beriladi. O‘zbek, xitoy va ko‘p tilli korpus loyihalarini qiyosiy tahlil qilish asosida ikki tilli resurs sifatini belgilovchi tipologik, texnologik va tashkiliy omillar aniqlangan: o‘zbek tilining agglyutinativ tuzilishi va ikki yozuvli an’anasi, xitoy tilida so‘z chegaralarining yo‘qligi va hisob so‘zlari tizimi, haqiqiy tarjimalarning kamligi hamda sintetik ma’lumotlar xavfi. Reprezentativlik va muvozanat, autentiklik va tarjima yo‘nalishi, ko‘p bosqichli tekislash, o‘zaro mos annotatsiya, sifat nazorati, huquqiy-axloqiy shaffoflik, ochiqlik va barqarorlikdan iborat yetti tamoyil hamda inson ekspertizasini vektor tasvirlarga asoslangan tekislash vositalari bilan uyg‘unlashtiruvchi olti bosqichli model taklif etiladi.

Kalit so‘zlar: parallel korpus, o‘zbek tili, xitoy tili, gaplarni tekislash, korpus annotatsiyasi, kam resursli tillar, LaBSE, mashina tarjimasi, lingvistik resurslar, korpus dizayni.

ПРИНЦИПЫ РАЗРАБОТКИ ЛИНГВИСТИЧЕСКИХ РЕСУРСОВ НА ОСНОВЕ ПАРАЛЛЕЛЬНОГО КОРПУСА ДЛЯ УЗБЕКСКОГО И КИТАЙСКОГО ЯЗЫКОВ

Аннотация: В настоящей статье обоснованы принципы разработки лингвистических ресурсов на основе параллельного корпуса для узбекско-китайской языковой пары, которая, несмотря на стремительное развитие политических, экономических и образовательных связей между Узбекистаном и Китаем, остаётся одной из наименее обеспеченных ресурсами пар в компьютерной лингвистике Центральной Азии. На основе сравнительного анализа современных узбекских, китайских и многоязычных корпусных проектов выявлены типологические, технологические и организационные факторы, определяющие качество двуязычного ресурса: агглютинативный строй и двухалфавитная традиция узбекского языка, отсутствие границ слов и система счётных слов в китайском языке, дефицит аутентичных переводов и риски синтетических данных. В работе предложены семь принципов, а именно: репрезентативность и сбалансированность, аутентичность и направленность перевода, многоуровневое выравнивание, совместимая разметка, контроль качества, правовая и этическая прозрачность, открытость и устойчивость, – а также шестиэтапная модель разработки, объединяющая экспертную работу человека с инструментами выравнивания на основе векторных представлений.

Ключевые слова: параллельный корпус, узбекский язык, китайский язык, выравнивание предложений, корпусная разметка, малоресурсные языки, LaBSE, машинный перевод, лингвистические ресурсы, проектирование корпуса.

INTRODUCTION

Over the last decade, the relationship between Uzbekistan and China has moved from occasional cooperation to a dense network of trade, investment, transport and educational projects. By the end of 2024, Chinese investments in Uzbekistan exceeded 10 billion US dollars, bilateral trade turnover reached 13.1 billion dollars, and 3,357 enterprises with Chinese capital were operating in the country, which made China Uzbekistan’s largest trade and investment partner [1]. Each of these figures implies a growing volume of contracts, technical documentation, news, legal acts and educational materials that must be translated between Uzbek and Chinese.

The state language policy of Uzbekistan has, at the same time, turned the digital readiness of the Uzbek language into a strategic task. Resolution No. PQ-358 of 14 October 2024 approved the Strategy for the development of artificial intelligence technologies until 2030 and an action plan for 2024–2026, linking the widespread introduction of AI in social and economic sectors with the availability of national data sets [2]. For language technologies, the most valuable data set of this kind is a parallel corpus, that is, a collection of original texts aligned with their translations at the level of documents, paragraphs, sentences and, ideally, words.

The Uzbek–Chinese pair, however, is still in a pre-corpus stage. A study presented at the 24th China National Conference on Computational Linguistics notes that, apart from the FLORES-200 benchmark, there are practically no publicly available high-quality parallel data sets between Chinese and the languages of Central Asia [3, p. 489]. The pair is also typologically distant: Uzbek is an agglutinative Turkic language with suffixal morphology and

the subject–object–verb order, whereas Chinese is an isolating Sino-Tibetan language that writes words without spaces, expresses grammatical meaning through word order, particles and classifiers, and uses the subject–verb–object order. **The aim of this article** is to formulate a coherent system of principles for developing parallel corpus-based linguistic resources for Uzbek and Chinese and to propose a practical development model that can be implemented by Uzbek and Chinese universities with limited resources.

LITERATURE REVIEW AND METHODOLOGY

Parallel corpus studies provide a well-developed theoretical framework for this task. M. Mikhailov and R. Cooper define a parallel corpus as a collection of source texts and their translations, and show that its research value depends on decisions taken before the first text is collected: the direction of translation, the proportion of genres, the treatment of indirect translations, and the format in which metadata are stored [4].

Uzbek computational linguistics has accumulated a considerable body of experience in recent years. B. Elov and M. Amirkulov describe the creation of an Uzbek–English parallel corpus as a sequence of stages that includes text collection, cleaning, extralinguistic tagging, sentence-level alignment, validation, tokenization, lemmatization, part-of-speech tagging and word-level alignment, and they implement the linguistic stages in the web resource uznatcorpara.uz [5, p. 72, 76]. The Uzbek–Kazakh corpus compiled by B. Allaberdiev and co-authors demonstrates the scale of human effort required for a quality resource: of its 121,138 sentence pairs, more than 80% were produced by manual translation of 100,000 Uzbek sentences, each sentence was translated by at least two of the 18 bilingual students involved, and disagreements were resolved by an expert [6].

Annotation resources for Uzbek are developing in parallel. M. Sharipov and colleagues proposed a morphological and syntactic tagset for Uzbek together with a web-based annotation tool [7], while B. Mansurov and A. Mansurov released UzBERT, the first publicly available monolingual Uzbek language model, pretrained on a news corpus of about 142 million words [8]. The BBPOS benchmark has shown how much modern models depend on reliable training data: on a manually annotated set of 500 sentences (5,831 words) labelled with the Universal Dependencies part-of-speech tags, UzBERT reached 91.6% accuracy, compared with 75.6% for a rule-based tagger [9]. A specific Uzbek problem is the coexistence of the Cyrillic alphabet and the Latin alphabet, which officially replaced Cyrillic in 1993, as well as the newly announced revised Latin alphabet; U. Salaev, E. Kuriyozov and C. Gómez-Rodríguez addressed it with an open-source transliteration tool that supports all three writing systems [10].

On the Chinese side, corpus infrastructure is mature. The BCC corpus of Beijing Language and Culture University contains texts amounting to tens of billions of characters and offers a formal query language that combines characters, words and grammatical tags [11]. Because Chinese writing does not mark word boundaries, any word-level processing of Chinese begins with segmentation; the PKUSEG toolkit demonstrated that domain-specific segmentation models (news, medicine, tourism, web texts) are noticeably more accurate than a single general model [12]. The Universal Dependencies framework, which now provides consistent morphosyntactic annotation for more than 100 languages, offers a common tagset that can link Uzbek and Chinese annotation layers despite their typological distance [13].

Multilingual machine translation research supplies both evaluation data and alignment methods. The No Language Left Behind project released FLORES-200, a benchmark of 3,001 professionally translated sentences in 204 language varieties, which includes Northern Uzbek in Latin script (uzn_Latn) and Simplified Chinese (zho_Hans) [14]. Language-agnostic sentence embeddings such as LaBSE, which covers 109 languages, map sentences with the same meaning in different languages to nearby points in a vector space [15]. Vecalign uses such embeddings to align sentences in linear time and space and outperformed the previous state-of-the-art method by 5 F1 points [16], while Bertalign, developed for Chinese–English literary texts, handles one-to-many and many-to-many correspondences that are typical of translations between typologically distant languages [17]. Open infrastructures such as OPUS-MT, which provided more than 1,000 pretrained translation models trained on openly available parallel data, show that corpora become most useful when they are released together with tools and models [18].

Finally, the first large-scale attempt to build a Chinese–Central Asian corpus is directly relevant to our topic. Q. Yuan and A. Murat crawled monolingual news from official Kazakh, Uzbek and Tajik sources, translated it into Chinese with a large language model and filtered the pseudo-parallel pairs with LaBSE using a cosine-similarity threshold and a margin score [3, p. 492–493]. For the Uzbek–Chinese pair, 35,989 candidate pairs were reduced to 29,048 after filtering (19.3% removed), and LaBSE distinguished parallel from non-parallel pairs with an F1 score of 0.991, against 0.956 for LASER and only 0.157 for multilingual BERT [3, p. 494–495]. These results confirm the reliability of embedding-based filtering, but they also show the limitation of synthetic data: the Chinese side of such a corpus is machine output rather than authentic human translation.

Methodology. The study is qualitative and design-oriented. It combines (a) a comparative analysis of the corpus projects listed above with respect to data sources, alignment methods, annotation layers, quality control and licensing; (b) a contrastive typological analysis of Uzbek and Chinese aimed at identifying units that cause alignment and annotation difficulties; and (c) modelling, in which the identified requirements are generalized into principles and arranged into a sequence of development stages.

RESULTS AND DISCUSSION

The comparative analysis showed that the main difficulties of an Uzbek–Chinese corpus are concentrated at three levels. **At the level of units**, one Uzbek word form often corresponds to several Chinese words, and vice versa. In the sentence Talabalar kutubxonada uchta kitob o‘qishdi and its translation 学生们在图书馆读了三本书 (“The students read three books in the library”), the locative suffix in kutubxona-da corresponds to the preposition-like word 在 placed before 图书馆; the numeral uchta corresponds to the numeral 三 plus the classifier 本, which has no counterpart in Uzbek; and the past-tense suffix of o‘qi-shdi corresponds to the verb 读 followed by the aspect particle 了. **At the level of order**, the verb closes the Uzbek sentence but follows the subject in Chinese, so word alignment links cross each other. **At the level of text**, translators frequently split long Uzbek sentences with chains of participles and converbs into two or three Chinese sentences, which produces one-to-many sentence correspondences.

These observations, together with the experience of the projects reviewed, allow us to formulate seven principles for developing Uzbek–Chinese parallel corpus-based resources.

1. The principle of representativeness and balance. A resource intended for general use must reflect the real spheres of Uzbek–Chinese communication rather than the texts that are easiest to download. News and official texts dominate the existing data [3], whereas trade contracts, technical standards, education, tourism, medicine, fiction and the language of Chinese-language learners in Uzbekistan are almost absent. We therefore recommend a target genre matrix with explicit quotas, for example 30% socio-political and legal texts, 25% economic and technical texts, 20% educational and scientific texts, 15% fiction and 10% spoken and web texts.

2. The principle of authenticity and directionality. Every text pair must be accompanied by metadata on the original language, the translation direction, the translator, the date and whether the text was translated directly or through Russian or English. In Uzbek–Chinese practice, mediated translation through Russian or English is common, and if such texts are mixed with direct translations, contrastive conclusions become unreliable [4]. Machine-generated pairs, which are useful for training models, should be stored in a separate, clearly labelled subcorpus and never presented as human translation.

3. The principle of multi-level alignment. Alignment should proceed from documents to paragraphs, sentences and words, with each level verified before the next one begins. At the sentence level, embedding-based aligners such as Vecalign and Bertalign are preferable because they support one-to-many and many-to-many correspondences [16; 17], and LaBSE scores can be used to rank candidate pairs [15]. Word-level alignment should be carried out not between Uzbek word forms and Chinese characters but between Uzbek morphemes or lemmas and segmented Chinese words, which is consistent with the example given above.

4. The principle of interoperable annotation. Uzbek texts should be normalized to the official Latin alphabet with the original script preserved in metadata [10], tokenized and annotated for lemma, part of speech and morphological features; Chinese texts should be segmented with domain-appropriate models [12] and annotated with part-of-speech tags. Both layers should be mapped onto the Universal Dependencies tagset [13], while language-specific categories such as Uzbek case and possessive suffixes or Chinese classifiers and aspect particles should be kept as additional features. Storage in XML (TEI) or JSON together with the TMX format used by translation memory systems ensures compatibility with corpus managers and computer-assisted translation tools.

5. The principle of quality assurance. Automatic filtering should combine length-ratio checks, language identification, duplicate removal and embedding similarity with thresholds selected on a manually verified development set [3]. Human validation remains indispensable: following the Uzbek–Kazakh experience, at least two qualified bilinguals should check each manually translated or post-edited segment, and inter-annotator agreement should be reported [6].

6. The principle of legal and ethical transparency. Copyright, licence and personal data issues must be resolved before texts are collected. Publicly funded official documents and openly licensed materials can be distributed in full, whereas copyrighted fiction may be provided as shuffled sentence pairs, as was done for the books in the Uzbek–Kazakh corpus [6], or as a searchable but non-downloadable collection.

7. The principle of openness and sustainability. The resource should be versioned, documented and released with persistent identifiers under clear licences, together with the scripts used to build it. Open releases allow the corpus to be evaluated on FLORES-200 [14], reused for training translation models in the manner of OPUS-MT [18], and extended by other universities.

On the basis of these principles, the following model for developing an Uzbek–Chinese parallel corpus-based resource is proposed:

1. Design – defining users and tasks, the genre matrix, the metadata scheme and the licensing policy.

2. Collection – gathering authentic Uzbek–Chinese translations and, separately, monolingual texts for manual or machine translation.

3. Pre-processing – encoding normalization (UTF-8), script normalization of Uzbek, Chinese segmentation, sentence splitting and deduplication.

4. Alignment – document, paragraph and sentence alignment with embedding-based aligners, followed by word alignment of lemmas and segmented words.

5. Annotation and validation – morphological and part-of-speech annotation mapped to Universal Dependencies, double human checking and agreement measurement.

6. Release and application – publication with documentation and licences, query interface, evaluation on FLORES-200 and use in lexicography, terminology, teaching and machine translation.

The model differs from purely automatic pipelines in two respects. **First**, it treats synthetic pairs produced by large language models as an auxiliary layer: they can enlarge training data for machine translation, but authentic human translations remain the core of the corpus and the basis of linguistic research. **Second**, it places annotation and validation between alignment and release, so that the resource can serve linguists, translators and teachers, not only developers of translation systems. At the same time, the Uzbek–Kazakh experience shows that the decisive factor is organized human work: a team of trained bilingual students under expert supervision produced more than 100,000 sentence pairs within two months [6].

CONCLUSION

The Uzbek–Chinese language pair combines high practical demand with a near absence of open, high-quality parallel data. The analysis has shown that the typological distance between agglutinative Uzbek and isolating Chinese, the two-script tradition of Uzbek, the lack of word boundaries in Chinese and the shortage of authentic direct translations require a special methodology rather than the mechanical transfer of solutions developed for other languages.

The seven principles proposed in the article, namely representativeness and balance, authenticity and directionality, multi-level alignment, interoperable annotation, quality assurance, legal and ethical transparency, and openness and sustainability, together with the six-stage development model, form a methodological basis for building Uzbek–Chinese parallel corpus-based resources. Further research will focus on creating a pilot corpus of authentic Uzbek–Chinese translations, developing a morpheme-level word alignment scheme and evaluating the corpus in machine translation experiments.

REFERENCES

1. Uzbekistan reveals volume of Chinese investments in country [Electronic resource] // Trend News Agency. – 2025. – 20 January. – URL: https://www.trend.az/casia/uzbekistan/3995236.html (accessed: 01.11.2026).
2. Resolution of the President of the Republic of Uzbekistan No. PQ-358 of 14.10.2024 “On approval of the Strategy for the development of artificial intelligence technologies until 2030” [Electronic resource]. – URL: https://lex.uz/docs/7159258 (accessed: 01.11.2026).
3. Yuan Q., Murat A. **Method for Constructing Parallel Corpora for Central Asian Low-Resource Languages Based on Large Language Models and Cross-Lingual Embeddings** // Proceedings of the 24th China National Conference on Computational Linguistics (CCL 2025). – Jinan: Chinese Information Processing Society of China, 2025. – P. 488–499. – URL: https://aclanthology.org/2025.ccl-1.36/ (accessed: 01.11.2026).
4. Mikhailov M., Cooper R. **Corpus Linguistics for Translation and Contrastive Studies: A Guide for Research.** – London; New York: Routledge, 2016. – 258 p.
5. Elov B.B., Amirkulov M. **Uzbeki-English Parallel Corpus Algorithm and Alignment Problem** // Central Asian Journal of Literature, Philosophy and Culture. – 2023. – Vol. 4, Issue 6. – P. 71–79.
6. Allaberdiev B., Matlatipov G., Kuriyozov E., Rakhmonov Z. **Parallel texts dataset for Uzbek-Kazakh machine translation** // Data in Brief. – 2024. – Vol. 53. – Article 110194. – DOI: 10.1016/j.dib.2024.110194.
7. Sharipov M., Mattiev J., Sobirov J., Baltayev R. **Creating a morphological and syntactic tagged corpus for the Uzbek language** // arXiv preprint arXiv:2210.15234. – 2022. – URL: https://arxiv.org/abs/2210.15234 (accessed: 01.11.2026).
8. Mansurov B., Mansurov A. **UzBERT: pretraining a BERT model for Uzbek** // arXiv preprint arXiv:2108.09814. – 2021. – URL: https://arxiv.org/abs/2108.09814 (accessed: 01.11.2026).
9. Bobojonova L., Akhundjanova A., Ostheimer P., Fellenz S. **BBPOS: BERT-based Part-of-Speech Tagging for Uzbek** // arXiv preprint arXiv:2501.10107. – 2025. – URL: https://arxiv.org/abs/2501.10107 (accessed: 01.11.2026).
10. Salaev U., Kuriyozov E., Gómez-Rodríguez C. **A machine transliteration tool between Uzbek alphabets** // CEUR Workshop Proceedings. – 2022. – Vol. 3315. – P. 42–50.
11. Xun E., Rao G., Xiao X., Zang J. **The construction of the BCC corpus in the age of big data [荀恩东, 饶高琦, 肖晓悦, 臧娇娇. 大数据背景下BCC语料库的研制]** // Corpus Linguistics [语料库语言学]. – 2016. – Vol. 3, No. 1. – P. 93–109.
12. Luo R., Xu J., Zhang Y., Ren X., Sun X. **PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation** // arXiv preprint arXiv:1906.11455. – 2019. – URL: https://arxiv.org/abs/1906.11455 (accessed: 01.11.2026).
13. de Marneffe M.-C., Manning C.D., Nivre J., Zeman D. **Universal Dependencies** // Computational Linguistics. – 2021. – Vol. 47, No. 2. – P. 255–308. – DOI: 10.1162/coli_a_00402.
14. NLLB Team, Costa-jussà M.R., Cross J. et al. **No Language Left Behind: Scaling Human-Centered Machine Translation** // arXiv preprint arXiv:2207.04672. – 2022. – URL: https://arxiv.org/abs/2207.04672 (accessed: 01.11.2026).

15. 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.
16. Thompson B., Koehn P. **Vecalign: Improved Sentence Alignment in Linear Time and Space** // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). – Hong Kong, 2019. – P. 1342–1348. – DOI: 10.18653/v1/D19-1136.
17. Liu L., Zhu M. **Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts** // Digital Scholarship in the Humanities. – 2023. – Vol. 38, No. 2. – P. 621–634. – DOI: 10.1093/llc/fqac089.
18. 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.
19. Internet sources
20. https://aclanthology.org/
21. https://uznatcorpara.uz/