НОВАЯ ЭРА НАУКИ И ЗНАНИЙ
МЕЖДУНАРОДНЫЙ НАУЧНЫЙ ЖУРНАЛ
global-science-review.com
editor@global-science-review.com

A TYPOLOGY-INFORMED METHODOLOGICAL MODEL FOR BUILDING AN UZBEK–CHINESE PARALLEL CORPUS

Zilola Ma'murjon qizi Soliyeva[1], S. T. Mustafayeva[2]

[1] Master's student, Tashkent State University of Oriental Studies, Tashkent, Uzbekistan

[2] DSc, Associate Professor (Docent), Tashkent State University of Oriental Studies, Tashkent, Uzbekistan

Corresponding author e-mail: soliyeva481@gmail.com
ORCID:0009000688484645

Abstract

Uzbek and Mandarin Chinese lie at opposite ends of the morphological spectrum: Uzbek is an agglutinative Turkic language with SOV order and a rich case system, whereas Chinese is an isolating Sino-Tibetan language with SVO order, function words and numeral classifiers, written without word boundaries. Our search did not identify a systematic, openly available Uzbek–Chinese parallel corpus; the only directly relevant study found builds a filtered, partly LLM-generated pseudo-parallel dataset from news crawls. This paper proposes a typology-informed methodological model for building such a corpus. It has five components: (1) a layered design separating a human-verified gold core, an automatic extension and a labelled synthetic layer, with provenance metadata; (2) asymmetric linguistic processing linked through Universal Dependencies; (3) multi-level alignment down to the morpheme, with correspondence-type labels; (4) a "unit–set of equivalents" lexical database; and (5) retrieval-augmented integration with large language models. A correspondence table and two author-composed sentence pairs show that most grammatical correspondences are not one-to-one. An evaluation protocol is specified for future empirical work.

Keywords: parallel corpus; Uzbek; Chinese; linguistic typology; sentence and word alignment; low-resource machine translation; retrieval-augmented generation; corpus annotation

1. Introduction

Parallel corpora underpin contrastive linguistics, translation studies, bilingual lexicography and machine translation (MT). Their construction methodology has matured from length-based sentence alignment (Gale & Church, 1993) to aligners that operate on multilingual sentence embeddings (Thompson & Koehn, 2019; Feng et al., 2022).

The Uzbek–Chinese pair is in a very different position. Searches in English, Chinese, Uzbek and Russian did not identify a systematic, openly available Uzbek–Chinese parallel corpus. The only directly relevant work we found is the study by Yuan and Murat (2025), which builds Uzbek–Chinese, Kazakh–Chinese and Tajik–Chinese data from crawled news sites and LLM-generated pseudo-parallel pairs. Meanwhile, large language models (LLMs) still lag behind dedicated MT systems for most lower-resourced languages (Robinson et al., 2023), and a 2025 presidential decree requires an “Uzbek language corpus” to be created by the end of 2026 and introduced into at least five open LLMs (Decree No. PF-189, 2025). Curated bilingual data are needed both for training and as verifiable reference material.

Such a corpus cannot simply reuse a Chinese–English pipeline. In Uzbek one word form can carry several grammatical meanings, each expressed by its own suffix; in Chinese the same meanings are spread over separate function words, written without spaces. An aligner that ignores this asymmetry produces links that are formally correct but linguistically opaque.

This paper therefore proposes a typology-informed methodological model for an Uzbek–Chinese parallel corpus. Its contributions are a five-component corpus model, a grammatical correspondence table that defines the inventory of alignment

labels, an illustrative analysis of author-composed sentence pairs and an evaluation protocol. The paper is methodological, and no experimental results are reported.

2. Related Work

2.1 Parallel corpora involving Chinese

Among Chinese–English projects, Sun et al. (2002) targeted more than 500,000 sentence pairs with a fixed genre distribution (news 10%, literature 30%, government reports 25%, science and technology 35%); their length-based alignment with a lexicon check reached about 95% overall precision and 98.83% for 1:1 alignments. In the Peking University corpus, automatic alignment was checked manually with an accuracy above 98%, and texts carry global and structural metadata (Chang, 2004). The Russian–Chinese corpus of the Russian National Corpus, the closest analogue for pairing Chinese with a morphologically rich language, comprises roughly 4.5 million words in more than 1,000 documents and adds markup specific to Chinese texts (*Русско-китайский корпус НКРЯ*, n.d.). These projects show three practices that matter for a new pair: explicit text-level metadata, controlled genre composition and manual verification of automatic alignment.

2.2 Uzbek and Uzbek–Chinese resources

Basic Uzbek tools exist, each validated on small data: a rule-based stemmer with 454 affixes in 15 classes (Sharipov & Yuldashov, 2022), the first Universal Dependencies treebank for Uzbek with 500 sentences (Akhundjanova & Talamo, 2025) and a transliteration tool between Uzbek alphabets with micro-F1 between 0.89 and 0.94 (Salaev et al., 2022). For Kazakh, KazParC shows that a multilingual corpus built with human translators is feasible (Yeshpanov et al., 2024).

Yuan and Murat (2025) crawled official news sites of Central Asian countries and generated pseudo-parallel pairs with DeepSeek. Filtering by LaBSE cosine similarity and margin score reduced the Uzbek–Chinese set from 35,989 to 29,048 pairs (19.3% removed), and they reported a gain of 0.65 BLEU over back-

translation. Their focus is MT training data; a linguistically annotated, human-verified resource for the pair is still lacking. Uzbek–Chinese contrastive studies are not corpus-based but supply hypotheses for annotation: Yo‘ldoshev (2025) compares SVO and SOV word order, Qurolova (2022) analyses verb–object order and Uzbek learners’ errors with 把被sentences, and Turniyozova (2022) examines the markers 被 受 挨and 让and their Uzbek equivalents.

2.3 LLMs and retrieval-augmented translation

NLLB covers 200 languages and about 40,000 translation directions (NLLB Team, 2024), whereas general-purpose LLMs performed below traditional MT for 84.1% of the 204 languages tested (Robinson et al., 2023). Retrieval-augmented generation (RAG) combines a model’s parametric memory with a non-parametric store queried at inference time (Lewis et al., 2020), and adding bilingual dictionary hints to prompts yielded roughly 0.9–1.1 BLEU in low-resource translation (Ghazvininejad et al., 2023). For an under-resourced pair, a parallel corpus is thus valuable twice: as training data and as retrievable evidence.

3. Typological Challenges of the Uzbek–Chinese Pair

Pinyin follows standard dictionary forms, including neutral tones; tone sandhi is not marked.

Agglutinative versus isolating morphology. Uzbek stacks suffixes, each usually carrying a single meaning, whereas Chinese words are largely invariant and grammatical relations are expressed by word order and function words. The single word *kitob-lar-imiz-dan* (book-PL-POSS.1PL-ABL) corresponds to 从我们的书中 (cóng wǒmen de shū zhōng) “from among our books”, five Chinese words; the plural *-lar* has no overt counterpart.

Script and word boundaries. Uzbek texts circulate in Latin and Cyrillic script, and the letters *o‘* and *g‘* are typed with several apostrophe-like characters (‘ ’ ' ‘ ’), which breaks exact-match search. Chinese texts may be simplified or traditional (书/書 shū “book”) and have no spaces, so segmentation is a task in its own right:

研究生命科学 is correctly segmented as 研究/生命/科学 (yánjiū/shēngmìng/kēxué) “study life sciences” but can be wrongly split as 研究生/命/科学 (yánjiūshēng/mìng/kēxué) “graduate student/fate/science”. Transcribed Uzbek names such as 乌兹别克斯坦 (Wūzībiékèsītǎn) “Uzbekistan” are likely to be missing from general segmentation lexicons.

SOV versus SVO. In Chinese the verb precedes the object, whereas the Uzbek predicate is clause-final (Yo‘ldoshev, 2025), so alignment links regularly cross and position-based heuristics are unreliable.

Case versus prepositions and localizers. The locative *-da* corresponds to the discontinuous 在… 上/里, as in 在桌子上 (zài zhuōzi shang) “on the table”; the ablative *-dan* to 从 (cóng) for a source and to 比 (bǐ) in comparisons; and the accusative *-ni* either to zero or to the 把 (bǎ) construction (Qurolova, 2022).

Aspect and classifiers. Uzbek tense–aspect suffixes (*-di*, *-gan*, *-yap*) have no one-to-one counterparts among the Chinese markers 了 (le), 过 (guo), 着 (zhe) and the progressive 在 (zài), and Chinese often expresses past time lexically, e.g. 昨天 (zuótiān) “yesterday”. Chinese also requires a classifier between numeral and noun: 三本书 (sān běn shū) “three books”. Uzbek uses the counting suffix *-ta* (*uch-ta kitob*), so 本 may align with *-ta*, while 个 (ge) often has no Uzbek counterpart.

Phraseology, realia and lacunae. Chengyu (成语 chéngyǔ) must be aligned as units. 一石二鸟 (yī shí èr niǎo), literally “one stone, two birds”, has a functional Uzbek equivalent with a different image, *bir o‘q bilan ikki quyonni urmoq* “to hit two hares with one bullet”, whereas 画蛇添足 (huà shé tiān zú) “to draw a snake and add feet” has no idiomatic counterpart. Realia such as Uzbek *mahalla* or Chinese 红包 (hóngbāo) “red envelope with money” lack direct equivalents. Aligning such items word by word creates spurious dictionary entries.

4. The Proposed Model

Figure 1 shows how the model turns these challenges into design decisions: parallel document-to-token hierarchies with language-specific annotation, a central

alignment layer linking them at every level, a shared semantic layer for named entities, idioms, realia and lacunae, and outputs such as a parallel concordance, a lexical database, a translation memory and MT/LLM training data.

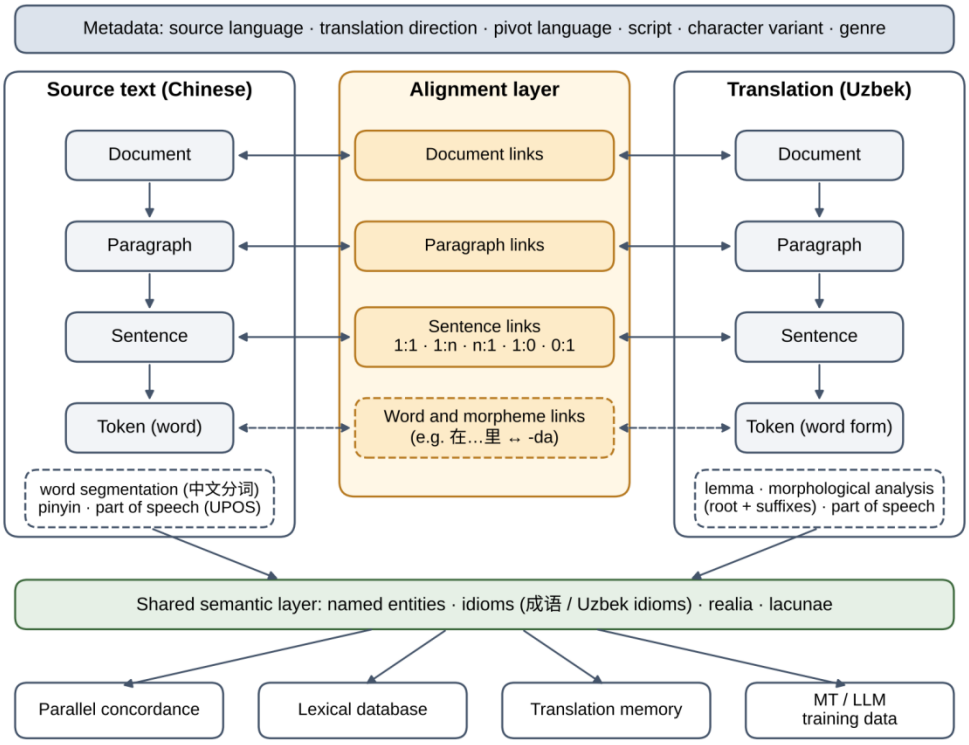


Figure 1. Conceptual model of the Uzbek–Chinese parallel corpus

4.1 Layered corpus design and metadata

The corpus has three physically separated layers: a *gold core* of human-verified sentence pairs, the only source of test sets and lexicon evidence; an *extension* of automatically aligned pairs that passed the confidence thresholds; and a *synthetic layer* of MT- or LLM-generated pairs, labelled provenance=synthetic and never merged into the gold core. The corpus is bidirectional (uz→zh and zh→uz). Mandatory document metadata record the translation direction, the pivot language (for indirect translations via Russian or English), the provenance, the Uzbek script and the Chinese character variant, together with genre, rights status and translator. Missing values are recorded explicitly as “unknown” rather than left empty or inferred. Metadata are stored in the TEI P5 header.

4.2 Asymmetric linguistic processing

Each side is processed to the depth its structure requires. Uzbek text is normalised to one apostrophe character, transliterated from Cyrillic to Latin with the original kept (Salaev et al., 2022), stemmed and lemmatised (Sharipov & Yuldashov, 2022), tagged and given a separate morpheme-segmentation layer (*kitob-lar-imiz-dan*), because morphemes are the units of word-level alignment. Chinese text is converted from traditional to simplified characters with the original preserved, segmented with a growing user lexicon of transcribed Uzbek names, and given word-level pinyin, since polyphonic characters are disambiguated only in context (行 xíng “walk” vs. háng “row; trade”). Both sides are mapped onto Universal Dependencies UPOS tags and features (de Marneffe et al., 2021) in CoNLL-U; a classifier flag in MISC compensates for the lack of a classifier UPOS tag.

4.3 Multi-level alignment with correspondence-type labels

Length-based alignment (Gale & Church, 1993) fits this pair poorly, because a Chinese character usually corresponds to one syllable and often to one morpheme, whereas an Uzbek word contains several of each. Sentence alignment therefore combines LaBSE embeddings (Feng et al., 2022) with Vecalign (Thompson & Koehn, 2019) and dictionary anchors such as numbers, dates and transcribed names. Chinese commas joining independent clauses are secondary boundary candidates, so one Chinese sentence can link to several Uzbek sentences. A confidence score routes each link to automatic acceptance, human review or a rejected log, with thresholds calibrated on the gold core. Word-level alignment starts with SimAlign, which needs no parallel training data (Jalili Sabet et al., 2020), and operates on Uzbek morphemes rather than whitespace tokens.

Every link carries a correspondence-type label. Ratios are always given in Uzbek:Chinese order: 1:1, 1:n, n:1, 1:0 (Uzbek material without a Chinese counterpart) and 0:1 (Chinese material without an Uzbek counterpart). Further flags

mark word↔affix links and discontinuous units such as 在…上 The inventory is derived from the correspondences in Table 1, few of which are plain 1:1.

Table 1. Cross-lingual grammatical correspondences between Uzbek and Chinese

Category	Uzbek means	Chinese means	Example (pinyin)	Correspondence type
Word order	SOV; predicate clause-final	SVO; verb before object	Men kitob o‘qiyman ↔ 我看书(Wǒ kàn shū) “I read books”	1:1 links, crossing
Locative	-da; postpositional noun ust(i)da	在+ noun + 上/里	stol ustida ↔ 在桌子上(zài zhuōzi shang) “on the table”	1:n, discontinuous
Ablative	-dan	从(source), 比 (comparison)	kutubxonadan ↔ 从图书馆(cóng túshūguǎn) “from the library”	word↔affix; reversed order
Accusative	-ni (definite object)	zero, or 把before the object	kitobni ↔ 把书(bǎ shū)	word↔affix (functional) or 1:0
Tense–aspect	-di, -gan, -yap	了, 过, 着, 在 lexical time words	Xitoyda bo‘lganman ↔ 我去过中国 (Wǒ qùguo Zhōngguó) “I have been to China”	1:n; word↔affix
Numeral classifier	-ta; optional dona, nafar	obligatory 量词: 本, 个, 张	uchta kitob ↔ 三本书(sān běn shū) “three books”	本↔ -ta word↔affix; 个0:1
Passive	-il-/-in-/-n-	被before the verb	o‘g‘irlandi ↔ 被偷了 (bèi tōu le) “was stolen”	1:n; word↔affix

Source: compiled by the authors on the basis of Yo‘ldoshev (2025), Qurolova (2022) and Turniyozova (2022); examples composed by the authors.

4.4 Lexical database model

Because correspondences are often one-to-many, the lexicon follows a “unit–set of equivalents” model rather than a “word–word” model. Each Uzbek lemma or Chinese word is linked to corpus-attested equivalents that record characters and tone-marked pinyin, part of speech, frequency by genre, supporting sentence pairs and an *equivalence degree*: *full* (*hamkorlik* ↔ 合作 hézuò “cooperation”), *partial* when usage or cultural load differs (*chuchvara* ↔ 饺子jiǎozi “dumplings”), or *zero*, when translators resort to description, transcription or calque. Classifiers, idioms, realia, terminology and function-word correspondences have separate layers. A status field (candidate, approved or rejected) ensures that only approved entries are used downstream.

4.5 Integration with LLMs and RAG

The corpus serves as non-parametric memory for LLM-based translation (Lewis et al., 2020). For a source sentence, a hybrid lexical and dense retriever

returns verified parallel examples, approved glossary entries and short grammar notes derived from Table 1, for example that a numeral requires a classifier. The evidence is placed in separate prompt sections, in the spirit of dictionary-based prompting (Ghazvininejad et al., 2023), and the output passes a terminology check and human post-editing before an approved pair returns to the corpus with its provenance label. The risks of LLM-generated data (hallucination, translationese and test-set leakage) are countered by provenance labels, the separate synthetic layer, test sets built only from human translations and deduplication of all stores against the test sets.

5. Illustrative Analysis

The two pairs below were composed by the authors for illustration and are not corpus data. Uzbek is segmented into morphemes; ratios are Uzbek:Chinese.

Example 1. *Men kecha kutubxona-dan ikki-ta kitob ol-di-m* (I yesterday library-ABL two-CNT book take-PST-1SG) ↔ **我昨天从**图书馆借了两本书。(Wǒ zuótiān cóng túshūguǎn jièle liǎng běn shū.) “Yesterday I borrowed two books from the library.” The links *men* ↔ **我**, *kecha* ↔ **昨天** and *kitob* ↔ 书 are 1:1. *Kutubxonadan* ↔ **从**图书馆 is 1:n at word level and splits at morpheme level into *kutubxona* ↔ 图书馆 (1:1) and *-dan* ↔ **从** (word↔affix, in reverse order). *Ikkita* ↔ **两本** splits likewise, with *-ta* ↔ the classifier **本**; ’two’ is **两** (liǎng), the form used before classifiers, not **二** (èr). *Oldim* ↔ **借了** is also 1:n, combining a partial lexical equivalence (*ol-* “take” vs. **借** “borrow”), *-di* ↔ **了** (word↔affix) and *-m* (1SG) ↔ ∅ (1:0), since person is already expressed by **我** The SOV/SVO difference makes the links cross.

Example 2. *Velosiped-im o‘g‘irla-n-di* (bicycle-POSS.1SG steal-PASS-PST) ↔ **我的自行车**被偷了。(Wǒ de zìxíngchē bèi tōu le.) “My bicycle was stolen.” The single suffix *-im* corresponds to two words, **我的** (1:n). Inside the verb, *-n-* ↔ **被** and *-di* ↔ **了** are word↔affix links, while the root *o‘g‘irla-* ↔ 偷 is 1:1. Two Uzbek words thus map to six Chinese words.

Taken together, and counting ʔ as a separate word, 5 of the 8 word-level links in these two pairs are not 1:1. After morpheme segmentation, every Uzbek lexical root aligns 1:1, and every grammatical morpheme falls into the word↔affix, 1:n or 1:0 types. It is the grammatical material, not the vocabulary, that requires morpheme-level alignment and typed links.

6. Evaluation Protocol

***Annotation*:** two annotators independently align and label a stratified gold-core sample, a third expert adjudicates, and Cohen's kappa is reported separately for link type, equivalence degree and the semantic layers. *Processing*: Chinese segmentation F1 and Uzbek lemmatisation and tagging accuracy are measured against gold samples. *Alignment*: sentence alignment is scored by precision, recall and F1 per link type, and word or morpheme alignment by AER (Och & Ney, 2003), reported separately for word↔affix and discontinuous links. *Translation*: both directions are evaluated on the FLORES-200 devtest (uzn_Latn, zho_Hans) and on held-out gold-core sentences, using sacreBLEU with full signatures (Post, 2018), chrF (Popović, 2015), COMET (Rei et al., 2020) and human error annotation. NLLB as a dedicated MT baseline is compared with an LLM without retrieval and with RAG using examples, glossary entries and grammar notes.

7. Discussion

This work shows how the typological distance between the two languages translates into specific design decisions. Asymmetric processing, morpheme-level alignment and correspondence-type labels make visible what a generic pipeline would hide, and the same labels serve as query filters for contrastive research, constraints on lexicon extraction and grammar notes for retrieval-augmented translation. The layered design allows LLM-generated data to be used without contaminating the gold standard.

Limitations are clear. No empirical evaluation has been carried out, so the thresholds, tool choices and expected benefits of retrieval remain hypotheses to be

tested with the protocol in Section 6. Direct human Uzbek–Chinese translations are scarce, and many go through Russian or English, which limits the size of the gold core. Copyright may restrict parts of the corpus to concordance views. The Uzbek tools have been validated on small datasets, and the correspondence table is not exhaustive. Future work will build a pilot gold core, evaluate segmentation and alignment against it, and release the annotation guidelines under an open licence.

8. Conclusion

This paper has proposed a typology-informed methodological model for an Uzbek–Chinese parallel corpus, a pair for which our search found no systematic open resource. The correspondence table and the illustrative analysis indicate that lexical roots tend to align one-to-one, whereas grammatical correspondences do not, which calls for morpheme-level alignment on the Uzbek side. The evaluation protocol defines how these claims can be tested empirically.

References

Akhundjanova, A., & Talamo, L. (2025). Universal Dependencies treebank for Uzbek. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025)* (pp. 1–6). University of Tartu Library.

Chang, B. (2004). Chinese-English parallel corpus construction and its application. In *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation (PACLIC 18)* (pp. 283–290). Waseda University.

Decree No. PF-189 of the President of the Republic of Uzbekistan “Sun’iy intellekt texnologiyalarini yanada rivojlantirishga oid qo‘shimcha chora-tadbirlar to‘g‘risida” [On additional measures for the further development of artificial intelligence technologies]. (2025, October 22). https://lex.uz/uz/docs/-7789403

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 547(2), 255–308. https://doi.org/10.1162/coli_a_00402

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 878–891). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.62

Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, *19*(1), 75–102.

Ghazvininejad, M., Gonen, H., & Zettlemoyer, L. (2023). *Dictionary-based phrase-level prompting of large language models for machine translation* (arXiv:2302.07856) [Preprint]. arXiv.

Jalili Sabet, M., Dufter, P., Yvon, F., & Schütze, H. (2020). SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1627–1643). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.147

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)* (pp. 9459–9474).

NLLB Team. (2024). Scaling neural machine translation to 200 languages. *Nature*, *630*(8018), 841–846. https://doi.org/10.1038/s41586-024-07335-x

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–51. https://doi.org/10.1162/089120103321337421

Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 392–395). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3049

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 186–191). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319

Qurolova, S. T. (2022). A comparative analysis of verb-object word order in Chinese and Uzbek. *Oriental Renaissance*, 256–259. https://oriens.uz/en/conference/article/a-comparative-analysis-of-verbobject-word-order-in-chinese-and-uzbek/

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2685–2702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Robinson, N., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023). ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 392–418). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.40

Русско-китайский корпус НКРЯ, или Как превратить ботвинью в гаспачо [The Russian–Chinese corpus of the RNC, or how to turn botvinya into gazpacho]. (n.d.). Системный Блокъ. Retrieved October 6, 2026, from https://sysblok.ru/oriental/russko-kitajskij-korpus-nkrja-ili-kak-prevratit-botvinju-v-gaspachcho/

Salaev, U., Kuriyozov, E., & Gómez-Rodríguez, C. (2022). A machine transliteration tool between Uzbek alphabets. In *ALTNLP 2022: CEUR Workshop Proceedings* (Vol. 3315, pp. 42–50).

Sharipov, M., & Yuldashov, O. (2022). UzbekStemmer: Development of a rule-based stemming algorithm for Uzbek language. In *ALTNLP 2022: CEUR Workshop Proceedings* (Vol. 3315, pp. 137–144).

Sun, L., Xue, S., Qu, W., Wang, X., & Sun, Y. (2002). Constructing of a large-scale Chinese-English parallel corpus. In *COLING-02: The 3rd Workshop on Asian Language Resources and International Standardization*.

Thompson, B., & Koehn, P. (2019). Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 1342–1348). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1136

Turniyozova, M. Ya. (2022). Xitoy va o‘zbek tillarida kauzativlikni ifodalanishi [The expression of causativity in Chinese and Uzbek]. *Oriental Renaissance*, 77–82. https://oriens.uz/en/conference/article/xitoy-va-ozbek-tillarida-kauzativlikni-ifodalanishi/

Yeshpanov, R., Polonskaya, A., & Varol, H. A. (2024). KazParC: Kazakh parallel corpus for machine translation. In *Proceedings of LREC-COLING 2024* (pp. 9633–9644). ELRA and ICCL.

Yo‘ldoshev, E. (2025). Xitoy va o‘zbek tillarida so‘z tartibi: Sintaktik strukturalarning qiyosiy tahlili [Word order in Chinese and Uzbek: A comparative analysis of syntactic structures]. *Science and Education*, *6*(10), 707–711.

Yuan, Q. [袁琦, & Murat, A. [阿力木·木拉提. (2025). 基于LLM与跨语言嵌入的中亚低资源语言平行语料库构建方法 [A method for constructing parallel corpora for low-resource Central Asian languages based on LLMs and cross-lingual embeddings]. In *Proceedings of the 24th China National Conference on Computational Linguistics (CCL 2025)* (pp. 488–499). Chinese Information Processing Society of China. https://aclanthology.org/2025.ccl-1.36/