Ilmiy tadqiqotlar va ularning yechimlari jurnali
Aoka guvohnoma № 323759 ISSN: 3060-4818

THEORETICAL AND METHODOLOGICAL FOUNDATIONS OF DEVELOPING UZBEK–CHINESE PARALLEL CORPORA: LINGUISTIC ALIGNMENT, TRANSLATION EQUIVALENCE, AND CROSS-LINGUISTIC STRUCTURAL DIFFERENCES

Xiaoqi You
National University of Uzbekistan, PhD
E-mail: 726366681@qq.com
https://doi.org/10.5281/zenodo.23858569

Abstract: The article examines the theoretical and methodological foundations of building Uzbek–Chinese parallel corpora. Although parallel corpora have become a core resource for translation studies, contrastive linguistics and machine translation, the Uzbek–Chinese language pair remains poorly resourced. The study analyzes three interrelated problems: the alignment of texts at the document, paragraph, sentence and word levels; the identification of translation equivalence; and the structural differences between an agglutinative language with subject–object–verb order and an isolating language with subject–verb–object order. Drawing on recent work on Uzbek corpus development and on embedding-based word alignment, the article proposes a multi-level framework for corpus construction that combines automatic alignment with linguistically informed annotation and human verification. The framework is intended to support contrastive research, translator training and the development of language technologies for the Uzbek–Chinese pair.

Keywords: parallel corpus, Uzbek language, Chinese language, sentence alignment, word alignment, translation equivalence, agglutinative morphology, isolating language, contrastive linguistics, corpus annotation.

INTRODUCTION

Parallel corpora, that is, collections of source texts aligned with their translations, are among the most valuable resources in modern linguistics. They allow researchers to observe how meanings are expressed in two languages, to extract translation equivalents, to test hypotheses about cross-linguistic differences, and to train and evaluate machine translation systems. For widely studied language pairs, large parallel corpora have been available for many years. For less-resourced pairs, however, the lack of aligned data remains a serious obstacle to both research and practical applications.

The Uzbek–Chinese pair belongs to this second group. Economic, educational and cultural cooperation between Uzbekistan and China has expanded considerably, and the demand for qualified translators and reliable translation tools has grown accordingly. Yet most corpus work on Uzbek has focused on Uzbek–English or on monolingual resources. Recent studies have addressed the stages of building an Uzbek–English parallel corpus and the alignment problem associated with it [1], the creation of a morphologically and syntactically tagged Uzbek corpus [5], and the development of translation resources for Southern Uzbek [4]. A dedicated methodological framework for Uzbek–Chinese corpora is still lacking.

The difficulty of building such a corpus lies not only in collecting texts but in the structural distance between the two languages. Uzbek is an agglutinative Turkic language in which grammatical meanings

are expressed through chains of suffixes attached to the stem, and its basic word order is subject–object–verb. Chinese is an isolating language with little inflection, in which grammatical relations are expressed mainly through word order, function words and particles, and its basic word order is subject–verb–object. These differences directly affect the way sentences and words can be aligned and the way translation equivalence can be defined.

The aim of this article is to formulate the theoretical and methodological foundations for developing Uzbek–Chinese parallel corpora. The objectives are: to define the levels of alignment relevant to this language pair; to clarify the concept of translation equivalence in a corpus setting; to describe the main structural differences that affect alignment; and to propose a construction framework that combines automatic tools with linguistic annotation and human verification.

LITERATURE REVIEW AND METHODOLOGY

Parallel corpus development for Uzbek

Corpus work on Uzbek has developed rapidly in recent years. Amirkulov [1] describes the stages of creating an Uzbek–English parallel corpus and proposes a model algorithm based on the achievements of world corpus studies, while highlighting alignment as one of the central technical problems. This work demonstrates that the construction of a parallel corpus is not a single operation but a sequence of stages, each of which requires both linguistic and technical decisions.

Sharipov and colleagues [5] address another essential component: annotation. They develop a new part-of-speech and syntactic tagset for Uzbek, describe a web-based tagging application, and report the first stage of building a tagged corpus. For a parallel corpus, such annotation is crucial, because the alignment of an agglutinative language with an isolating one becomes much more reliable when morphological boundaries and syntactic functions are explicitly marked.

Mamasaidov and colleagues [4] show what a focused resource effort can achieve for a less-resourced variety. Working on Southern Uzbek, spoken by about five million people in Afghanistan, they release a 997-sentence evaluation set aligned with the FLORES+ benchmark, a collection of 39,994 parallel sentences, and a translation model fine-tuned from NLLB-200. Their work illustrates the practical value of combining existing benchmarks with newly aligned data and the importance of script-specific post-processing.

Word alignment methods

Word alignment, the identification of corresponding words or expressions across a sentence pair, has traditionally relied on statistical models trained on large parallel corpora. Such models perform poorly when parallel data are scarce, which is precisely the situation for Uzbek–Chinese. Recent embedding-based methods offer an alternative. Jalili Sabet and colleagues [3] propose SimAlign, which derives word alignments from multilingual embeddings trained only on monolingual text, without any parallel data or dictionaries. They report that embedding-based alignments are better than traditional statistical aligners for four language pairs and comparable for two, and that for English–German contextualized embeddings achieve an F1 score five percentage points higher than a strong statistical aligner trained on 100,000 parallel sentences.

Dou and Neubig [2] take a complementary approach. Their aligner, known as AWESOME, fine-tunes pre-trained multilingual language models on parallel text in order to improve alignment quality. In

experiments on five language pairs, it consistently outperforms previous state-of-the-art aligners. Together, these two methods suggest a practical strategy for low-resource pairs: begin with alignment from multilingual embeddings when parallel data are limited, and refine the model through fine-tuning as the verified portion of the corpus grows.

Methodology

The study adopts a theoretical and methodological design. First, it synthesizes the existing literature on Uzbek corpus development and on word alignment methods [1–5]. Second, it applies contrastive analysis to the grammatical systems of Uzbek and Chinese in order to identify the structural features that affect alignment. Third, it uses illustrative sentence pairs to show how these features produce one-to-one, one-to-many and many-to-many correspondences. Finally, it integrates these findings into a multi-level framework for corpus construction.

The illustrative examples in this article were constructed for analytical purposes. They are intended to demonstrate typical alignment configurations rather than to report corpus frequencies. Quantitative validation of the framework, including measurements of alignment accuracy on a sample of verified sentence pairs, is identified as a task for subsequent empirical work.

RESULTS AND DISCUSSION

Levels of alignment

The analysis distinguishes four levels of alignment in an Uzbek–Chinese parallel corpus. At the document level, a source text is paired with its translation, and metadata such as genre, translator, direction of translation and publication date are recorded. At the paragraph level, paragraphs are matched, which is usually straightforward in literary and official texts but can be complicated in texts where the translator has merged or divided paragraphs.

At the sentence level, alignment becomes more demanding. Sentence boundaries do not always coincide, because a long Uzbek sentence containing several participial or converbial clauses may be rendered in Chinese as two or more shorter sentences, and vice versa. Length-based alignment methods, which assume that a long sentence corresponds to a long translation, are sensitive to such insertions and divisions. As the work on Uzbek alignment shows [1], this problem must be solved before reliable word-level analysis is possible.

At the word level, the correspondence between units is least direct. One Uzbek word form may correspond to several Chinese words, and one Chinese word may correspond to a stem with several suffixes in Uzbek. For this reason, the framework treats the morpheme rather than the orthographic word as the minimal unit of alignment on the Uzbek side, while Chinese text is first segmented into words, since Chinese writing does not mark word boundaries with spaces.

Structural differences affecting alignment

The first major difference concerns morphology. Consider the Uzbek word form “maktabda”, consisting of the stem “maktab” (school) and the locative suffix “-da”. In Chinese, the same meaning is expressed by a combination of a preposition and a noun, as in “在学校” (zài xuéxiào, “at school”). A single Uzbek word thus corresponds to two Chinese words, and the case suffix aligns with a separate

preposition. If alignment is performed at the level of orthographic words, this correspondence is either lost or recorded imprecisely.

The second difference concerns word order. In the Uzbek sentence “Men maktabda o‘qiyman” (I study at school), and in its Chinese equivalent “我在学校学习” (wǒ zài xuéxiào xuéxí), both languages happen to place the verb after the locative phrase. The orders diverge as soon as a direct object appears: Uzbek “Men kitob o‘qiyman” (I read a book) places the object before the verb, whereas Chinese “我看书” (wǒ kàn shū) places it after the verb. Alignment links therefore frequently cross each other, and alignment tools must be able to handle non-monotonic correspondences.

The third difference concerns grammatical categories that have no direct counterpart. Chinese uses numeral classifiers, as in “三本书” (sān běn shū, “three books”), where the classifier “本” has no equivalent in the Uzbek phrase “uchta kitob”. Conversely, Uzbek expresses possession, person and number through suffixes, as in “kitobim” (my book), which corresponds to the Chinese phrase “我的书” (wǒ de shū) with a separate pronoun and a structural particle. Aspectual meanings are also distributed differently: Chinese particles such as “了” (le) mark completed action, whereas Uzbek expresses tense and aspect through verbal suffixes, as in “o‘qidim” (I read, completed).

These differences produce three typical alignment configurations. In one-to-one alignment, a unit in one language corresponds to exactly one unit in the other, which is common for lexical stems. In one-to-many alignment, a single Uzbek word form corresponds to several Chinese words, or a single Chinese word corresponds to several Uzbek morphemes. In null alignment, a unit has no counterpart, as with Chinese classifiers or structural particles. A corpus design that records all three configurations explicitly provides a much more accurate basis for contrastive analysis than one that assumes word-to-word correspondence.

Translation equivalence in the corpus

Translation equivalence in a parallel corpus cannot be reduced to formal correspondence. The analysis distinguishes three levels. Formal equivalence holds when the structures of the source and target expressions correspond closely, which is rare between Uzbek and Chinese because of their structural distance. Semantic equivalence holds when the propositional content is preserved even though the structure differs, as in the locative example above. Pragmatic equivalence holds when the communicative function is preserved, for example in idiomatic expressions, politeness formulas or culturally specific references.

For corpus construction, this distinction has a practical consequence. The annotation scheme should allow alignment links to be labelled by type, so that researchers can separate literal correspondences from paraphrases and culturally adapted renderings. Such labelling turns the corpus from a simple collection of aligned sentences into a resource for studying translation strategies, and it also allows machine translation developers to filter training data according to the degree of correspondence.

A multi-level framework for corpus construction

On the basis of the analysis, a framework consisting of five stages is proposed. The first stage is text selection and documentation. Texts are collected from genres in which Uzbek–Chinese translations exist, such as official documents, news, literary works and educational materials, and each pair is documented

with metadata on genre, translator and direction of translation. Recording the direction of translation is essential, because translated language differs systematically from original language.

The second stage is preprocessing. Uzbek texts are normalized with respect to script and spelling, since Uzbek is written in both Latin and Cyrillic alphabets, and Chinese texts are segmented into words. Uzbek word forms are then divided into stems and suffixes using a morphological analyzer, and both sides are annotated with part-of-speech tags. The tagset developed for Uzbek [5] provides a basis for this step.

The third stage is sentence alignment. An automatic aligner proposes sentence correspondences, which are then checked against paragraph boundaries and length ratios. Pairs with unusual length ratios or one-to-many correspondences are flagged for manual review, since these are the cases in which length-based methods are most likely to fail [1].

The fourth stage is word and morpheme alignment. In the initial phase, when verified parallel data are limited, embedding-based alignment without parallel training data is applied [3]. As the verified portion of the corpus grows, the aligner is fine-tuned on this portion to improve accuracy [2]. Alignment on the Uzbek side is performed at the morpheme level wherever possible, so that case suffixes, possessive markers and verbal affixes can be linked to the corresponding Chinese function words and particles.

The fifth stage is human verification and quality control. A sample of aligned sentences is checked by bilingual annotators, and the agreement between annotators is measured. Errors detected during verification are used to correct the automatic output and to retrain the alignment model. This cycle of automatic processing and human verification, which has proved effective in related corpus projects, gradually increases both the size and the reliability of the corpus.

Applications and limitations

A corpus built according to this framework can serve several purposes. In contrastive linguistics, it enables systematic study of how Uzbek case and possessive suffixes correspond to Chinese prepositions, particles and word order. In translation studies, it allows researchers to compare strategies used by different translators and in different genres. In translator training, aligned examples provide authentic material for teaching students how to handle structural mismatches. In language technology, verified aligned data can be used to train and evaluate machine translation systems for a pair that currently lacks such resources, following the example of recent work on Southern Uzbek [4].

The framework also has limitations. Its effectiveness depends on the availability of Uzbek morphological tools and Chinese segmentation tools of sufficient quality, and errors at these stages propagate to alignment. The amount of human verification required is considerable, especially in the early phases. Finally, the illustrative analysis presented here must be confirmed by quantitative evaluation on a verified sample, which is the next step of the research.

CONCLUSION

The article has outlined the theoretical and methodological foundations for developing Uzbek–Chinese parallel corpora. The analysis shows that the structural distance between agglutinative Uzbek and isolating Chinese makes word-to-word alignment inadequate and requires alignment at several levels, with the morpheme as the minimal unit on the Uzbek side and segmented words on the Chinese side. Differences in morphology, word order and grammatical categories produce one-to-one, one-to-many and null alignment configurations, all of which must be represented explicitly in the corpus.

Translation equivalence in the corpus should be understood at formal, semantic and pragmatic levels, and alignment links should be labelled accordingly. Recent advances in embedding-based word alignment [2; 3] and in Uzbek corpus development [1; 4; 5] make it possible to build such a corpus even with limited initial resources, provided that automatic processing is combined with linguistic annotation and human verification.

Further research should implement the proposed framework on a pilot corpus, measure alignment accuracy and inter-annotator agreement, and expand the corpus across genres. Such work would provide a foundation for contrastive studies, translator education and language technologies for the Uzbek–Chinese language pair.

REFERENCES

1. Amirkulov B. Uzbeki-English parallel corpus algorithm and alignment problem // Central Asian Journal of Literature, Philosophy and Culture. – 2023. – URL: https://bibsonomy.org/bibtex/2ddd0c49052c9654c8c51bc33865a91a4/centralasian_20 (accessed: 01.11.2026).
2. Dou Z.-Y., Neubig G. Word alignment by fine-tuning embeddings on parallel corpora // Proceedings of EACL 2021. – 2021. – URL: https://arxiv.org/abs/2101.08231 (accessed: 01.11.2026).
3. Jalili Sabet M., Dufter P., Yvon F., Schütze H. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings // Findings of EMNLP 2020. – 2020. – URL: https://arxiv.org/abs/2004.08728 (accessed: 01.11.2026).
4. Mamasaidov M., Aral A., Shopulatov A., Inomjonov M. Filling the gap for Uzbek: Creating translation resources for Southern Uzbek [Preprint]. – 2025. – URL: https://arxiv.org/abs/2508.14586 (accessed: 01.11.2026).
5. Sharipov M., Mattiev J., Sobirov J., Baltayev R. Creating a morphological and syntactic tagged corpus for the Uzbek language [Preprint]. – 2022. – URL: https://arxiv.org/abs/2210.15234 (accessed: 01.11.2026).