Western European Journal of Linguistics and Education
Volume 3, Issue 8, August 2025
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed

KARAKALPAK AND UZBEK LANGUAGES PARALLEL CORPUS AND ITS SYNTACTIC ADAPTATION

Khudoyberganova Munisa Shonazar qizi
Master's Student, National University of Uzbekistan
munisaxudayberganova13@gmail.com

Annotation: This article examines the creation and syntactic adaptation of a parallel corpus for Karakalpak and Uzbek languages. It describes the role of parallel corpora in Natural Language Processing (NLP), machine translation, and linguistic research. The study focuses on the PARATRANSLATOR platform, which integrates multilingual corpora and enables contextual translations. It analyzes morphological and syntactic tagging, alignment techniques, and structural correspondences between the two languages. Similarities and differences in grammar, phonology, and lexicon between Karakalpak and Uzbek are highlighted. The research emphasizes the importance of such resources for preserving low-resource languages and improving AI-based translation technologies.

Keywords: Parallel corpus, NLP, Machine translation, Syntactic adaptation, Language resources.

A parallel corpus is a collection of texts written in two or more languages that correspond in meaning. In such a collection, the texts for each language are arranged in translation pairs, meaning that each sentence or phrase in one language corresponds to its equivalent in another language. This corpus is an important tool in the field of Natural Language Processing (NLP), particularly in machine translation, terminological research, and multilingual linguistic analysis.

The most important aspect of a parallel corpus is linking equivalent sentences in two languages, ensuring its accuracy and reliability [N. Abduraxmonova, G. Shamsiyeva. 2025]. Therefore, the collected data must be verified by linguists and translators. The following evaluation methods are used for this purpose [Brown, P.F., Cocke, J., Della Pietra, S.A., Mercer, R.L, 1993]:

- ✓ BLEU(Bilingual Evaluation Understudy);
- ✓ METOR;
- ✓ TER(Translation Edit Rate);
- ✓ N-gramstatisticalmodels.

In NLP, machine translation, linguistic studies, and many other research areas, parallel corpora are of great significance and are widely used in:

- Machine translation;
- Linguistic research;
- Artificial intelligence and NLP systems;
- Substitution and adaptation techniques.

In machine translation, the main lexicographic resources are dictionaries and parallel corpora (bitexts). Parallel corpora serve as a primary source for corpus-based machine translation, containing original texts and their translations. In computational linguistics,

Western European Journal of Linguistics and Education
Volume 3, Issue 8, August 2025
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed

translation corpora have been used since the 1980s [N. Abduraxmonova, G. Shamsiyeva. 2025]. One of the first electronic sources, the Canadian Hansard, was initially used for sentence alignment, which has now become a standard feature in applications such as translation memories. Within the framework of the practical project "PARATRANSLATOR – Creation of an Electronic Contextual Translation Dictionary Platform Based on a Parallel Corpus," a multilingual parallel corpus was created for Uzbek, English, Russian, Turkish, Japanese, French, Spanish, and Karakalpak languages [https://paratranslator.uz/en/]. PARATRANSLATOR is an online program that integrates large multilingual corpora, allowing users to search for translations in context. It not only translates words but also provides paired examples of their usage in context across two languages in various styles. This project enables users to more accurately determine the meaning of various expressions and phraseological units in nine languages (Uzbek, Turkish, Karakalpak, English, French, Russian, Spanish, Japanese, and Korean) during translation. It also demonstrates natural language features through examples from the Uzbek electronic corpus, covering lexical (homonymy, synonymy, antonymy) and grammatical characteristics, and helps in understanding texts [https://paratranslator.uz/en/]. For the first time, translation from Karakalpak to Uzbek is implemented via a parallel corpus in this platform.

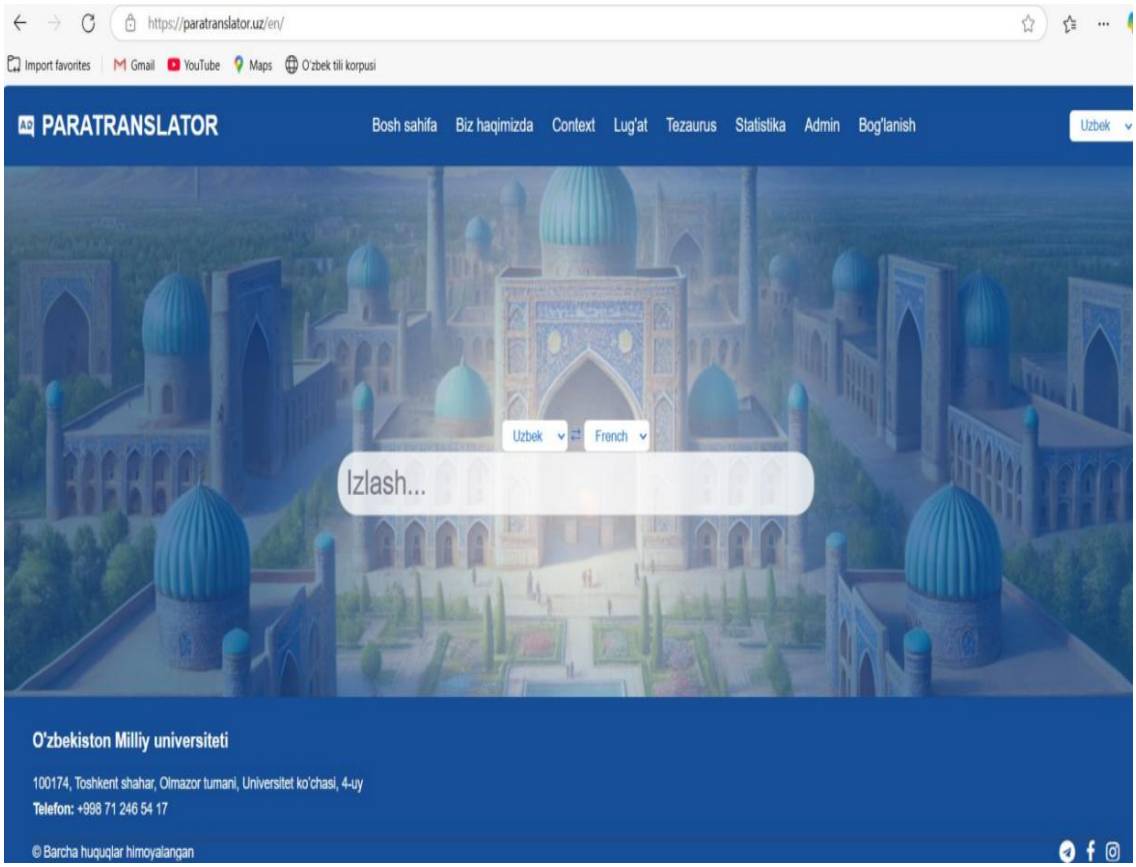

Figure 1. "PARATRANSLATOR" Platform.

The creation of the PARATRANSLATOR platform is of great scientific and practical significance for the digitalization of both Uzbek and Karakalpak languages. While there are hundreds of electronic dictionaries based on English worldwide, including contextual parallel

Western European Journal of Linguistics and Education
Volume 3, Issue 8, August 2025
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed

corpus-based platforms for translation from English to Turkic languages, creating digital resources for Uzbek and Karakalpak is especially important for AI-powered applications. This project serves practical purposes such as improving translation processes, developing terminological resources for Uzbek, and creating tools for language teaching for a wide audience.

Karakalpak is considered a low-resource language. Building a parallel corpus for Karakalpak and Uzbek helps preserve the modern state of the Karakalpak language, its lexical and syntactic features, and provides a valuable source of information for learners, researchers, and translators. Parallel corpora are essential resources for automatic language processing and translation systems.

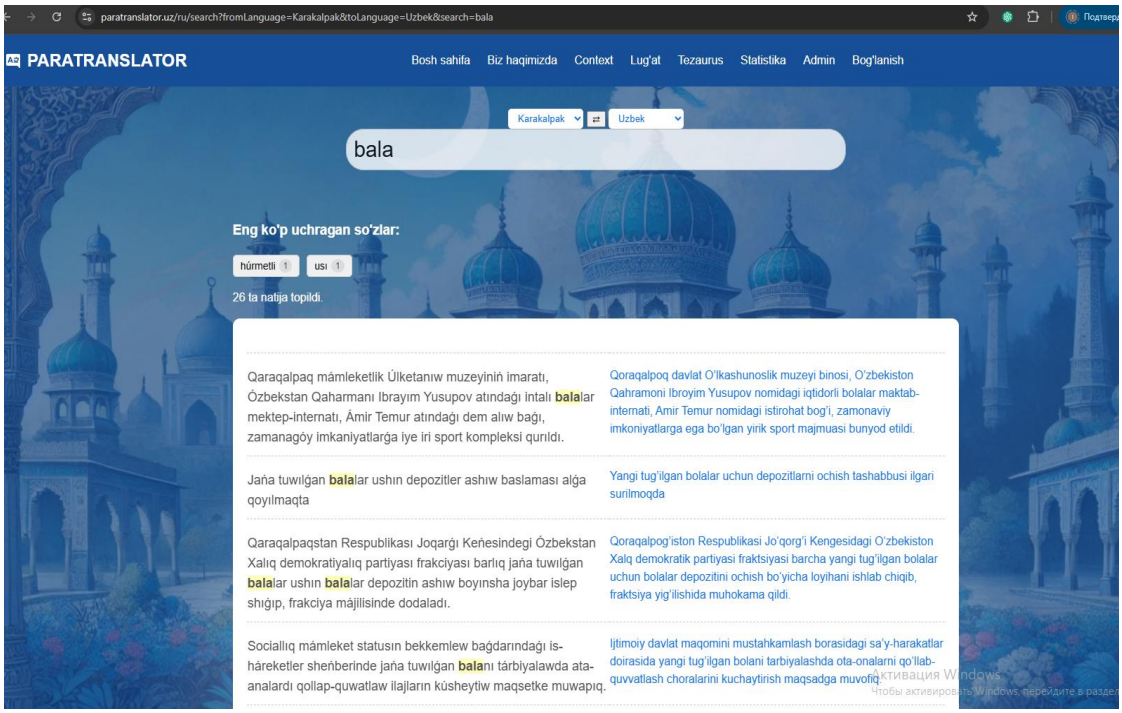

Figure 2. Paratranslator Platform Search Interface

Through Karakalpak–Uzbek translation pairs, translation systems can study and optimize cross-linguistic semantic, morphological, and syntactic relations. In constructing NLP models, the parallel corpus allows for the analysis of grammatical structure, word choice, and meanings in both languages, resulting in accurate and reliable language models.

Karakalpak and Uzbek belong to the Turkic language family and share similarities and differences in their syntactic structures and morphological features. Syntactic adaptation is the process of ensuring the structural correspondence of sentence components (subject, predicate, object, embedded clauses, etc.) in the parallel corpus. This process involves the following main tasks:

- *Morphological and syntactic tagging:* For each language, elements of the text (words and sentence parts) are identified using tags. Morphological tagging specifies parts of speech, affixes, and grammatical features, while syntactic tagging identifies the sentence's syntactic

Western European Journal of Linguistics and Education
Volume 3, Issue 8, August 2025
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed

structure, i.e., subject, predicate, object, and additional elements.
- *Building syntactic tree structures:* Each sentence is represented as a tree structure, showing the relationships between its components.

- *Alignment:* Corresponding syntactic elements are identified between sentences in the two languages. The main elements — subject, predicate, and object — are compared to determine structural correspondence.

In implementing machine translation between Karakalpak and Uzbek, attention should be given to the following similarities and differences:

Similarities:
- Both languages, although belonging to the Kipchak (Karakalpak) and Karluk (Uzbek) subgroups, share a common Turkic grammatical base [Qudaybergenov M., Qazaqbaeva A., Kurtchaev A. 2016].

- Word formation is carried out through suffixation (agglutinative languages).

- The typical word order is S+O+V (Subject + Object + Verb):

Men kitob o'qiyapman / Men kitap o'qıp atirman.

- Many Turkic root words exist in both languages: *ko'z, quloq, suv, qo'l, qish, yoz, yurak, til, ona, ota, etc.*

- Many words in both languages differ only by phonetic changes: *yurak ↔ jürek bola ↔ bala, ko'p ↔ köp, yangi ↔ jańa*

- Both languages have hard and soft consonants: *k/g, q/g', t/d.*

- Karakalpak has the sound *"ń,"* which in some cases corresponds to Uzbek *"ng": jańa ↔ yangi*

- Both languages use Cyrillic and Latin scripts, with an ongoing transition to Latin.

Differences are shown in the following table:

Field	Karakalpak	Uzbek
Language group	Kipchak	Karluk
Phonetics	ń, ı, w, q sounds active	ng, i, v, g' sounds active
Vocabulary	Dominantly Turkic words	More Persian and Arabic borrowings
Pronunciation	Stronger, typical of Kipchak dialects	Softer, influenced by Karluk and Persian
Verb forms	Barıw, kelıw, otırıw	Bormoq, kelmoq, o'tirmoq

While both belong to the Turkic family and generally share grammatical rules and sentence structures (SOV), some morphological and syntactic differences exist:
- *Morphological differences:* In Karakalpak, suffixes play a key role in grammatical adaptation. Uzbek also uses suffixes, but their forms and functions differ. For example, the Karakalpak

Western European Journal of Linguistics and Education
Volume 3, Issue 8, August 2025
https://westerneuropeanstudies.com/index.php/2
ISSN (E): 2942-190X Open Access| Peer Reviewed

plural suffix -lar/-ler corresponds to Uzbek -lar, but in Karakalpak, -ler conforms to vowel harmony: mektepler, balalar (Uzbek: maktablar, bolalar) [Qudaybergenov M., Qazaqbaeva A., Kurtchaev A. 2016].

- *Syntactic precision:* Although the basic element order is similar, the position of additional sentence elements, case markers, and other syntactic units may differ.

In conclusion, the syntactic adaptation of the Karakalpak–Uzbek parallel corpus, along with morphological and syntactic tagging, serves as a foundation for research in linguistics, machine translation, and NLP. This corpus helps improve translation systems and language models, while enabling a deeper understanding of the languages' development and unique features. Future studies can expand the use of this parallel corpus to further develop automatic translation and NLP technologies.

The Karakalpak–Uzbek parallel corpus serves to:

- ❖ enrich, preserve, and develop language resources;
- ❖ improve machine translation and NLP technologies;
- ❖ provide deep analysis of linguistics and language evolution;
- ❖ advance cultural and educational initiatives.

References:

1. Brown, P.F., Cocke, J., Della Pietra, S.A., Mercer, R.L, 1993 The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 1993(2), pp. 263–311.
2. N. Abduraxmonova, G. Shamsiyeva. Machine Translation Based on Parallel Corpus. Globeedit. 2025. – p. 57.
3. Qudaybergenov M., Qazaqbaeva A., Kurtchaev A. (2016). Theoretical Grammar of the Karakalpak Language: A Textbook. Nukus, 2016, 150 pages.
4. https://paratranslator.uz/en/