ORIGINAL ARTICLE

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE

ISSN: 2692-5206, Impact Factor: 12,23
American Academic publishers, volume 05, issue 04, 2025

Journal: https://www.academicpublishers.org/journals/index.php/ijai

CREATING A TURKIC LANGUAGES PARALLEL CORPUS FOR THE UZBEK LANGUAGE CORPUS

Bekchanova Zebo Baxramovna
1st-year Master's student in Computational Linguistics, Urgench State University.
Ko‘palova Shahrizoda Otabek qizi

1st-year Master's student in Computational Linguistics, Urgench State University.

Annotation: This article highlights the main stages and challenges of creating a parallel corpus. It also provides an in-depth analysis of the lexical, syntactic, semantic, and pragmatic features of the language using corpora. Important issues such as taking into account the lexical differences between Uzbek and other Turkic languages, and the creation of a parallel corpus, are addressed.

Keywords: language, language family, Turkic languages, lexical difference, parallel corpus, lexical, syntactic, semantic, and pragmatic features of the language.

Introduction.

Corpus linguistics is one of the most advanced fields of modern linguistics, playing an essential role in the scientific study and analysis of language. In recent years, corpus linguistics has been rapidly developing. The use of parallel corpora not only broadens the scope of applied linguistic research but also fosters new philosophical perspectives for language teaching studies. Through corpora, the lexical, syntactic, semantic, and pragmatic features of a language can be deeply analyzed. Such research is often carried out using parallel corpora. Parallel corpora are collections of texts with the same content in multiple languages, and their analysis helps identify similarities and differences between languages. Especially, creating a parallel corpus between Turkic languages contributes significantly not only to improving language teaching and translation processes but also to the development of machine translation systems. The scientific, practical, and technological importance of creating a parallel corpus between Uzbek and other Turkic languages is substantial.

Main Body.

Creating parallel corpora offers linguists and translators opportunities not only to improve translation quality but also to better understand linguistic relationships and differences. According to R. Karimov, parallel corpora are a key tool for studying syntactic, semantic, and pragmatic similarities between languages, and serve as a foundation for tasks such as linguistic and extralinguistic tagging, developing parallel corpus algorithms, identifying formal and informal registers of usage in morphology, syntax, and translation studies. They also enhance the reliability and objectivity of research that would otherwise rely heavily on linguistic intuition, enabling the creation of new-generation corpus-based dictionaries and grammars. This highlights the main purpose of creating parallel corpora. Turkic languages are closely related, making the creation of a parallel corpus between them highly beneficial. The first step in creating a parallel corpus is to study the grammatical systems of the languages involved. M. Imrad notes: grammatical analysis is a crucial stage in the creation of a parallel corpus, enabling the comparison of syntactic structures across languages. Syntactic and morphological similarities among Turkic languages influence the methodologies used in translation between these languages. The most reliable data for parallel corpora are human-translated texts. Texts translated by humans increase the accuracy of the corpus and serve as a high-quality database for machine translation, offering users higher quality translations compared to existing machine translation tools. Moreover, parallel corpora are important tools for identifying lexical relationships between languages. Researchers note that with parallel corpora, it is possible to identify lexical and semantic similarities, helping linguists improve translation systems. These corpora assist in analyzing the semantic layers of language and ensuring accurate, appropriate translations. Considering the lexical differences between Uzbek and other Turkic languages is crucial in creating a parallel corpus. Another theoretical and practical aspect of creating parallel corpora is their contribution to a deeper understanding of linguistic elements in the language learning and translation processes. These corpora also make it possible to analyze pragmatic differences between languages, as each language has unique contextual features. Bilingual and multilingual corpora have immense value across various fields. Their importance in comparative analysis is supported by the views of scholars such as Aijmer and Altenberg. One important aspect of creating a parallel corpus is maintaining text quality. Accurate and precise translations ensure the effectiveness of the corpus. It is essential that semantic and syntactic features of the language are correctly represented in translations. Such corpora are not only useful for scientific analysis but also for practical applications like machine translation systems.

Corpus Creation Process.

The following main stages can be identified in creating a parallel corpus between Turkic languages: in the first stage, the grammatical systems and lexical bases of the languages are analyzed; in the second stage, translated texts are collected and analyzed; in the third stage, texts are aligned and refined with the help of specialists to ensure accurate translation. According to researchers, this process allows for the study of different layers of language and improves translation quality. When collecting texts, literary and scientific materials are especially preferred, as they reflect accurate and clear forms of language. Studying the differences between literary and scientific texts enables comprehensive linguistic analysis when creating a parallel corpus.

Conclusion.

Creating a parallel corpus between Turkic languages is of great significance in linguistics, translation, and computational linguistics. Through these corpora, it is possible to deeply analyze similarities and differences between languages. Additionally, parallel corpora simplify the language learning process and enhance the effectiveness of translation systems. Today, corpora have become essential tools that save time and effort. Corpus-based language education, dependency-based parsing, FST technology in morphological analysis, national corpus development methodologies, corpus-based morphological and semantic analyzers, and neural technologies based on parallel corpora for machine translation, as well as the development of an educational corpus for the Uzbek language, are all being actively researched. Parallel corpora are useful not only in linguistics but also in translation studies, bilingual lexicography, and any field where language comparison is necessary.

References:

1. Abdurakhmonova N, Tuliyev U. Morphological analysis by finite state transducer for Uzbek-English machine translation. Foreign Philology: Language. Literature, Education. 2018(3):68.
2. Abdurakhmonova N, Urdishev K. Corpus-based teaching Uzbek as a foreign language. Journal of Foreign Language Teaching and Applied Linguistics (JFLTAL). 2019;6(1-2019):131-137.
3. Abduraxmonova, N. Z. "Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) dissertation abstract." (2018).

4. Abdurakhmonova N. The bases of automatic morphological analysis for machine translation. Izvestiya Kyrgyzskogo gosudarstvennogo tekhnicheskogo universiteta. 2016;2(38):12-17.

5. John Hutchins. Machine translation and human translation: in competition or in complementation. International Journal of Translation, 13(1-2):5–20, 2001.

6. Johansson, S. Seeing through Multilingual Corpora. Amsterdam: Benjamins. 2007.

7. Imrad, M. Parallel Korpuslar Yaratish: Nazariy va Amaliy Asoslar. Tashkent: Ma'naviyat nashriyoti. 2014.

8. Karimov Rustam. O‘zbek-ingliz parallel korpusini tuzishning lingvistik va dasturiy masalalari. Dissertation. Bukhara – 2022.

9. Q.F. Wen, L.F. Wang, M.C. Liang. Spoken and Written English Corpus of Chinese Learners. Beijing: Foreign Language Teaching and Research Press, 2005. (In Chinese)