JOURNAL OF MULTIDISCIPLINARY SCIENCES AND INNOVATIONS
VOLUME 05, ISSUE 01
MOPNTHLY JOURNALS
ISSN NUMBER: 2751-4390
IMPACT FACTOR: 9,08

THE THEORETICAL FOUNDATIONS FOR CREATING A TURKIC LANGUAGES PARALLEL CORPUS FOR THE UZBEK LANGUAGE CORPUS

Bekchanova Zebo Baxramovna
2nd-year Master's student of Computational Linguistics,
Urgench State University;

Abstract

The article discusses the main stages and challenges in the process of creating a parallel corpus. It provides an in-depth analysis of the lexical, syntactic, semantic, and pragmatic features of language using corpora. Important issues such as accounting for lexical differences between the Uzbek language and other Turkic languages and the creation of a parallel corpus are addressed.

Key words

language, language family, Turkic languages, lexical differences, parallel corpus, lexical, syntactic, semantic, and pragmatic features of language.

Introduction

One of the global challenges of the 21st century is the preservation of the national identity of natural languages. The creation and development of electronic corpora of world languages, along with systematic research in NLP and language technologies, has become an urgent task[1]. Scientific and practical research conducted in foreign corpus linguistics has proven that corpora are not only essential for specialists working with language, but also play a significant role in national development.

Corpus linguistics is one of the most rapidly developing fields of modern linguistics and plays an important role in the scientific study and analysis of language. In recent years, corpus linguistics has been developing intensively. The application of parallel corpora not only expands the scope of applied linguistic research but also forms new philosophical approaches to foreign language teaching[2]. Using corpora, it is possible to deeply study the lexical, syntactic, semantic, and pragmatic features of language. Such studies are often conducted using parallel corpora.

In Uzbekistan, one of the most important tasks facing applied linguistics today has been elevated to the level of state policy—namely, the creation of a national corpus of the Uzbek language. In particular, to enhance the prestige and status of the Uzbek language in society and internationally, it is necessary to create an electronic national corpus that integrates all scientific, theoretical, and practical information related to the Uzbek language. Other priorities include promoting the Uzbek language on the global Internet information network, ensuring its proper representation, developing Uzbek-language software applications, implementing computer programs for teaching Uzbek on a large scale, and creating software for editing Uzbek-language texts[3]. Considerable research efforts are currently underway in this area.

---

[1] Abduraxmonova N. O`zbek tili electron korpusining kompyuter modellari (monografiya) /Toshkent: Muharrir, 2021, 202 b

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

[3] O'zbekiston Respublikasi Prezidentining 2019-yil 21-oktabrdagi "O'zbek tilining davlat tili sifatidagi nufuzi va mavqeyini tubdan oshirish chora tadbirlari to'g'risida"gi PF-5850-son

In the fields of Russian and English corpus linguistics, numerous scholars such as V. Zakharov, A. Sedov, A. Baranov, R. Potapova, V. Rykov, U. Francis, N. Leontyeva, V. Martin, S. Kübler, A. Laurence, E. Atwell, S. Hunston, L. Boizou, McEnery, J. Graffmiller, J. Grieve, N. Groom, S. Hansson, M. Mahlberg, P. Milin, A. Murakami, R. Paice, A. Schembri, P. Thompson, B. Winter, and G. Lynch have conducted significant research[4]. In Turkology, corpus linguistics research has been carried out by scholars such as Aksan, Deniz, Zeyrek, Kemal Oflazer, Umut Özge (Turkish); Yusup Aibaidulla, Kim-Teng Lua (Uyghur); I. A. Buskunbaeva, Z. Sirazitdinov (Bashkir); Sheymovich (Khakas); J. Suleymanov, A. Gatiatullin, O. Nevzorova, R. Gilmullin, B. Hakimov (Tatar); L. Kubedinova (Crimean Tatar); and Salchak (Tuvan).

Among Uzbek scholars, B. Mengliyev, Sh. Shahobiddinova, Z. Xolmanova, S. Karimov, N. Abdurakhmonova, L. Raupova, Sh. Hamroyeva, M. Abjalova, G. Toirova, G. Ikromova, J. Jumabayeva, G. Ergasheva, and A. Eshmo‘minov have contributed notable research. The concept of the Uzbek National Corpus is currently being developed by a research team led by B. Mengliyev.

Main Part

The parallel corpus, as an autonomous component of an electronic corpus, is significant due to its ability to accumulate a vast amount of valuable information. Specially formatted multilingual corpora designed for side-by-side comparison in machine translation are known as aligned parallel corpora. One of the earliest examples of parallel texts dates back to 196 BC—the Rosetta Stone, discovered in 1799 near the city of Rosetta in the Nile Delta, containing inscriptions in two languages (Greek and Egyptian). Information on the structure, composition, and capabilities of parallel corpora can be found in the works of D. O. Dobrovolsky, Yu. Tao, V. Zakharov, A. A. Kokoreva, and E. P. Sosnina.

A parallel corpus, consisting of original texts and their translations, can be used in various ways for translation studies, machine translation, linguistics, computational linguistics, and human translators. In computational linguistics, translation corpora began to be used in the early 1980s for machine translation, as well as for term extraction and word sense disambiguation. Early parallel corpora included avalanche reports collected in German, French, and Italian in Switzerland, and weather reports published in English and French by Canadian media in the late 1980s and early 1990s.

One of the first electronic resources was the Canadian Hansard corpus, initially used for sentence alignment (Gale & Church, 1991), a feature that has since become standard in applications such as translation memories. Parallel corpora are also used as databases for multilingual grammar induction, automatic lexicography, cross-lingual information retrieval, and other language processing tasks. The ultimate goal of most projects developed during this period was to create machine translation systems.

The creation of parallel corpora enables linguists and translators not only to improve translation quality but also to gain a clearer understanding of interlingual relationships and differences. According to R. Karimov, parallel corpora are an important tool for studying syntactic, semantic, and pragmatic similarities between languages. Priority tasks such as linguistic and extralinguistic tagging and developing algorithms for parallel corpus construction enhance objectivity in linguistic research, support the identification of formal and informal registers, and provide opportunities for creating next-generation corpus-based dictionaries and grammars[5]. This perspective highlights the core purpose of creating parallel corpora. Given the

[4] Abdurakhmonova N. Kompyuter lingvistikasi (darslik) / Globe edit publishing, 2020, 395 b.

[5] Karimov Rustam. O‘zbek-ingliz parallel korpusini tuzishning lingvistik va dasturiy masalalari. Dissertasiya. Buxoro – 2022

close relationship among Turkic languages, creating a parallel corpus for these languages can be particularly beneficial.

The first step in creating a parallel corpus is the study of grammatical systems across languages. Syntactic and morphological similarities among Turkic languages influence the methodologies applied in translation. The most reliable sources for parallel corpora are human-translated texts. Human translations increase corpus accuracy, serve as high-quality datasets for machine translation, and offer users translations superior to those produced by existing automatic translation tools.

Parallel corpora also play a key role in identifying lexical relationships between languages. Researchers emphasize that such corpora help identify lexical and semantic similarities, supporting the development of translation systems. They facilitate semantic analysis and ensure accurate and contextually appropriate translations. Accounting for lexical differences between Uzbek and other Turkic languages is essential in creating a parallel corpus. Another theoretical and practical aspect of parallel corpora is their role in deepening linguistic understanding during language learning and translation processes. These corpora also allow for the analysis of pragmatic differences arising from language-specific contextual features.

The importance of bilingual and multilingual corpora in various domains is undeniable. The significance of parallel corpora in comparative analysis is well established[6]. Comparative analysis involves the systematic comparison and description of similarities and differences between two or more languages. Aijmer and Altenberg have explained the advantages of parallel corpora in comparative linguistic analysis. One of the key aspects of creating a parallel corpus is maintaining text quality. Accurate and faithful translations ensure the effectiveness of parallel corpora. Semantic and syntactic aspects of language must be properly reflected in translations. Such corpora can be applied not only in academic research but also in practice, such as in automatic translation systems.

The Process of Corpus Creation

In creating a parallel corpus among Turkic languages, the following main stages can be identified:

1. Analysis of grammatical systems and lexical bases across languages;
2. Collection and analysis of translated texts;
3. Alignment of texts and expert validation to ensure translation accuracy.

According to researchers, this process enables the study of different layers of language and improves translation quality.

During text collection, literary and scientific texts are especially preferred, as they reflect accurate and standardized language usage. Studying differences between literary and scientific texts provides broader analytical opportunities in parallel corpus creation.

Conclusion

The creation of parallel corpora among Turkic languages holds great significance for linguistics, translation studies, and computational linguistics. Such corpora enable deep analysis of similarities and differences between languages, simplify language learning processes, and improve the efficiency of translation systems. Today, corpora have become tools that save time and effort. Corpus-based language education, dependency-based parsing, FST technology in morphological analysis, author corpora, software and linguistic frameworks for national corpus creation, morphological and semantic analyzers, and the development of neural machine translation technologies based on parallel corpora are actively being researched. Parallel corpora

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

are valuable not only in linguistics but also in translation studies, bilingual lexicography, and fields requiring language comparison.

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):131-7.
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) il dis. aftoref." (2018). Abdurakhmonova N. The bases of automatic morphological analysis for machine translation. Izvestiya Kyrgyzskogo gosudarstvennogo tekhnicheskogo universiteta. 2016;2 (38):12-7.
4. John Hutchins. Machine translation and human translation: in competition or in complementation. International Journal of Translation, 13(1-2):5–20, 2001.

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

6. Imrad, M. *Parallel Korpuslar Yaratish: Nazariy va Amaliy Asoslar*. Tashkent: Ma'naviyat nashriyoti. 2014.
7. Karimov Rustam. O‘zbek-ingliz parallel korpusini tuzishning lingvistik va dasturiy masalalari. Dissertasiya. Buxoro – 2022.

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