

TURKOLOGIYADA PARALLEL KORPUSLARNING EVOLYUTSIYASI

Aybibi ISKANDAROVA

Associate Professor, PhD in Philology, Department of Computational Linguistics and Applied Linguistics, National University of Uzbekistan

E-mail: aybibiskandarova1962@gmail.com,

Hakimova Nargiza Feruz qizi

Second-year student, Tashkent State Transport University

hakimovanargiza1234@gmail.com

Annotatsiya: Ushbu maqolada turkologiyada ilk korpuslarning yaratilishi, ularning shakllanish davri haqida ma'lumot beriladi. Parallel korpuslar doirasida filologik tadqiqotlarni olib borishdagi imkoniyatlari haqida fikr bildiriladi. Shuningdek, parallel korpus doirasidagi yutuq va muammolar haqida fikr bildiriladi.

Kalit soʻzlar: korpus lingvistikasi, tarjima birliklari bazasi, matn segmentatsiyasi, qidiruv tizimi, konkordans, parallel korpuslarning ahamiyati.

ЭВОЛЮЦИЯ ПАРАЛЛЕЛЬНЫХ КОРПУСОВ В ТЮРКОЛОГИИ

Аннотация: В данной статье представлена информация о создании первых корпусов в тюркологии и периоде их формирования. Рассматриваются возможности проведения филологических исследований в рамках параллельных корпусов. Также освещаются достижения и проблемы, связанные с параллельными корпусами.

Ключевые слова: корпусная лингвистика, база переводческих единиц, сегментация текста, поисковая система, конкорданс, значение параллельных корпусов.

THE EVOLUTION OF PARALLEL CORPORA IN TURKOLOGY

Abstract: *This article provides information on the creation of the first corpora in Turkology and the period of their formation. It also discusses the possibilities of conducting philological research within parallel corpora. In addition, it considers the achievements and challenges in the field of parallel corpora.*

Keywords: *corpus linguistics, translation unit database, text segmentation, search system, concordance, significance of parallel corpora.*

Introduction: In the late 1990s and the early 2000s, the theoretical foundations of computer linguistics in Turkology began to take shape. The idea of creating corpora of Turkic languages emerged during this period. The initial efforts to develop corpora for Uzbek, Kazakh, Kyrgyz, Tatar, Azerbaijani, and Turkish were experimental in nature. During this time, specialists carried out manual alignment of texts for the corpora. In this regard, European and Russian language corpora served, to some extent, as models in the practice of corpus development.

We classify the stages of the development of corpus linguistics in Turkology as follows:

Traditional stage	Late 20th century, 1990s
	During this period, corpora of Turkic languages had not yet been created. However, extensive research was conducted in the field of comparative linguistics based on translated texts and bilingual dictionaries.

Digitization stage	<p>1990–2005</p> <p>During this period, under the influence of Russian and European corpus linguistics practices, ideas for creating electronic corpora for Turkic languages developed. Work was carried out on collecting and digitizing texts of various styles, as well as on forming small-scale research corpora. At this stage, parallel text alignment was not yet fully developed.</p>
Modern parallel corpora stage	<p>2005–2015</p> <p>During this period, efforts to develop corpora within Turkic languages intensified. The design of the Turkish National Corpus was developed, and texts such as newspapers, literary works, and internet materials were collected for the corpus linguistic base. For the Uzbek language corpus, electronic libraries and text collections were formed, and research corpora were created. Languages such as Turkish, Uzbek, Kazakh, and Azerbaijani were included in the international OPUS parallel corpus. Technologies such as morphological analyzers, POS tagging, lemmatization, and corpus management systems were further improved.</p>
Digital and artificial intelligence stage	<p>From 2015 to the present</p> <p>At present, Turkic parallel corpora are developing in close integration with artificial intelligence and natural language processing (NLP) technologies. They are actively used in machine translation systems. Bilingual parallel corpora, such as Uzbek–English and Tatar–Russian, have been created, and their use has increased significantly. Neural machine translation systems have emerged. Parallel corpora serve as a resource for machine translation and for training artificial intelligence models.</p>

Literature review and methodology: In her study “Computer Models of the Uzbek Language Electronic Corpus,” Professor N. Abdurakhmonova notes a number of works carried out within the Turkic languages corpus domain. In particular, scholars from the Institute of Applied Semiotics of Tatarstan developed a morphologically, syntactically, and semantically annotated tagging system for a linguistically annotated corpus of Turkic languages. Researchers from Istanbul Technical University investigated issues such as adapting a two-stage morphological analyzer to corpus analysis, sentence segmentation, tokenizer development, the universal dependency theory of syntactically annotated corpora, and the development of standards for multi-component words. Scholars from the Akhmet Baitursynov Institute of Linguistics and the Institute of Artificial Intelligence of Kazakhstan developed a Kazakh national corpus manager based on speech styles. Researchers from the National University of Uzbekistan, Tashkent State University of Uzbek Language and Literature, Tashkent University of Information Technologies, and the Tashkent branch of Samarkand State University carried out a number of theoretical and practical studies on corpus-based language teaching, dependency theory-based parsing, FST technology in corpus morphological analysis, author corpora, software and linguistic support for national corpus development, morphological and semantic analyzers of corpora, the development of neural machine translation technologies based on parallel corpora, and the formation of educational and web corpora of the Uzbek language [1.7].

In Turkology, the works of scholars such as Aksan, Deniz Zeyrek, Kemal Oflazer, and Unut Ozge Bular on the Turkish language corpus; Yusup Aibaidulla and Kim-Teng Lua on the Uyghur language; L.A. Buskunbaeva and Z. Sirazitdinov on the Bashkir language; Sheymovich on the Khakas language; J. Suleymanov, A. Gatiatullin, O. Nevzorova, R. Gilmullin, and B. Hakimov

on the Tatar language; L. Kubedinova on the Crimean Tatar language; and A. Salchak on the Tuvan language are noteworthy [1.8].

In Turkology, studies aimed at developing machine translation and parallel corpora for Turkish, Kazakh, Kyrgyz, and Tatar languages have been conducted by R. Khusainov [3], R. Gatiatullin [4], and R. Yeshpanov [5]. Their works are focused on developing effective methodologies for low-resource languages within the framework of parallel corpora. R. Yeshpanov and co-authors have conducted research aimed at creating a large-scale, open, and structured parallel corpus for the Kazakh language. M. Jumashev and co-authors substantiate the importance of collections of parallel texts in the development of machine translation for the Kyrgyz language. The authors of these studies scientifically prove that, rather than creating large-scale corpora, it is more appropriate to first develop small but high-quality, reliable, and cleaned parallel datasets, and that small yet high-quality parallel corpora are of strategic importance in developing machine translation for low-resource languages (low-resource). Studies conducted by A. Valeev and co-authors substantiate the effectiveness of transfer and multilingual approaches for Russian–Tatar neural machine translation.

In Uzbek linguistics, scholars such as S. Muhamedova, N. Abdurakhmanova, Sh. Xamroyeva, A. Eshmuminov, G. Toirova, A. Rakhmonova, M. Xolova, R. Karimov, and N. G‘ulomova have conducted research in corpus linguistics. A number of textbooks and manuals related to the field by scholars such as A. Rahimov, N. Abdurakhmanova, Z. Xolmanova, B. Mengliyev, and Sh. Xamroyeva have also been published [6.8].

A parallel corpus serves as a linguistic base for philological research. It is a segmented electronic analogue of texts used for observing interlingual equivalence boundaries, comparative analysis of source and translated texts, understanding translation strategies, and studying the representation of linguocultural units.

Sh. Xamroyeva emphasizes that a parallel corpus serves to observe the structural organization of the target language of translation and the contextual potential of idioms and words, that is, it is significant in analyzing the structures of two languages, searching for equivalents of original texts in translation studies, automatic translation, linguodidactics, and lexicography [7.81].

Creating a corpus based on parallel texts is not an easy task. First, corresponding text fragments in two languages are paired. Various methods have been developed for automatic alignment of texts, and grammatical constructions in texts of the two languages—words, phrases, and sentences—are aligned. Manually edited bitext variants are included in the parallel corpus. Lexical units for translation memory are formed.

The novel “Anna Karenina” by L. Tolstoy in its original and translated versions into the target language was included in the parallel corpus as follows:

«Ну что за охота спать!» —сказал Степан Аркадьич, после выпитых за ужином нескольких стаканов вина пришедший в свое самое милое и поэтическое настроение.	Stepan Arkadiyich ovqat mahalida bir necha stakan vino ichganidan keyin odatdagi dilbar, shoirona kayfiyatiga kelib: – Ey, uyquga balo bormi! – dedi.
---	--

Вронский ценил это, сделавшееся единственной целью ее жизни, желание не только нравиться, но служить ему, но вместе с тем и тяготился теми любовными сетями, которыми она старалась опутать его.	Anna hayotning birdan-bir maqsadi bo‘lib qolgan bu narsani, ya’ni faqat Vronskiyga yoqishgina emas, unga xizmat ham qilish tilagini Vronskiy qadrlardi, ammo shu bilan uni o‘rab olish harakatida yurgan Annaning sevgi to‘rlari orasida qiynalardi ham. Vaqt qancha ko‘p o‘tgan sari, o‘zini shu to‘r bilan qancha o‘ralganini ko‘rgan sari, bu to‘rdan qutulish
--	---

	emas, bu to‘r erkiga halaqit berayotganmi-yo‘qligini tekshirib ko‘rish istagi ham shuncha kuchayib borardi.
Когда княгиня через пять минут вошла в комнату, она нашла их уже совершенно помирившимися.	Besh daqiqachadan keyin knyaginya uyga kirganda ular allaqachon apoq-chapoq bo‘lib ketishgan edi.
Через три минуты, не глядя на часы, чтобы не растравлять раны, Левин бегом бежал по коридору.	Toza ko‘ylak olib kelishga Kuzmaning fahmi yetmabdi. Levin uch daqiqadan so‘ng, yarasiga tuz sepib o‘tirmaslik uchun soatiga ham qaramasdan, yo‘lakda chopib chiqib ketdi.

The translation units stored in the translation memory of a parallel corpus display different forms of the searched word in concordance, that is, in its contextual environments within the corpus window. In literary texts, examples such as: *милое поэтическое настроение* – delightful poetic mood; *любовными сетями* – nets of love; *ну что за охота спать* – oh, what’s the use of sleeping; *совершенно помирившимися* – completely reconciled, on very friendly terms; *не растравлять раны* – not to rub salt into the wound are included in the translation memory in their corresponding forms.

Analysis and results: With the participation of members of the Department of Computational Linguistics and Applied Linguistics at the National University of Uzbekistan (project leader Doctor of Philological Sciences, Professor N. Z. Abdurakhmonova; members PhD, Associate Professor A. R. Iskandarova; Z. G. Qodirova; project duration: May 1, 2024 –

April 30, 2025), the innovative project No. IL-7823051524 titled “PARATRANSLATOR: creation of a contextological electronic translation dictionary platform based on a parallel corpus” was implemented. The aim of this project is to create an electronic dictionary platform that provides context-sensitive translation and is focused on developing a multilingual translation system.

The PARATRANSLATOR database includes 10 languages: Uzbek, English, Russian, Turkish, French, Spanish, Karakalpak, Japanese, Korean, and German, and contains 960,000 aligned texts belonging to artistic, scientific, official, and journalistic styles. The most important feature of the electronic translation dictionary created within the project, compared to traditional dictionaries, is that it provides translation variants of words based on real textual data. The working mechanism of the PARATRANSLATOR platform is as follows: a user enters a word or phrase; the system identifies relevant micro-contexts and displays translation variants in the corpus window, along with their frequency, concordance, usage location, and contextual form. As a result, the user selects the most appropriate translation equivalent.

The difference between the PARATRANSLATOR contextological dictionary and conventional dictionaries is that a conventional dictionary translates at the word level, whereas a contextological dictionary operates at the text level; in a conventional dictionary, the number of translation variants is limited, while in a context-based electronic translation dictionary, a large set of real-text-based translation units is available and displayed in the system interface. This expands the possibilities for scientific observation, selection, and analysis for researchers and translators.

The work carried out within the “PARATRANSLATOR: creation of a contextological electronic translation dictionary platform based on a parallel corpus” project enriches the digital resources of the Uzbek language, contributes to the development of national translation infrastructure, and serves

the advancement of corpus linguistics, translation studies, linguodidactics, and comparative linguistics.

Conclusions: In order to transform Turkic language corpora into a richly resourced open-platform linguistic base, it is necessary to increase the number of highly qualified specialists in the field of computational linguistics. Only in this way will it be possible to include texts from all subcodes of national languages in corpora. Joint research projects between national and international scholars should be further developed and centralized.

The problems related to the size of Turkic language corpora and the capabilities of search systems are connected to the limited number of specialists in the field. An open-platform, richly resourced electronic corpus of Turkic languages preserves the naturalness of language, increases the number of learners and researchers of Turkic languages abroad, enhances their international prestige, and serves as a linguistic base for large-scale scientific research.

REFERENCES:

1. Abduraxmonova N.Z. O‘zbek tili elektron korpusining kompyuter modellari. Filol.fan.dok.(DSc) diss.avtoref.–T.2021.–B.32.
2. O‘sha manba. –B.32.
3. Хусаинов Р. Татарский национальный корпус: структура и принципы построения. – Казан: Институт языка, литературы и искусства им. Г. Ибрагимова, 2012;
4. Гатиатуллин Р. Создание и использование параллельных корпусов для тюркских языков: Компьютерная лингвистика и интеллектуальные технологии, 2014, №13: с.120-128;
5. Ешпанов Р. KazParC: Kazakh Parallel Corpus for Machine Translation. LREC-COLING 2024. – ELRA, 2024, 9633–9644 b .

6. Abduraxmonova N.Z. O‘zbek tili elektron korpusining kompyuter modellari. Filol.fan.dok.(DSc) diss.avtoref.–T.2021.–B.32.
7. V.Zaxarov, B.Mengliyev, Sh.Xamroyeva. Korpus lingvistikasi: korpus tuzish va undan foydalanish. O‘quv qo‘llanma.–T.2021. –185 b.
8. Лев Толстой. Анна Каренина.Т. 2019. 1-4 қисм– 640 б; 5-8 қисм–560 б.