

**TURKOLOGIYADA PARALLEL KORPUSLARNING
INTERFEYISI, QIDIRUV TIZIMI****Aybibi ISKANDAROVA**

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

E-mail: aybibiiskandarova1962@gmail.com,

Hakimova Nargiza Feruz qizi

Second-year student, Tashkent State Transport University

hakimovanargiza1234@gmail.com

Annotatsiya: Ushbu maqolada turkologiyada ilk korpuslarning yaratilishi, ularning shakllanishi haqida ma'lumot beriladi. O'zbek, qozoq, qirg'iz, tatar, turk tili korpuslarining interfeysi, qidiruv tizimi tahlil qilinadi, parallel korpuslar doirasida filologik tadqiqotlarni olib borishdagi imkoniyatlari haqida fikr bildiriladi.

Kalit so'zlar: korpus lingvistikasi, parallel korpus, interfeys, qidiruv tizimi, tarjima birliklari bazasi.

**ИНТЕРФЕЙС ПАРАЛЛЕЛЬНЫХ КОРПУСОВ В
ТЮРКОЛОГИИ, ПОИСКОВАЯ СИСТЕМА**

Аннотация: В данной статье представлена информация о создании и формировании первых корпусов в тюркологии. Анализируется интерфейс корпусов узбекского, казахского, кыргызского, татарского, турецкого языков, в поисковой системе высказывается мнение о возможностях проведения филологических исследований в рамках параллельных корпусов.

Ключевые слова: корпусная лингвистика, параллельный корпус, интерфейс, поисковая система, база единиц перевода.

INTERFACE AND SEARCH SYSTEM OF PARALLEL CORPORA IN TURKOLOGY

Abstract: This article provides information on the creation of the first corpora in Turkology and their formation process. It analyzes the interface and search system of Uzbek, Kazakh, Kyrgyz, Tatar, and Turkish language corpora, and presents views on the possibilities of conducting philological research within parallel corpora.

Keywords: corpus linguistics, parallel corpus, interface, search system, translation unit's database.

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

Literature Review and Methodology. A parallel corpus serves as a linguistic database for philological research. It is a segmented electronic analogue of texts used for observing interlingual equivalence boundaries, comparatively studying original and translated texts, understanding translation strategies, and investigating the contextual realization of linguocultural units.

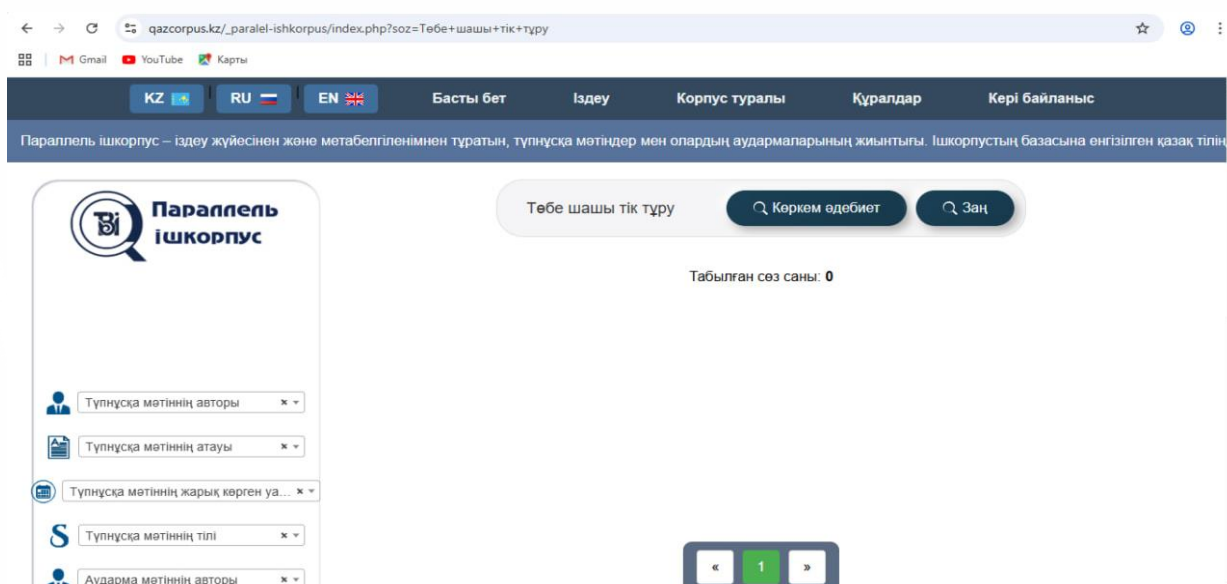
According to Sh. Khamroyeva, a parallel corpus enables the observation of the structural organization of the target language in translation, as well as the

contextual potential of idioms and lexical units. In other words, it is significant in analyzing the structures of two languages, searching for equivalents of original texts in translation studies, machine translation, linguodidactics, and lexicography [1.81].

In her article entitled “Корпус параллельных текстов в обучение иностранному языку,” A.A. Kokoreva discusses the importance of parallel corpora in the educational process. The scholar emphasizes that in learning professional vocabulary, the following possibilities of a parallel corpus are particularly useful: “The typical difference of a parallel corpus from other language corpora is its multilingual nature. This feature is beneficial in developing lexical skills in language learners. Using a parallel corpus, it is possible to effectively teach professional vocabulary to students by taking into account the following possibilities” [2.59]:

- a) the corpus contains professionally oriented written sources;
- b) the corpus includes original texts and their translations into other languages;
- c) the corpus develops skills of translating from the target language into the native language and vice versa;
- d) the corpus includes samples of genres created in a scientific style; therefore, translation skills in scientific style can be developed through the use of a parallel corpus;
- e) illustrative corpora should be involved in the educational process (corpora may be research-oriented or illustrative depending on their purpose), because the availability of numerous illustrative examples for professional vocabulary learning clarifies the studied issue;
- f) modern professional lexical units can easily be found in the corpus.

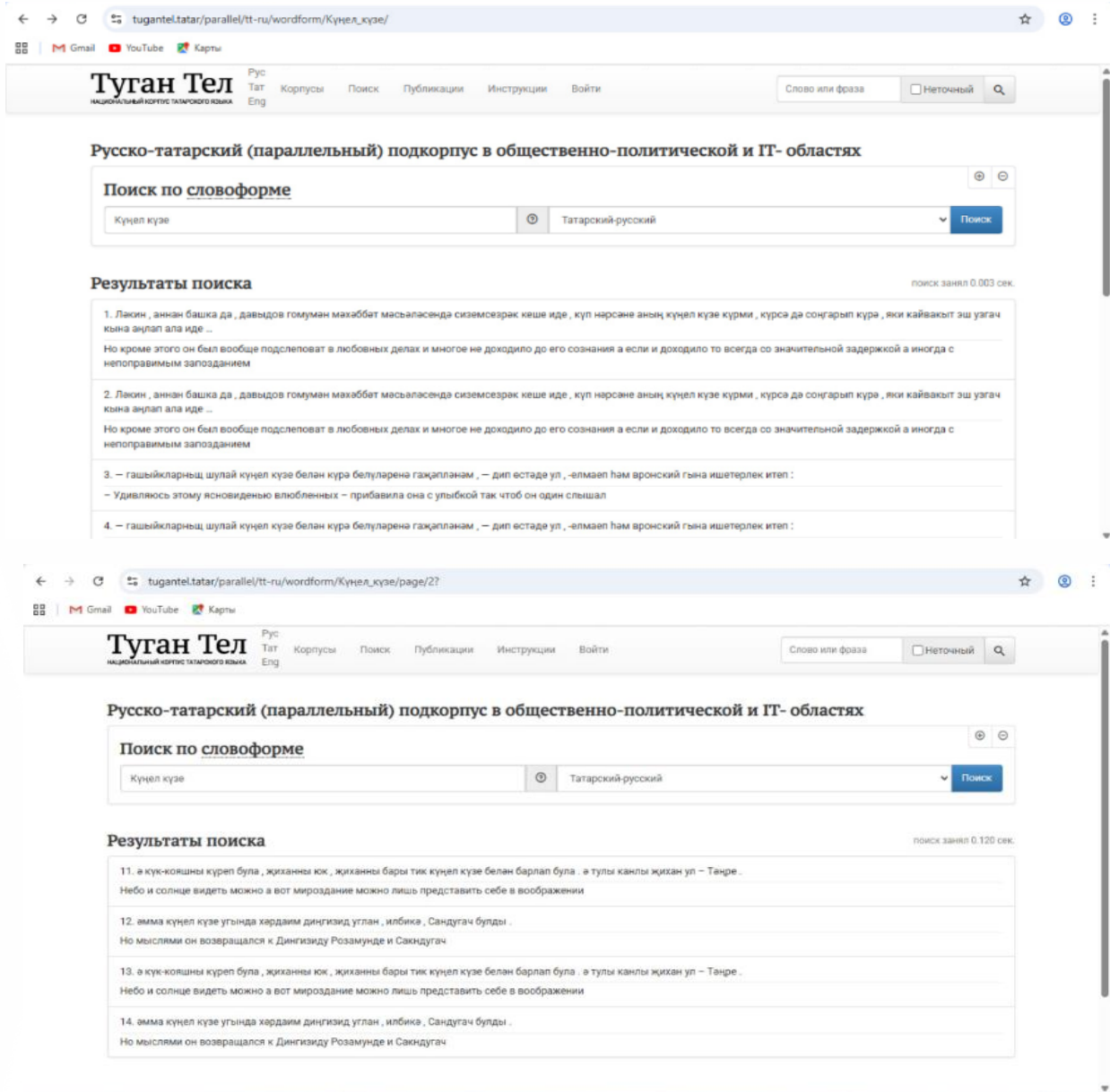
Analysis and Results: When analyzing language units through the search system of Turkic language corpora, it became evident that their capabilities are, to some extent, limited. In the Turkish language corpus interface (<https://www.tnc.org.tr/>), there is no information about the existence of a parallel subcorpus, and it can be concluded that a parallel corpus has not been created for this language corpus.



In the Kazakh language parallel corpus search system (<https://qazcorpus.kz/indexru.php>), when the phrase “Төбе шашы тік тұру,” “Tepa sochi tikka bo‘ldi” was entered, the statistics showed the number 0, meaning that this expression was not found among the uploaded content. This indicates that the Kazakh parallel corpus contains a limited amount of textual data.

The Tugan Til corpus is one of the more developing corpora among Turkic language corpora. When the phrase “Күңел күзе” / “ko‘ngil ko‘zi” was entered into the Tugan Til parallel corpus search system (<https://tugantel.tatar/>), statistical data showed that there are 14 Tatar and Russian forms of this expression.





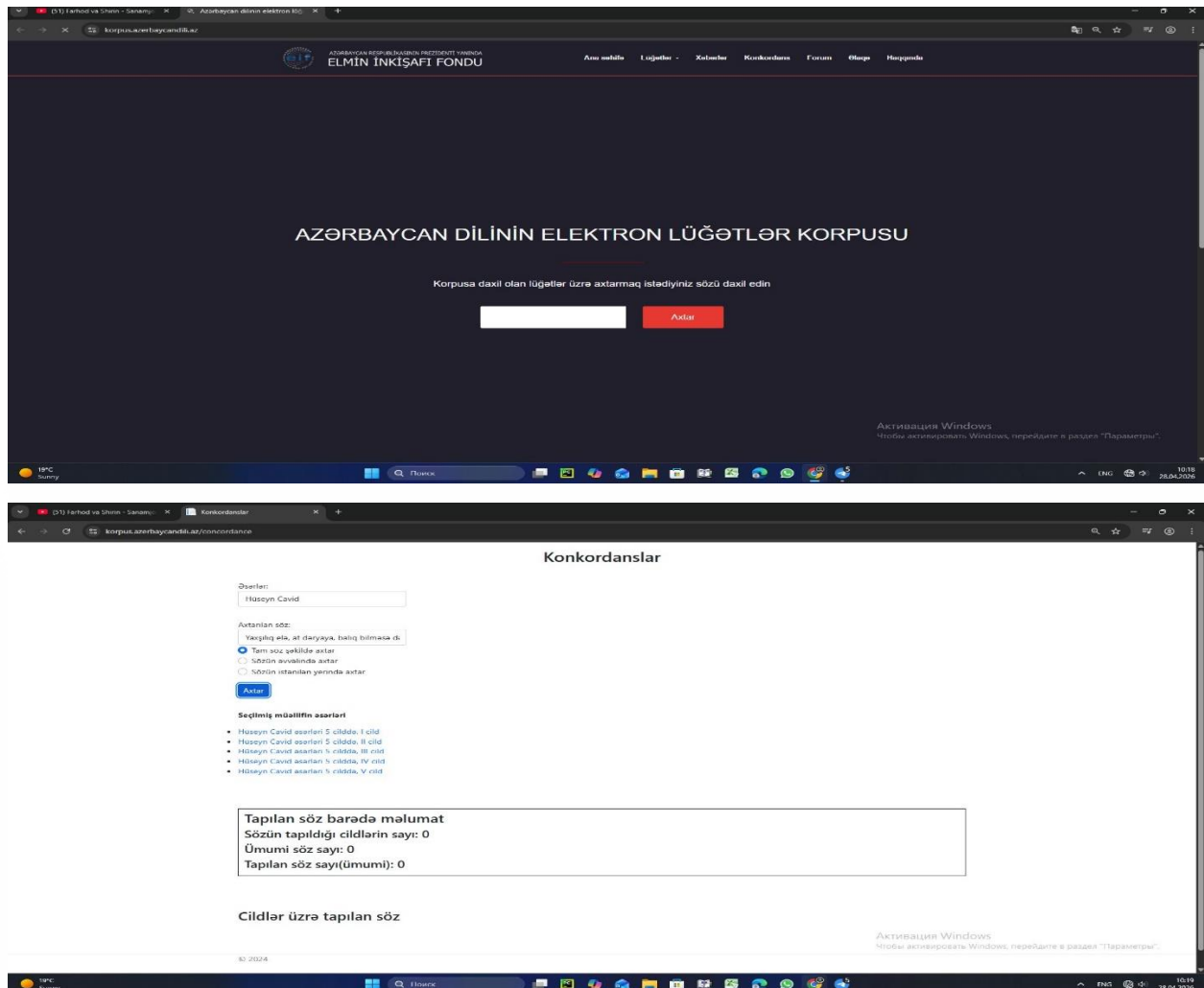
The screenshot shows the Tugan Tel website interface. The top navigation bar includes links for 'Рус', 'Тат', and 'Eng' languages, and buttons for 'Корпусы', 'Поиск', 'Публикации', 'Инструкции', and 'Войти'. The main heading is 'Русско-татарский (параллельный) подкорпус в общественно-политической и IT- областях'. Below this is a search bar with the text 'Поиск по словоформе' and a dropdown menu set to 'Татарский-русский'. The search results are displayed under the heading 'Результаты поиска' and show four items, each with a numbered list of search results. The first item is '1. Лакин, аннан башка да, давидов гомуман махаббат масъаласында сиземсезрак кеше иде, күп нәрсәне аның күңел күзе күрми, күрсә да сонгарып күрә, яки кайвакыт эш узгач кына аңлап ала иде...'. The second item is '2. Лакин, аннан башка да, давидов гомуман махаббат масъаласында сиземсезрак кеше иде, күп нәрсәне аның күңел күзе күрми, күрсә да сонгарып күрә, яки кайвакыт эш узгач кына аңлап ала иде...'. The third item is '3. – гашыйкларның шулай күңел күзе белән күрә белүләренә гаҗапланам, – дип естаде ул, –елмаеп һәм аронский гына иштерлек итеп: – Удивляюсь этому ясновиденью влюбленных – приваила она с улыбкой так чтоб он один слышал'. The fourth item is '4. – гашыйкларның шулай күңел күзе белән күрә белүләренә гаҗапланам, – дип естаде ул, –елмаеп һәм аронский гына иштерлек итеп:'. The search results are displayed in a table format with columns for the search results and the corresponding text.

This corpus is one of the continuously updated corpora among Turkic language corpora, and it demonstrates the possibility of observing the real usage of words, expressions, and grammatical forms, as well as conducting empirical analysis of their various forms.

In the Azerbaijani electronic dictionary corpus search system, when the proverb “Yaxşılıq elə, at dəryaya, balıq bilməsə də, Xalq bilər,” “Yaxshilik



qıl-u daryoga ot, baliq bilmasa ham, Xoliq biladi” was entered, no information about the searched unit was found.

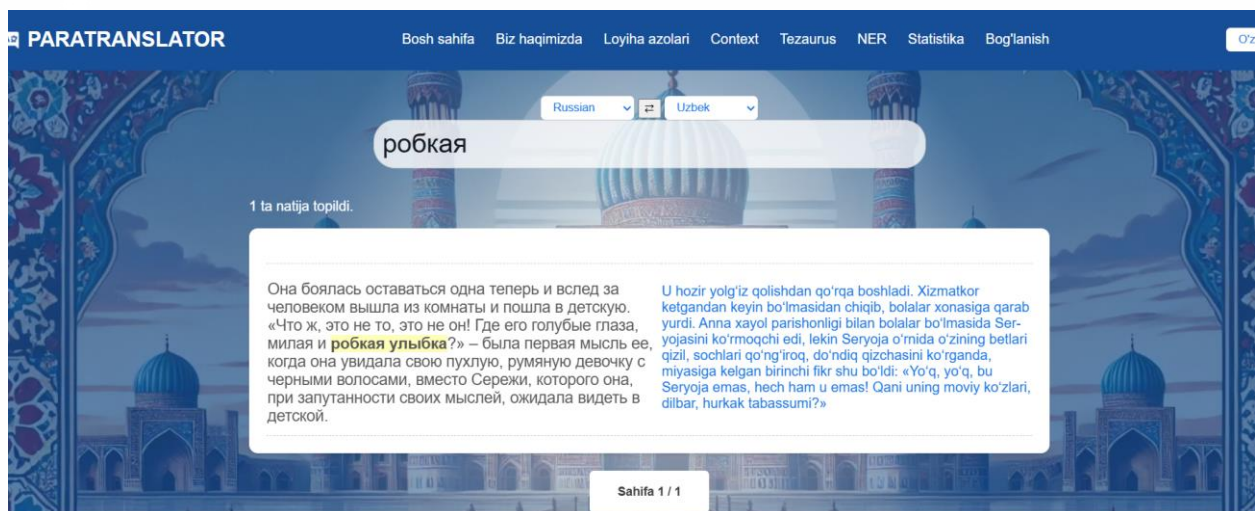


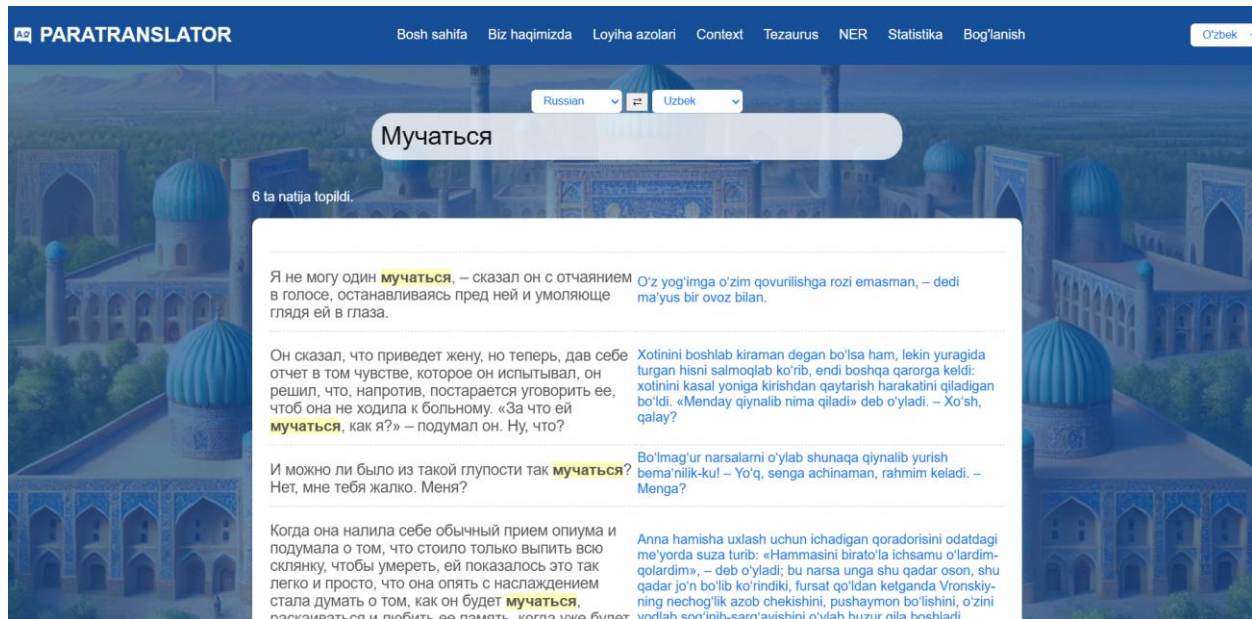
The Azerbaijani electronic dictionary corpus does not include a parallel corpus. Its interface consists of sections such as Home, Dictionaries, News, Concordance, Forum, Contact, and About.

The Uzbek language corpus (<https://uzbekcorpus.uz/education>), in the project participants section, indicates that it was developed in 2020 and consists of the following structure: Home, Project Members, Search, Morphological Analyzer, Linguistic Analyzer, Parallel Corpus, Educational Corpus, Author Corpus, Literary Corpus, Linguistic Resources, Learning Dictionaries, Thesaurus, User Guide, About the Corpus, and Scientific Publications.



The parallel subcorpus of the Uzbek language corpus contains a relatively small volume of segmented texts. With the participation of the Department of Computational Linguistics and Applied Linguistics at the National University of Uzbekistan (project leader: Dr. N. Z. Abdurakhmonova; members: Assoc. Prof. A. R. Iskandarova, Z. G. Qodirova; project period: May 1, 2024 – April 30, 2025), the innovative project No. IL-7823051524 titled “PARATRANSLATOR: Creation of a Contextual Electronic Translation Dictionary Platform Based on a Parallel Corpus” was implemented. The aim of this project is to develop an electronic dictionary platform that provides context-based translation equivalents, focusing on the formation of a multilingual translation system. The Paratranslator database includes 10 languages: Uzbek, English, Russian, Turkish, French, Spanish, Karakalpak, Japanese, Korean, and German, with 960,000 aligned texts belonging to literary, scientific, official, and journalistic styles.





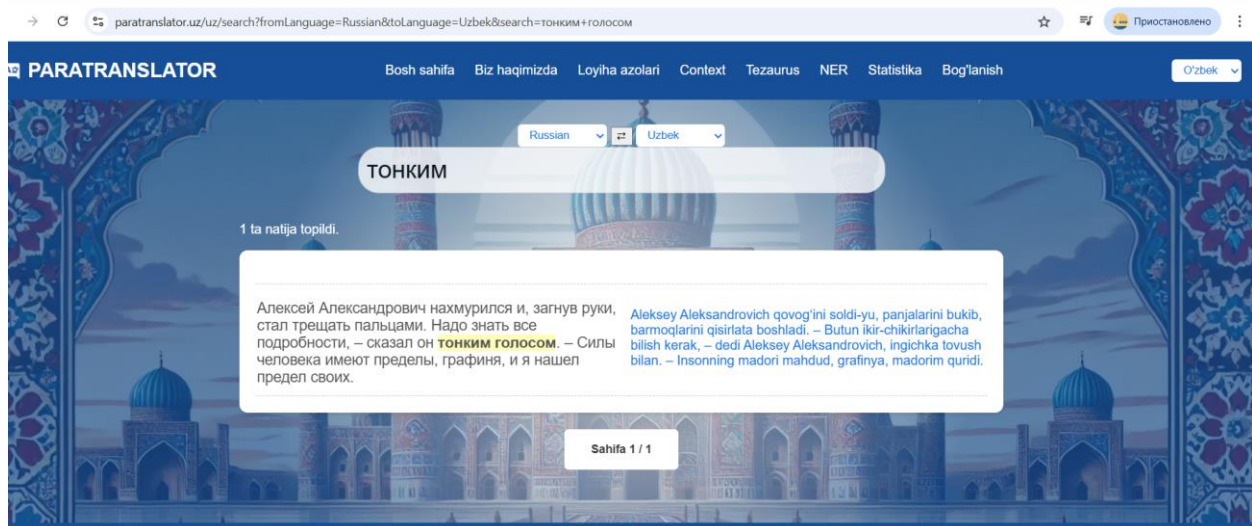
Within the framework of the project, the most important feature of the created electronic translation dictionary, compared to traditional dictionaries, is that it provides translation variants of words based on real textual data. The working mechanism of the Paratranslator platform is as follows: the user enters a word or phrase. The system retrieves relevant micro-contexts for the query and displays translation variants in the corpus window, along with their frequency, concordance—usage positions and contexts. As a result, the user selects the most appropriate translation equivalent.

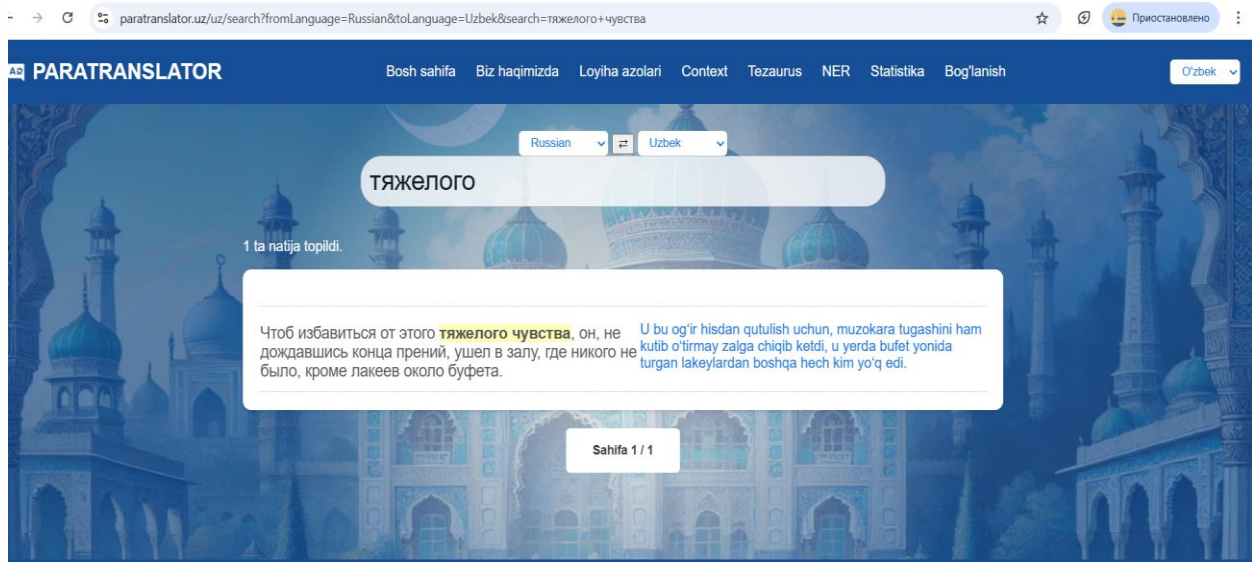
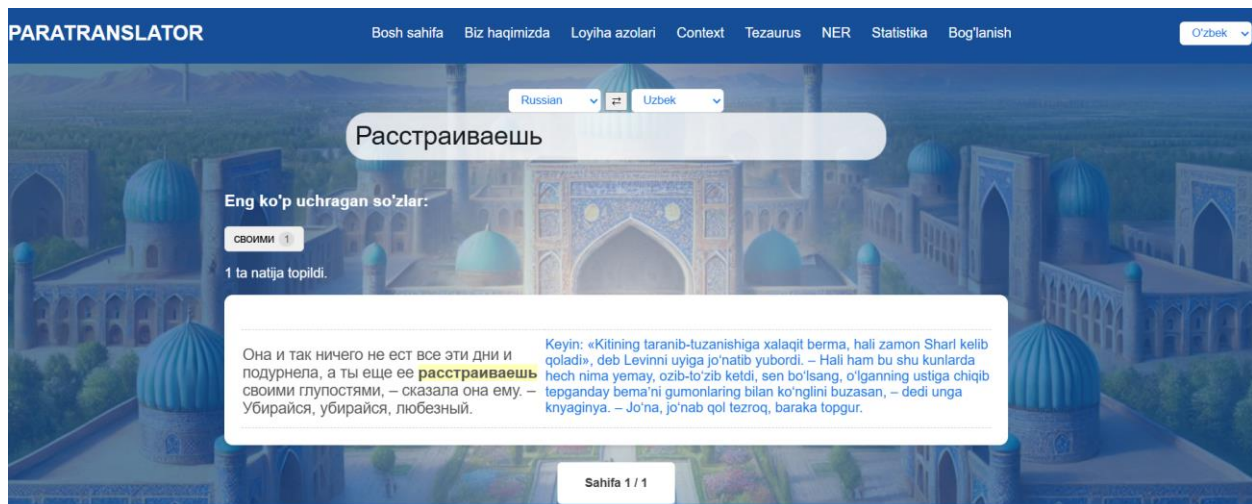
The difference between the Paratranslator contextual dictionary and conventional dictionaries is that traditional dictionaries translate at the word level, whereas the contextual dictionary operates at the text level; traditional dictionaries provide a limited number of translation equivalents, while the contextual electronic translation dictionary contains a large corpus-based set of real-text translation units, which are displayed in the system interface. This expands scientific observation and analytical possibilities for researchers and translators.



The work carried out within the framework of the project “PARATRANSLATOR: Creation of a Contextual Electronic Translation Dictionary Platform Based on a Parallel Corpus” contributes to the enrichment of Uzbek digital linguistic resources, the development of national translation infrastructure, and serves the advancement of corpus linguistics, translation studies, linguodidactics, and comparative linguistics.

In Paratranslator, Russian–Uzbek bitexts of the novel “*Anna Karenina*” in both original and translated versions have been included. When polysemous words and phrases are entered into the Paratranslator search system, they are displayed in the form of micro-texts.





Conclusions. In order to transform Turkic language corpora into a linguistically rich, resource-based, and open-platform system, it is necessary to increase the number of highly qualified specialists in the field of computational linguistics. In higher education institutions, bachelor's degree programs in Computational Linguistics within philology and language teaching fields should be established. At the master's level, the number of allocated quotas for this specialty should be increased. Only under these conditions will it be possible to include texts representing all subcodes of national languages into



corpora. It is also necessary to develop centralized national and international project activities in cooperation with researchers.

The problems related to the size of Turkic language corpora and the limitations of their search systems are directly connected to the shortage of specialists in this field. An open-platform, resource-rich electronic corpus of Turkic languages will help preserve linguistic naturalness, increase the number of learners and researchers of Turkic languages abroad, enhance their international prestige, and in the future serve as a linguistic database for scientific research.

REFERENCES

1. V.Zaxarov, B.Mengliyev, Sh.Xamroyeva. Korpus lingvistikasi: korpus tuzish va undan foydalanish. O'quv qo'llanma.–T.2021. –185 b.
2. Кокорева А.А.Корпус параллельных текстов в обучении иностранному языку. – Вестник ТГУ (Тамбовский государственный университет), выпуск 2 (118)б. 2013. –С. 57-62.
3. <https://www.tnc.org.tr/>
4. <https://qazcorpus.kz/indexru.php>
5. <https://tugantel.tatar/>
6. (<https://uzbekcorpus.uz/education>
7. Лев Толстой. Анна Каренина.Т. 2019. 1-4 қисм– 640 б; 5-8 қисм–560 б.