Vital Annex: International Journal of Novel Research in Advanced Sciences (IJNRAS)
Volume: 01 Issue: 05 | 2022 ISSN: 2751-756X
http://innosci.org

# Development of a Parallel Corpus of the Uzbek and Russian Languages

***Avezov Sukhrob Sobirovich***
*Lecturer, Bukhara State University, Uzbekistan, Bukhara*

**Annotation:** A description of the development of a parallel Uzbek-Russian corpus is presented. The general mechanism of the corpus, the structure of the database of texts, text processing algorithms, as well as automatic control of the corpus using the author's program Uz-Rus-Corp are considered. The development of a parallel corpus will contribute to the organization of machine translation of texts from Uzbek into Russian.

**Keywords:** computational linguistics, corpus linguistics, databases, development, MongoDB, NoSQL, lexicology, machine translation.

## 1. Introduction

The article discusses the results obtained in the framework of the research work "Basic principles for creating the Uzbek-Russian and Russian-Uzbek online platform of parallel corpora (based on the novel by A. Kadiri "Past Days")" of the Department of Russian Literary Studies of the Bukhara State University.

**2. The main part.** At the current stage of intensive development of our country, "the need for Uzbekistan to become competitive on a global scale in the field of science, intellectual potential, modern personnel, high technologies"[1] has defined new tasks for Uzbek linguistics, which consist in raising theoretical research to the level of the world standard. In this regard, studies on the creation of a parallel corpus of the Uzbek-Russian and Russian-Uzbek online platforms reveal not only the rich artistic potential of the two languages, but also directly reflect similar features and definitions of national cultures, national thinking and worldview. Being the state language of Uzbekistan, it is also common in Kazakhstan, Kyrgyzstan and Afghanistan. It is the language of ancient culture and centuries-old literary tradition [2, p. four]. Along with the Uzbek language, much attention is paid to the Russian language in Uzbekistan. It is one of the East Slavic languages, the national language of the Russian people, which is one of the most widely spoken languages in the world, the sixth among all the languages of the world in terms of the total number of speakers and the eighth in the number of countries using it as a second state language. A large number of scientific works are devoted to comparative study at different stages and language tiers. One of the main problems facing linguists is the creation of the National Corpus of the Uzbek language, as well as its parallel corpus. A large collection of parallel texts is called a "parallel corpus". The development of a parallel corpus requires alignment of the parallel text with identification of corresponding sentences in both halves of the parallel text. [3,c. 5].

In practice, parallel corpora are used to get a translation of a text in a specific format. From a scientific point of view, the formation of a parallel corpus makes it possible to implement important scientific and research tasks in the field of computational linguistics.

The development of the Uzbek-Russian parallel corpus will increase the productivity of relations between the Uzbek and Russian peoples, and will help the population of the two countries to increase their knowledge of both languages.

This paper discusses the description of the purpose, purpose and process of developing the Uzbek-Russian parallel corpus. Further ways of using the corpus in machine translation are proposed. Today, one of the most important tasks in the field of computational linguistics in Uzbekistan is machine translation from Uzbek into other languages and vice versa. In particular, for the translation of texts from Uzbek into Russian, it requires the creation of a parallel corpus, proving the relevance of the task.

Also, against the background of the created corpus, it becomes possible to conduct a number of statistical studies, processing textual information, and studying the elements of the languages used.

The parallel corpus includes texts from A. Kadiri's novel "Past Days". As part of the creation of a parallel corpus, the following tasks were performed:

1) a representative sample of the text;

2) pre-treatment

3) description of the sources of the text;

2) alignment of texts;

3) development of text processing algorithms;

4) creation of the Uz-Rus-Corp program with search capabilities;

5) input of texts into a parallel corpus;

6) statistical analysis of data;

7) development of an experimental machine translation module. [4, p. 3]

## 3. Corpus database structure

The database of the parallel corpus includes the following information: language, type, title of the document, author, name, translator, source and year of publication of the text. The database structure is developed using the MongoDB database management system, which allows you to store and process a large amount of data. The main parallel body table consists of the following fields:

sn is a serial number;

uz - text in Tajik;

rus - translation of the text in Russian;

tw – text type;

mainText is the name of the document;

author - author;

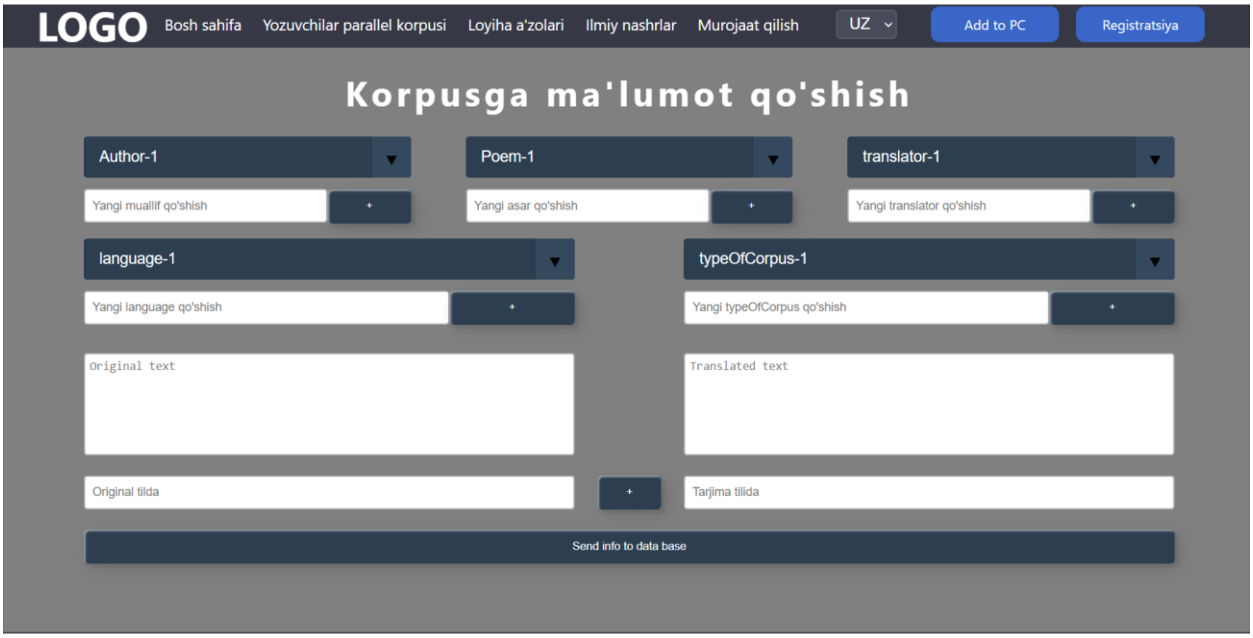
inter

textSource - text source;

datePubl is the year of publication.

## 4. Program Uz-Rus-Corp

To create a parallel corpus and directly form a database of texts, the author's program Uz-Rus-Corp is used. This program can also be used to create and form parallel corpora of other languages. With the help of the program, a basis of parallel texts in two or more languages is created with the control functions of these corpora. The program supports various fonts of corresponding languages in text and supports Unicode encoding. Provides semi-automatic comparison of texts and text alignment by words, by sentences or by paragraph. It is also possible to process text according to the proposed punctuation marks.



Korpusga ma'lumot qo'shish

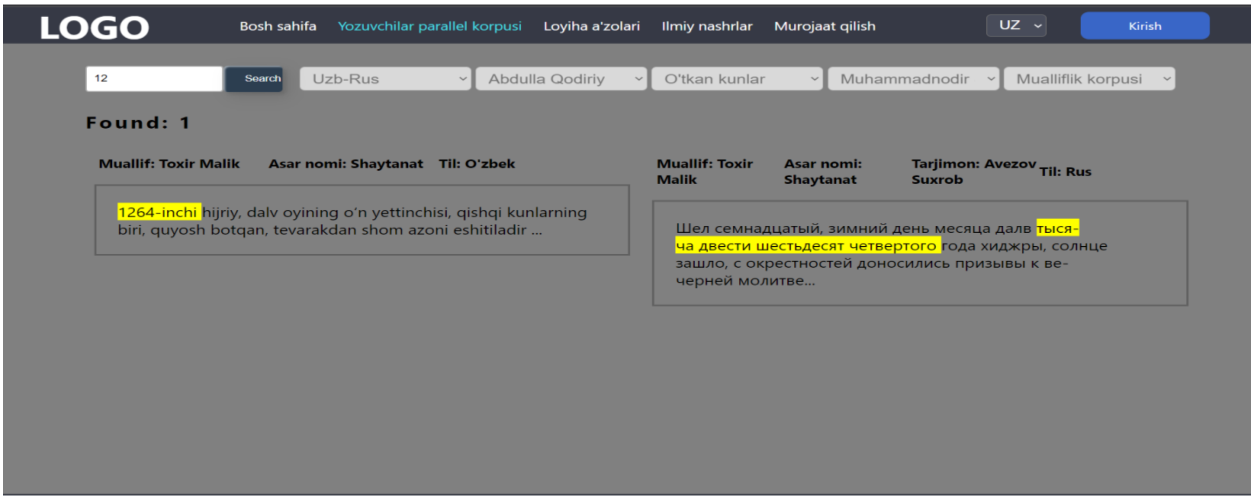
Author-1 | Poem-1 | translator-1

language-1 | typeOfCorpus-1

Figure 1. Interface for aligning sentences in the Uz-Rus-Corp program

It should be noted that the Uz-Rus-Corp program makes it possible to sort, filter and search for information in a parallel corpus. The listed control functions in a parallel package provide the possibility of machine translation. Software modules are developed using various algorithms for processing text data. Unigrams, bigrams, trigrams of words are chosen as search elements in the text. You can also use different output modes in TXT, XLSX and HTML formats. (See Figure 1.)

The most scientific and theoretical aspect of using a parallel corpus can be noted as statistical data analysis. Possibilities are provided to count the frequencies of elements in the text, such as syllables, words and phrases. The listed functions make it possible to obtain the necessary characteristics about the nature of corpus languages.



12 | Uzb-Rus | Abdulla Qodiriy | O'tkan kunlar | Muhammadnodir | Mualliflik korpusi

Found: 1

Muallif: Toxir Malik Asar nomi: Shaytanat Til: O'zbek

1264-inchi hijriy, dalv oyining o'n yettinchisi, qishqi kunlarning biri, quyosh botqan, tevarakdan shom azoni eshitiladir ...

Muallif: Toxir Malik Asar nomi: Shaytanat Tarjimon: Avezov Suxrob Til: Rus

Шел семнадцатый, зимний день месяца далв тысяча двести шестьдесят четвертого года хиджры, солнце зашло, с окрестностей доносились призывы к вечерней молитве...

Figure 2. Results interface in the Uz-Rus-Corp program

The Uz-Rus-Corp program uses the following search algorithms and methods: simple linear text search; advanced search using regular expressions; search by text elements; parallel search in two languages. The main search tool is the non-structural query language NoSQL. The possibilities of semi-automatic morphological analysis using regular expressions were also used. The search function is expanded with the ability to find word forms and word usages, keywords by lemmas [5, p. 5].

**4. Conclusion**

Summing up the results obtained, it can be noted that the developed parallel body requires improvement. We plan to improve the text preprocessing functionality in the future. It is also planned to use automatic morphological analysis. More powerful tools are being created, in particular algorithms and search methods for displaying results. Terminological and special dictionaries are formed on the basis of the parallel corpus. Various statistical methods of text analysis are being developed on the basis of more significant parameters and elements of textual information. All the results obtained will be implemented and based as a single model for the development of new parallel corpora related to the Uzbek language [6, c.2].

**Bibliography.**

1. Ш.Мирзиёев. Ўзбекистон Республикаси Президенти Ш.Мирзиёевнинг 2017 йил 22 декабрдаги Олий мажлисга мурожаатномаси // Халқ сўзи, 2017 йил 23 декабрь. – № 258 (6952).
2. Nigmatova L. K. Language and cultural issues in uzbek vocabulary //Scientific reports of Bukhara State University. – 2021. – Т. 5. – №. 1. – С. 30-49.
3. Захаров В., Богданова С. Корпусная лингвистика. – Litres, 2022.
4. Шарипов С.С. Таржимавий лексикографиянинг тарихий ва хронологик хусусиятлари //Центр научных публикаций. Бухара, 2022. – Т. 15. – №. 15.
5. Sobirovich A. S. Lecturer at the Department of Russian Language and Literature Bukhara State University //Scientific reports of bukhara state university. – С. 86.
6. Авезов С.С. О корпусной лингвистике , трудностях перевода и принципах организации параллельных корпусов текстов.// «Ўзбек тилининг миллий корпуси: муаммо ва вазифалар» //Халқаро илмий-амалий конференция материаллари. Тошкент ,2022.