

**TASHKENT STATE UNIVERSITY OF UZBEK LANGUAGE AND
LITERATURE**

**LEARNING CORPUS OF UZBEK LANGUAGE: STRUCTURE,
CONTENT, OPPORTUNITIES**

B.Mengliev, Sh.Khamroeva, D.Elova,
Tashkent, Uzbekistan.

Introduction. During the years of independence, although a number of studies were conducted in computer linguistics as to achieve automatic translation, natural language processing (Uzbek language), the practice of creating language corpus was not put into practice.

In recent years, the corpus of the Uzbek language, their features, types, principles of formation of the author's corpus [Khamroeva, 2018],

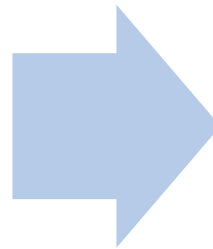
the problems of graphic analysis of Uzbek language units [Abjalova, 2019; 2018], the principles of creating a linguistic base of the language corpus [Eshmuminov, 2019] began to be studied in a monographic plan.

Also, a number of articles were published on the linguistic basis of semantic tagging of noun units for Uzbek language corpora [Akhmedova, 2019], the problems of creating an Uzbek-English parallel corpus [Mengliev, 2019].

As a result of such theoretical research, practical projects were developed, the results of which were directed to the development of programming and electronic products.




The main part. The process of studying the problems of computer and corpus linguistics in world linguistics has accelerated. The use of machine translation, thesaurus, virtual dictionaries has expanded, many types of language corporations have emerged that differ in purpose. Natural language processing, linguistic analysis program as lemmatizator, tagger, stemmer, parser, orthocorrector as well as the creation of their linguistic supply is a topical issue nowadays computer linguistics. Among them, learning corpuses with lingvodidactic properties have a special place.



At the Tashkent State University of Uzbek Language and Literature named Alisher Navai for the first time in the framework of the practical project AM-FZ-201908172 “Creation of the learning corps of Uzbek language” for 2020-2021 began work on the learning corps of Uzbek language. There is a discussion of the search system of the Uzbek language learning corps, which is the product of this project, its specific features and the results achieved.

The purpose of this project



is to create an learning corpus of Uzbek language, to attach to the Uzbek language learning corpus by placing educational dictionaries in the search engine.

Learning corps and its role in linguodidactics. Learning corpus is a linguistic corpus with lingvodidactic properties, the materials of which are focused on the study of a particular language. The Uzbek language learning corpus is a corpus of the Uzbek language, operating in the form of a special site, containing electronic texts of linguodidactic nature, aimed at teaching the possibilities of the Uzbek language.

The field of effective use of the corpus is lingvodidactics or teaching uzbek language; it is equally important in the study as the native language, as a foreign language. In language teaching, the corpus is very useful in showing an array of examples to fully demonstrate the richness of vocabulary, to explain the possibility of using a word through the grammatical construction.



The constant updating of the example, which is important for language teaching, and the ability and opportunity to reflect it are available only via corpus.



The teacher can find new, convincing, infinite and diverse examples here, it is not difficult to define the task, exercises, in a few minutes can prepare a task consisting of new examples on the topic.



It is possible to sort the texts in the corpus: the example can be separated not only from all the texts, but also from a fragment that is interesting and necessary for the researcher.



Thus, corpus texts allow you to select a specific period of, a specific type of text



Therefore, it opens up a wider range of focused learning opportunities.



One of the main features of the corpus is that it can be constantly enriched with texts of different themes / genres.

Composition and structure of the Uzbek learning corps. Official, scientific-popular, artistic and journalistic texts in modern Uzbek literary language were selected as the material of the Uzbek language learning corps. The corps fully demonstrates its representativeness: the materials it contains are diverse: the text's genre have the following proportions:

1. The artistic part of the texts of the Uzbek learning corps, which has been collected till this time includes works of genres such as short stories, novels, feuilletons, memoirs by Abdulla Qahhor, Said Ahmad, Togay Murod, Askad Mukhtor.

2. Texts in journalistic style include journalistic articles on websites as kun.uz, daryo.uz.

3. The official style materials of the corps are the texts of the Constitution of the Republic of Uzbekistan, Laws, Resolutions and Decrees of the President of the Republic of Uzbekistan, various codes available on the lex.uz. website.

4. Materials in the popular science style mainly consist of school textbooks: literature, physics, mathematics, chemistry, geography, biology, drawing, etiquette.

Fragmentation of texts, removal of characters (tables, pictures, diagrams) of uncounted parts of the text was carried out using the sublimetext program.

1. Texts of moral and educational content were selected for the Uzbek language learning corpus, placed in the corpus.

2. Existing educational dictionaries in Uzbek language are collected on one platform; subject to a convenient search engine.

In nowadays, the following practical results have been achieved in the development of the Uzbek language learning corpus:

3. Linguistic support of the Uzbek language learning corpus – a morpholexicon consisting of 32,000 lexemes placed to determine the set of words in the Uzbek language.

Currently, the “Learning Corpus of the Uzbek language” can be found at <http://uzschoolcorpara.uz/>.

As well known, a language corpus is a collection of texts placed in a written / oral, automated search engine that is stored electronically in a specific natural language.

Therefore, one of the most important parts of the language corpus is its search engine. There are several special software tools that help you get the information you need from the concordance (a relatively simple view of a search engine) or a corpus manager.

Corpus manager is a special search system that consists of several programs designed to retrieve corpus information, providing statistical information and search results in a user-friendly format. The search result is displayed in the form of a list of contexts attached to the source in the concordant-lexical link mode of the searched unit

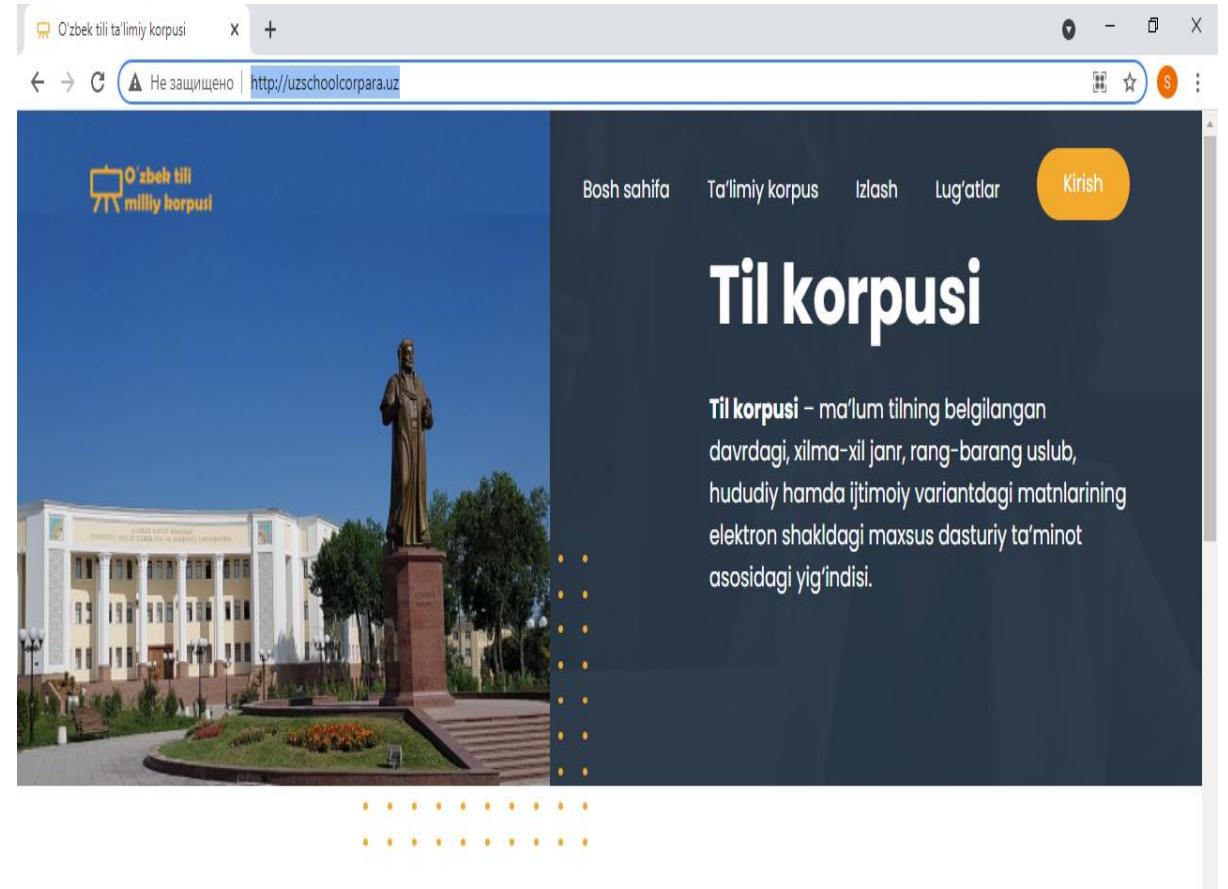
The search capabilities of the corpus manager include wordform and lemma, compound divided / undivided syntagm, compound form based on morphological features.

The most important component that introduces the user is the interface. The ease with which the user can work with the corpus depends on the interface being structured in a perfect, clear, unique way. In the development of the Uzbek language learning corps, along with the general requirements, special attention is paid to the features of the Uzbek language, the age-appropriate aspects of the student.

The Uzbek language learning corpus consists of a corpus interface, a search engine, about 75 million words, a database of more than 1 million texts and 15 electronic dictionary sources. Also, on the first page of the website of the Uzbek language learning corpus there is basic information about the corpus and its creators, you can go to any page in the menu on the right. The corpus menu consists of four sections.

The homepage consisted of a complete database of the site's search window, additional information about the text attached to it, parameters for commenting on language units for search, description of electronic lexicographic sources, team of last block corpus developers, program used, copyright of texts. Below we analyze the corpus interface (Figure 1).

Figure 1. The interface of Uzbek language learning corpus

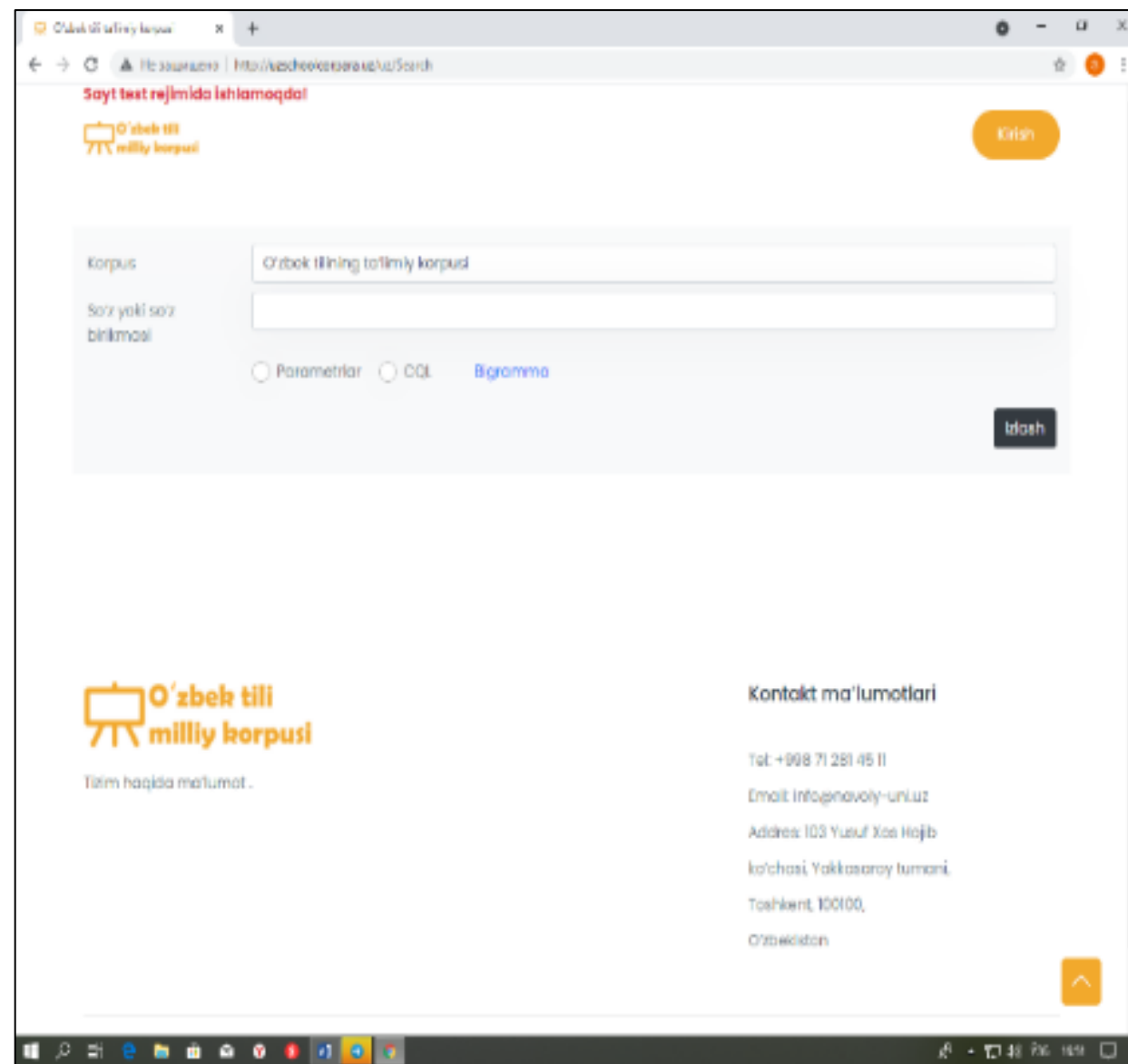


As you can see in the picture, the interface is divided into four parts: “Homepage”, “Learning corpus”, “search”, “dictionaries”.

The homepage is a common interface of Uzbek language learning corpus.

The part named *learning corpus* includes corpus information (general information, corpus map (structure and content), news archive, information about authors and project, text parameters, word frequency system, software, project published work, theoretical material) and instructions for use of the case.

The search button leads to the main page of the case. The corpus search page has the following view (Fig.2).



The corpus has a simple and complex (SQL) search form. For now, the simple search and (partially) bigram search engine is active. It is possible to search for a word or phrase by a simple search (Fig.3).

The screenshot displays a web browser window with the URL <http://uzschoolcorpara.uz/uz/Search?c=1&q=ватан>. The page header includes the text "Sayt test rejimida ishlamoqda!" and the logo for "O'zbek tili milliy korpusi". A "Kirish" button is located in the top right corner.

The search form contains the following fields and options:

- Korpus: O'zbek tilining ta'limiy korpusi
- So'z yoki so'z birikmasi: ватан
- Search options: Parametrlar, CQL, Bigramma
- Search button: Izlash

The search results are displayed under the heading "26 ta yozuvdan 20 tasi ko'rsatilmoqda!". The results are as follows:

Text snippet	Word	Context
ТОМОШАБОФ Ўтмишдан. Эй бус–бутун аёллати вайрон ўлон	ватан	, Ҳар гўшаси замонада зиндон ўлон ватан. «Рамузот»...
... нарсагина киради: 25 грамм кўрғошин!– Э, улуғ совет	ватанининг	соҳибқирони, бос тепкини!
... қизини олиб берсақ, оларсанми, шунинг билан ҳам	ватан–рўзғор	қиларсанми, чап тиззаси кўкрагинга тегса,...
... улар ўлмасин, Ўқли–қизи ебир–есир бўлмасин,	Ватани	беэга бўлиб қолмасин. Булар эмас, келиб бизга...
... Яхши ўйла, мени ёмон қилмагин, Бу зиндонни ўзинг	ватан	билмагин, Мард бўлсанг, номарднинг ишин қилмагин,...
... билан ўйнаб–куламан, Ростин айтсам, бирга	ватан	қиламан, Алп Қоражон билан даврон сураман,...
... эмас... миллат... дин... йўқ... бўлса... мен...	ватан	... қарши эмас... кетди... қарши эмасман... лекин...
... лабзи ҳалол йигитларимиз, қизларимиз, ёш–қари ҳамма	ватанпарвар	азаматлар Искандар замонасидан бери одамлар уриниб...
... бунчалик ғазаб, ғалабага шунчалик ишонч бўлган.	Ватан	деганда кўкрагида шунча меҳр–муҳаббат, ғайрат,...
... Тақдир насиб, шундай саодатга эришсалар, ўзларига	ватан	тутган маконларию жуфтларини жонларини жабборга...
... учди. Ботинда чалина бошлаган най навоси уни	Ватанга	бошлаб кетди...
... ишлайди, қувнаб яшайди. Бу шаҳарлар – бепоён	Ватанимизнинг	бир гўшаси бўлган Арманистоннинг эртанги куни...

This search function creates concordance and the results are in the wordsform and lemmas. SQL and bigram views of the search are also being developed. (Figure 4)

The screenshot shows a web browser window with two tabs: 'O'zbek tili ta'limiy korpusi' and 'n-grammalar'. The address bar shows the URL 'bigram.navoiy-uni.uz/bigrams'. The page header includes 'O'zbek tili' and a 'LOGIN' link. A sidebar on the left lists navigation options: 'O'zbek tili korpusi', 'Qidiruv', 'Bigrammalar', 'Trigrammalar', 'n-grammalar', 'Manbalar', and 'Sinonimlar'. The main content area is titled 'Bigramma bo'yicha qidiruv' and features two toggle switches: 'Leksika va grammatikani hisobga olish' and 'Punktuatsiyani hisobga olish'. Below these is a search input field labeled 'So'z izlash'. A section titled 'Leksik - grammatik qidiruv' contains a 'So'z' input field, a 'Gramm. belgilar' dropdown menu, and a search results box. The results box shows 'Gramm. belgilar' with a sub-item 'beta' and another 'Gramm. belgilar' entry. A 'Masofa' label is visible at the bottom of the search area.

In the corpus, a different view is set up by means of an SQL search engine: word search, word form search, compound search, semantic, morphological, stylistic features. When a morphological (also called a grammatical search) search is selected, a separate window opens and a search based on parameters derived from the grammatical features of the Uzbek language appears (Figure 5).

<p>Сўз туркумлари:</p> <input type="checkbox"/> От <input type="checkbox"/> Сифат <input type="checkbox"/> Сон <input type="checkbox"/> Олмош <input type="checkbox"/> Равиш <input type="checkbox"/> Феъл <input type="checkbox"/> Юклама <input type="checkbox"/> Боғловчи <input type="checkbox"/> Кўмакчи <input type="checkbox"/> Тақлид сўзлар <input type="checkbox"/> Модал сўзлар <input type="checkbox"/> Оралиқ сўзлар	<p>Келишик:</p> <input type="checkbox"/> Бош келишик <input type="checkbox"/> Қаратқич келишиги <input type="checkbox"/> Тушум келишиги <input type="checkbox"/> Жўнатлиш келишиги <input type="checkbox"/> Ўрин-пайт келишиги <input type="checkbox"/> Чиқиш келишиги	<p>Майл</p> <input type="checkbox"/> Хабар майли <input type="checkbox"/> Буйруқ-истак майли <input type="checkbox"/> шарт- майли	<p>Замон</p> <input type="checkbox"/> Ўтган замон <input type="checkbox"/> Ҳозирги замон <input type="checkbox"/> Келаси замон
<p>Атоқли отлар:</p> <input type="checkbox"/> Фамилия <input type="checkbox"/> Исм <input type="checkbox"/> шариф	<p>Сон:</p> <input type="checkbox"/> Бирлик <input type="checkbox"/> Кўплик	<p>Нисбат</p> <input type="checkbox"/> Аниқ нисбат <input type="checkbox"/> Орттирма нисбат <input type="checkbox"/> Ўзлик нисбат <input type="checkbox"/> Мажхул нисбат <input type="checkbox"/> Биргалик нисбат	<input type="checkbox"/> Бўлишли <input type="checkbox"/> бўлишсиз
		<p>Шахс-сон</p> <input type="checkbox"/> I шахс <input type="checkbox"/> II шахс <input type="checkbox"/> III шахс	<input type="checkbox"/> Ўтимли <input type="checkbox"/> ўтимсиз

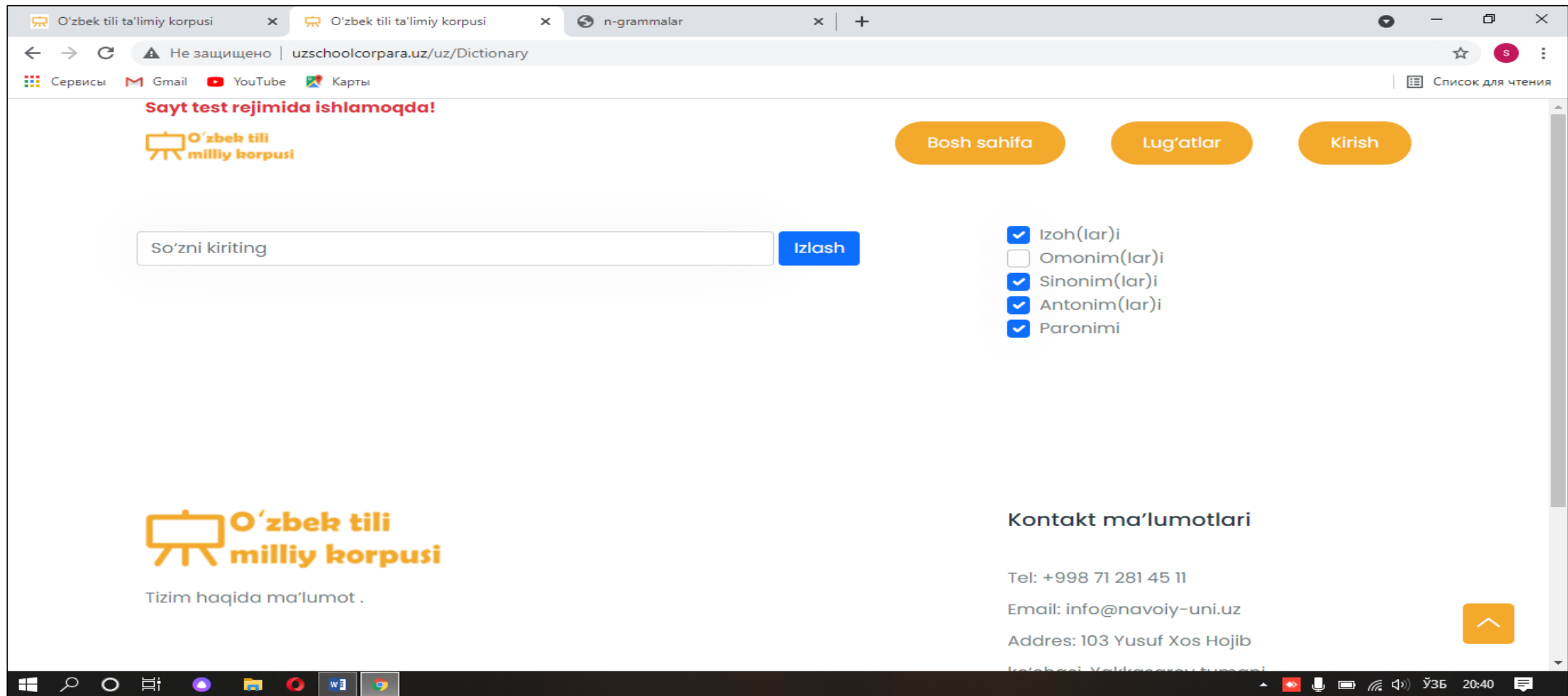
Figure 5. A grammar-based search window interface

The size of the Uzbek language learning corpus is dynamic, as it is in the processing. In the corpus search, only the word and word form search currently returns results (Figure 6).

... chiqqan nihol quyoshga intilgani kabi odamzot ham	kitobga	, bilimga tinimsiz talpinib yashaydi. Bosqinchi...
... Alisher Navoiyning «Xamsa» asarlari va boshqa	kitoblar	buyuk boylik hisoblanadi.
	Kitob	tufayli biz tarixdan, ulug' ajdodlarimizning...
... ba'zan hatto temir sandiqlarda saqlaganlar.	Kitobni	asrash va undan foydalanishning eng qulay yo'li...
... aniq vazifani bajaradi. Turli zallarda joylashgan	kitob	muzeyi, bolalar xonasi, kinomarkaz, bufet, dam...
... hamda turli sohalarga oid ilmiy, ilmiy-ommabop	kitoblar	, ularning bosma va elektron nusxalari, internet...
... brayil yozuvidagi, ya'ni ko'zi ojizlar uchun	kitoblar	, brayil klaviaturasi bilan jihozlangan kompyuterlar...
... ya'ni «Ko'ragoniyning yangi yulduzlar jadvali»	kitobi	nusxalari ham bizda katta qiziqish uyg'otadi. Bu...
<div style="border: 1px solid #ccc; padding: 5px; display: inline-block;"> ← 1 2 3 4 5 6 7 8 9 10 11 12 13 14 ... 119 120 → </div>		

Figure 6. Uzbek language learning corpus wordform search engine result window

The corpus also has an virtual library that allows automatic search. If the Dictionaries option is selected, the window shown in Figure 7 will open.



When you enter a word in the search box, you will find information about the word's description, (if any) synonym, homonym, antonym, spelling.

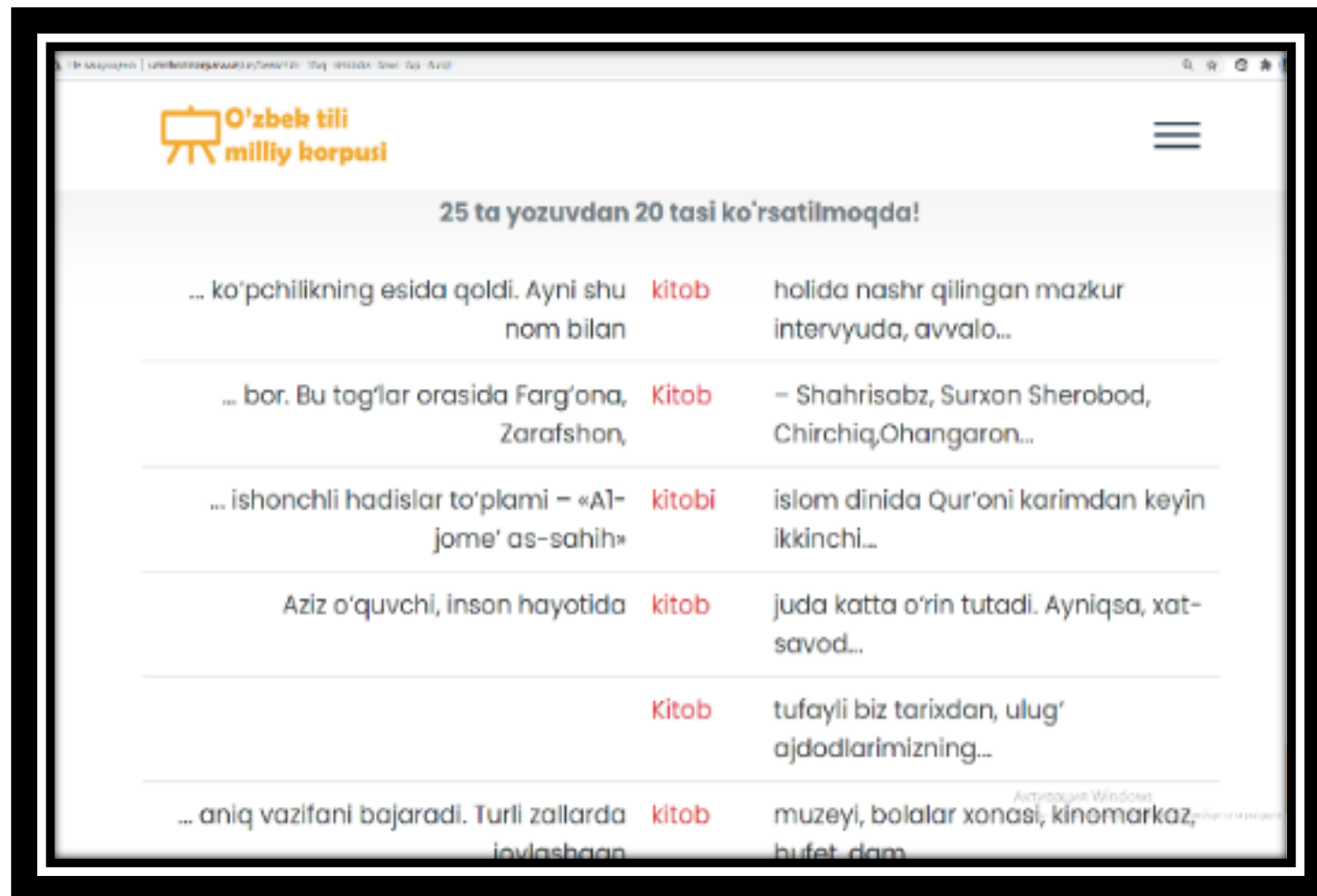
Corpus search allows the user to identify:

1) all forms of a particular word with an array of examples; see from which source the samples were taken;

2) the meaning and interpretation of the word in linguistic dictionaries;

3) a list of words that can be combined with the word given in the search by its right and left;

4) the frequency or statistics of the use of the same word by this or that author;



Conclusion. the learning corps compares the speech of the learner of the person who is its carrier; compares literary oral / written language; compares language styles in books, textbooks, publications, identify differences; serves as a modern information retrieval tool for processes such as determining word order in sentences. By using this tool in linguodidactics, the learner develops the ability to independently analyze language material (based on the corpus), critical and analytical thinking.

Despite its complexity, the learning corps can be an effective resource in the development of teaching material.

However, the site of the project "Uzbek language learning corps", the search engine still needs to be improved. Because when the search engine was tested, errors also occurred as a result of the error of some words. During the subsequent work on the learning corps of the Uzbek language, the search for grammatical, semantic symbols in the corps will be achieved, errors in the search function will be eliminated.

The image shows a screenshot of a search engine interface. A search result is displayed in a white box titled "Batafsil". The text of the result discusses the history of art in the East and Central Asia, mentioning the Middle Ages and the Silk Road. The word "hayot" is highlighted in red in the text. Below the text, there is a blue link: "Tasviriy san'at: Umumiy o'rta ta'lim maktablarining 6-sinfi uchun darslik 2017". Two red arrows point from the word "hayot" and the blue link to two callout boxes. The first callout box contains the text "Qidirilayotgan so'z" (Searched word) in red. The second callout box contains the text "Manba haqida ma'lumot" (Information about the source) in red.

Batafsil

Sharq mamlakatlaridan Xitoy, Hindistonda o'rta asrlar san'ati birmuncha erta boshlanib, XIX asrda ham davom etdi. O'rta Osiyoda O'rta asrlar san'ati VII asrdan boshlanib XVII asrgacha davom etdi. O'rta asrlar san'ati jahon xalqlari milliy madaniyatining gullab-yashnashida asosiy bosqich hisoblanadi. Shu davrdan boshlab milliy o'ziga xos san'at shakllanib, juda ko'p mahalliy maktablar paydo bo'la boshladi. O'rta asrlar san'ati asarlarida **hayot** go'zalligi tarannum etilib, tabiat latofati aks ettirildi.

[Tasviriy san'at: Umumiy o'rta ta'lim maktablarining 6-sinfi uchun darslik 2017](#)

Qidirilayotgan so'z

Manba haqida ma'lumot

1. Abjalova M.A. Linguistic modules of the program of editing and analyzing texts in Uzbek language (for the program of editing texts in official and scientific style): PhD dissertation. – Fergana, 2019. – 164 p.
2. Akhmedova D.B., Mengliev B.R. Semantic Tag Categories in Corpus Linguistics: Experience and Examination International Journal of Recent Technology and Engineering. (IJRTE) ISSN: 2277-3878, Volume-8, Issue-3S, October 2019. – P. 208-212.
3. Akhmetova K.Yu. Corpus approach in teaching a foreign language (Ахметова К.Ю. Корпусный подход в обучении иностранному языку) <https://scipress.ru/philology/articles/korpusnyj-podkhod-v-obuchenii-inostrannomu-yazyku.html>
4. Eshmuminov A.A. Synonym database of the Uzbek National Corps PhD dissertation. – Qarshi, 2019. – 140 p.
5. Karimov R., Mengliev B. Theoretical fundamentals of uzbek-english parallel corpus / Journal of critical reviews. ISSN- 2394-5125. – VOL 7, ISSUE 17, 2020. – P. 73-76.;
1. Karimov R.A., Mengliev B.R. The Role of the Parallel Corpus in Linguistics, the Importance and the Possibilities of Interpretation \ International Journal of Engineering and Advanced Technology (IJEAT). ISSN: 2249 – 8958, Volume-8, Issue-5S3 July 2019. – P. 388-391.
2. Khamroeva Sh. Linguistic bases of creation of the Uzbek language authorship corpus: PhD dissertation. – Bukhara, 2018. – 165 p.
3. Leech G. Teaching and language corpora: A convergence [Text] / G.Leech, A.Wichmann, S.Fligelstone, A.M.Menery. Knowles Teaching and Language Corpora. London: Longman, 1997. – P. 1-23.
4. Mardanshina R.M. National Corpus in the Practice of Linguistic Research and Language Teaching (Марданшина Р.М. Национальный корпус в практике лингвистических исследований и преподавании языка) <https://kpfu.ru/.../Nacioanlnyj.korpus.v.praktike.lingivisticeskikh.issledovaniy.i.prepod>.
5. Zakharov V.P. Corpus linguistics. Study guide. – St. Petersburg, 2005. – 48 с. – С. 34. (Захаров В.П. Корпусная лингвистика. Учебно-методическое пособие)