

KORPUSDAN QIDIRUV TIZIMI SIFATIDA FOYDALANISH

O‘roqova Dilnoza Odil qizi

Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti 2-bosqich magistranti.

E-mail: adilovnадilnoza09@gmail.com

Annotatsiya: Mazkur maqola korpus tushunchasi, uning tildagi o‘rni, korpusning qidiruv turlari, ishslash prinsipi va ularni qidirish, o‘rganishga bo‘lgan turlicha yondashuvlarni o‘z ichiga oladi.

Kalit so‘zlar: korpus, konteks, lug’at, intertekstuallik, yashirin modellar, tillarni taqqoslash, metama’lumot, matn razmetkasi va lingvistik annotatsiya

Abstract: This article includes the concept of the corpus, its place in the language, types of corpus searches, the principle of operation and different approaches to their search and study.

Keywords: corpus, context, vocabulary, intertextuality, implicit models, language comparison, metadata, text markup and linguistic annotation

Аннотация: В данной статье представлены понятие корпуса, его место в языке, виды корпусных поисков, принцип работы и различные подходы к их поиску и изучению.

Ключевые слова: корпус, контекст, словарный запас, интертекстуальность, неявные модели, языковое сравнение, метаданные, текстовая разметка и лингвистическая аннотация.

Dunyo tilshunosligida tilning imkoniyatlarini kengroq o‘rganish, til grammatikasining muammoli tomonlarini kontekstda aniqlash, tilda grammatik qoliplarni belgilash, ko‘p tarmoqli elektron lug’atlar yaratish ishini yengillashtirish, tilni o‘rganishda zamonaviy axborot texnologiyalardan foydalanish samaradorligini oshirish, tilda avtomatik tarjima, qidiruv va kompyuter tahlilini yo‘lga qo‘yish, elektron darsliklar va lug’atlar tayyorlash kabi masalalarini hal qilish uchun tillarda korpuslar yaratishning nazariy va amaliy asoslarini ishlab chiqish, tilning maxsus sohalar bo‘yicha korpusini qurish zaruratining mavjudligi maqolamizning dolzarbligini belgilaydi. Korpus matnlaming katta to‘plami bo‘lishi bilan birga, bu lingvistik tahlilga asoslanadigan yozma yoki og‘zaki materiallar asosi hamdir. Korpusda:

- 1) aniq bir yozuvchining yoki yozuvchilarning matnlari;
- 2) aniq o‘n yil yoki asrga oid matnlar;
- 3) muayyan mavzudagi zamonaviy matnlar;
- 4) til yoki jamiyatda yetarlicha mayjud bo‘lgan zamonaviy matnlardan foydalaniladi.

Yuqorida belgilangan matn turlaridan korpus shakllantirilganda quyidagilami qidirish mumkin:

- 1) kontekstdagi so‘zlamning barcha shakllari;
- 2) lug’atdagi o‘zgarish va izchillik;
- 3) eng ko‘p tanlangan so‘z birikib kela oladigan so‘zlar;
- 4) ikkita matnlar o‘rtasidagi eng muhim farqlar;



- 5) muayyan yozuvchining so‘z va iboralami qo‘llashdagi o‘ziga xoslik;
- 6) intertekstuallik: so‘zning ma’nosi uning qo‘llanilish yig‘indisi sifatida;
- 7) so‘z birliklaridan foydalanadigan yashirin modellar;
- 8) tillarni taqqoslash.

Korpusning qidiruv tizimi va uning ishslash prinsipi 3 turdag'i asosiy qismni o‘z ichiga oladi, bu korpusdagi ma’lumotlar bazasini tekshirishga yordam beradi. Bular: metama’lumot, matn razmetkasi va lingvistik annotatsiya (izohlash). Metama’lumotlar - matn kim tomonidan yozilgani, qachon nashr qilingani, qaysi tilda yozilgani to‘g‘risida ba’zi ma’lumotlami beradi. Metama’lumot korpus matnida kodlanishi yoki alohida hujjat sifatida ma’lumotlar bazasida saqlanishi mumkin. Matn razmetkasi yoki matnni belgilash matnni formatlashni ifodalash uchun ishlatiladi masalan, jumlaning boshlanishi va tugash nuqtasini belgilashda muhim hisoblanadi. Metama’lumot matndagi ma’ruzachilarni aniqlab beradi va har birining yoshi, jinsi haqida foydalii ma’lumot beradi. Matn razmetkasi keyinchalik har bir ma’ruzachi gapirishni boshlaganini va tugatganligini ko‘rsatish uchun ishlatiladi. Korpusga kiritilgan metama’lumot va matn razmetkasi birgalikda bir qator tadqiqot savollariga javob beradi. Matn korpusi ichida lingvistik ma’lumotlami kodlashimiz mumkin, keyinchalik bu tahvilni muntazam va aniq tarzda olib boriladi, bu holatda korpus analitik yoki lingvistik nuqtayi nazar bilan izohlanadi. Annotatsiya kodlash qoidalaridan matn razmetkasi sifatida foydalanadi, masalan, XML (extensible markup language) qirrali-briket teqlari ishlatilishi mumkin, bu yerda otli fraza boshlanadigan nuqtada (<np>) shakli va tugash nuqtasida (</np>) shakli ishlatiladi. Masalan, *np O‘quvchi </np> <np> doskaga </np> chiqdi <np> The children </np> sat in <np> the room </np>*[Abdurakhmonova, N. 2019,2021,2022].

Korpus lingvistikasida korpuslar foydalanuvchiga tez va ishonchli qidirishga imkon beradigan dasturlami qo‘llashda o‘zgarmasdir. Bu dasturlarning ba’zilari, chunonchi, konkordanserlar foydalanuvchiga matndan so‘zlarni qidirishga imkon beradi. Shunga o‘xshash dasturlarning ko‘philigi matnlardagi ma’lumotlar chastotasini yaratishga imkon beradi. Masalan, korpusda paydo bo‘ladigan barcha so‘zlarning ro‘yxati va har bir so‘zning maxsus o‘sha korpusda necha marta uchrashini ko‘rsatadigan so‘zlarning chastotali ro‘yxatini yaratadi. Konkordanser va chastota faktlari tahvilning ikki muhim tomoni hisoblanadi, ya’ni sifat va miqdoriy jihatli sifatida korpus lingvistikasining muhim qismlaridir.

Korpus qidiruv tizimidagi eng muhim saytlardan biri *uzbekcorpus.uz*. O‘zbek tili korpusi uchun lemme va token asosiy qidiruv birligi sanaladi. Lekin korpusning qidiruv menejeri ham token, ham so‘z birikmasi va gap qidiruviga mo‘ljallangan. Lingvistik ma’lumotlar bazasida o‘zbek tilining morfologik bazasi bo‘lib, barcha qidiruv algoritmi uchun asos vazifasini bajaradi. U orqali korpusda mavjud so‘zlarning statistikasini chiqarish imkoniyati tug‘iladi. Shuningdek, korpusning qidiruv tizimi orqali ixtiyoriy so‘zning n-gram bo‘yicha qo‘llanilishi natijasida so‘zlarning konkordanslarni hosil qilish mumkin. Lemma, so‘zning lug‘atdagi shakli bo‘lib, aksariyat tillar uchun yasalish va grammatik jihatdan shakllanish materiali bo‘la oladi. O‘zbek tilining turkiy tillarga xos xususiyatlardan biri morfem birliklarini kombinatsion shakllanishi bo‘yicha natijaga erishildi[Abdurakhmonova, N. 2019,2021,2022].

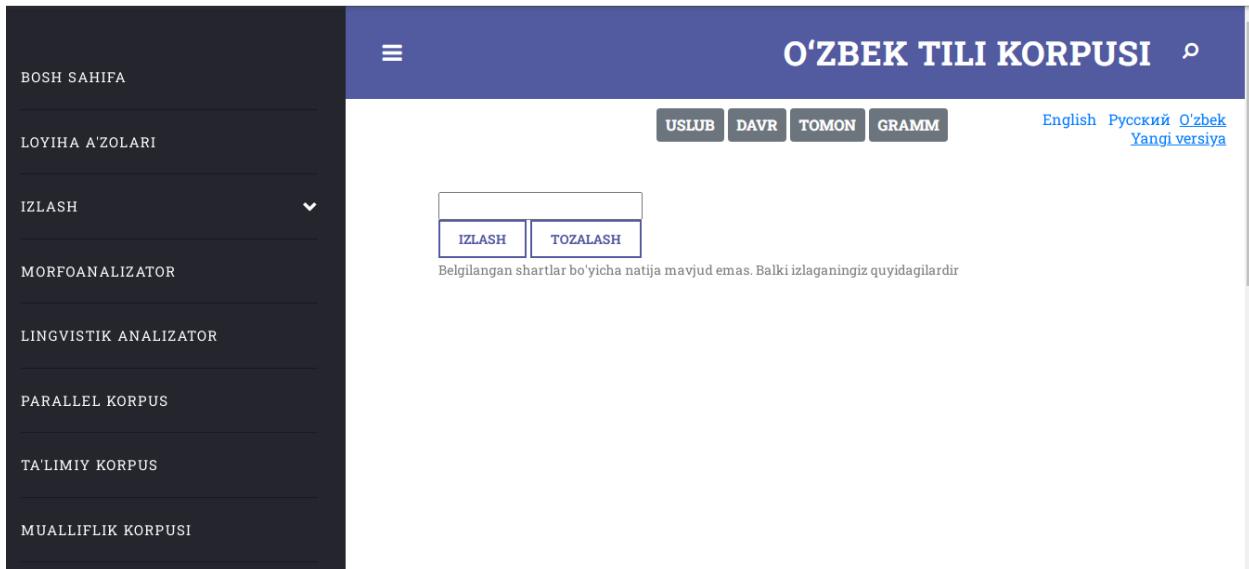
O‘zbek tili elektron korpusning funksiyalari quyidagidan iborat:

- 1) Bosh sahifa
- 2) Izlash (lemma, token, konkordans bo‘yicha)
- 3) Morfoanalizator



- 4) Subkorpuslar (ta'limiylar, parallel korpus)
- 5) Foydalanuvchilar yo'riqnomasi
- 6) Elektron lug'atlar
- 7) Tezaurus

Korpusning qidiruv tizimi uch asosiy birlik bo'yicha funksional vazifa bajaradi



The screenshot shows the homepage of the O'ZBEK TILI KORPUSI. The top navigation bar includes links for English, Russian, and O'zbek versions. The main menu on the left lists various corpus types: BOSH SAHIFA, LOYIHA A'ZOLARI, IZLASH, MORFOANALIZATOR, LINGVISTIK ANALIZATOR, PARALLEL KORPUS, TA'LIMIY KORPUS, and MUALLIFLIK KORPUSI. The right side features a search interface with tabs for IZLASH and TOZALASH, and a note about search conditions.

- 1) Matnlarni uslub bo'yicha qidirish
 - 2) Matnlarni xronologik ya'ni davr tartibida qidirish
 - 3) Konkordanslar uchun n-gram modeli bo'yicha o'ng va chap tarafdan qidirish
- Ushbu electron korpus statistic ma'lumotlarni quyidagi jihatlarga ko'ra chiqarish imkoniyatiga ega:
- 1) lemma bo'yicha qidirish – masalan, kitob so'zi qidirilganda, ushbu so'z o'zak holatda matn tarkibida aniqlanadi



The screenshot shows a search result for the word "kitob". The results are displayed in two columns. The first column contains information about a book by Tohir Malik: "Tohir Malik", "«Shaytanat 1-kitob»", "1989". The second column contains a quote from the book: "TOHIR MALIK SHAYTANAT Birinchi **kitob** MUQADDIMA Bismillohir Rohmanir Rohiyim. «Qachonki (bu) ish tugagach (ya'ni jannat ...ko'proq...»". Below this, another section shows a summary of the book: "Tohir Malik", "«Shaytanat 1-kitob»", "1989", followed by the text: "izlaridan dog' bo'lib ketgan bitta stol, oshiq-ma'shug'i arang ilinib turgan **kitob** javoni turgan tor xona Zohidga meros bo'lib qoldi. Stol tortmasidan ...ko'proq...»".



Lemma bo'yicha qidirishda so'zlarni lug'at shaklida 0 grammatik ko'rsatkichda yoki aynan grammatik qo'shimchalar bilan qidiruvda so'rovga berilgan shakliga qarab natija uzatiladi[Abdurakhmonova, N. 2019,2021,2022].

Lemma korpusdan ma'lumot qidirishning asosiy birligi sanaladi. Morfologik tahlil natijasida mustaqil so'z turkumining statistik natijasi olinadi. Morfologik shakllangan so'zlar kombinatsion jihatdan turli shakllarga ega bo'ladi.

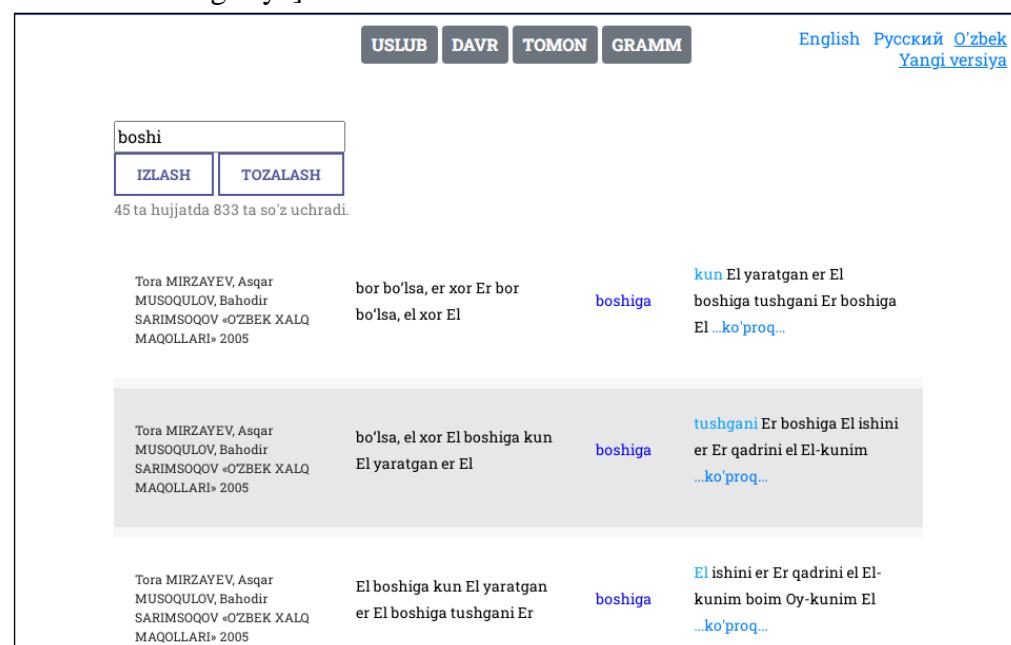
2) token bo'yicha qidirish – [asos + grammatik kategoriya] modeli bo'yicha so'rov yaratish. Mazkur holatda yasovchi qo'shimchalar yoki shu so'z qatnashgan birlik morfologik jihatdan tahlil qilinadi



The screenshot shows the homepage of the O'ZBEK TILI KORPUSI corpus. At the top, there is a navigation bar with three dots on the left and the title 'O'ZBEK TILI KORPUSI' in the center. Below the title are two buttons: 'USLUB' and 'DAVR'. To the right of these buttons are language links: 'English', 'Русский', 'O'zbek', and 'Yangi versiya'. On the left side, there is a sidebar with a box containing the word 'kutubxona' and two buttons: 'IZLASH' and 'TOZALASH'. The main content area displays a list of search results for 'kutubxona': 'Belgilangan shartlar bo'yicha natija mavjud emas. Balki izlaganingiz quyidagilardir'. Below this, a list of words is shown: 'kutubxona', 'kutubxonachi', 'kutubxonalararo', 'kutubxonachilik', 'kutubxonashunos', and 'kutubxonashunoslik'.

(asosan, Grammatik belgilar) bo'yicha toppish imkoniyati mavjud.

3) konkordans bo'yicha qidirish – n-gram left [$W_1+W_2+W_3+Lt$]/ n-gram right [$Lt + W_1 + W_2 + W_3 + W_n$]. Bunda W – so'z, W_1 qidirilayotgan so'zga birish masofadagi n-so'z, Lt – [lemma + Grammatik kategoriya]



The screenshot shows the search results for the word 'boshi'. At the top, there is a navigation bar with four buttons: 'USLUB', 'DAVR', 'TOMON', and 'GRAMM'. To the right of these buttons are language links: 'English', 'Русский', 'O'zbek', and 'Yangi versiya'. On the left side, there is a sidebar with a box containing the word 'boshi' and two buttons: 'IZLASH' and 'TOZALASH'. The main content area displays a list of search results for 'boshi': '45 ta hujjatda 833 ta so'z uchradi.' Below this, three examples of concordance lines are shown:

Tora MIRZAYEV, Asqar MUSQULOV, Bahodir SARIMSOQOV «O'ZBEK XALQ MAQOLLARI» 2005	bor bo'lsa, er xor Er bor bo'lsa, el xor El	boshiga	kun El yaratgan er El boshiga tushgani Er boshiga El ...ko'proq...
Tora MIRZAYEV, Asqar MUSQULOV, Bahodir SARIMSOQOV «O'ZBEK XALQ MAQOLLARI» 2005	bo'lsa, el xor El boshiga kun El yaratgan er El	boshiga	tushgani! Er boshiga El ishini er Er qadrini el El-kunim ...ko'proq...
Tora MIRZAYEV, Asqar MUSQULOV, Bahodir SARIMSOQOV «O'ZBEK XALQ MAQOLLARI» 2005	El boshiga kun El yaratgan er El boshiga tushgani Er	boshiga	El ishini er Er qadrini el El-kunim boim Oy-kunim El ...ko'proq...



Konkordans qidirishda qidiruv tizimidagi so‘zning o‘zidan oldingi yoki keying keluvchi birliklar bilan birikish holati n-gram bo‘yicha ko‘rsatiladi.



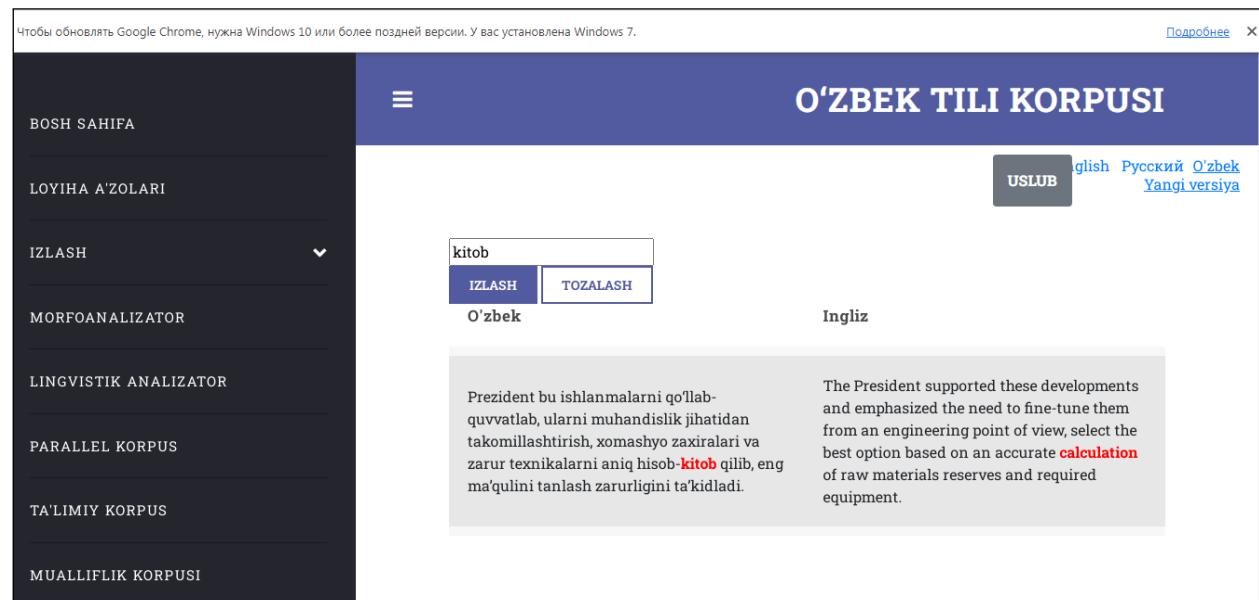
The screenshot shows a software interface for computational linguistics. On the left, a sidebar lists various linguistic tools: IZLASH, MORFOANALIZATOR, LINGVISTIK ANALIZATOR, PARALLEL KORPUS, TA'LIMIY KORPUS, MUALLIFLIK KORPUSI, and LINGVISTIK RESURSLAR. The main window has tabs for kutubxonachi (catalog) and IZLASH. A message at the top says "8 ta hujjatda 17 ta so‘z uchradi". Below this, there are two sections: one for a book by Nilufar Abdurakhmonova and another for a paper by N. Mahmudov et al. To the right, a morphological analysis tool displays the word "kulol" with its lemma "kulol", part of speech "lug'atda", and other details like "So‘z turkumi NOUN" and "Morfologik teg". There are also buttons for "kutubxonachi" and "kushlar" (catalogs).

Matndagi ixtiyoriy so‘zning lemmasi, uning turkumi va FST metodidan olingan morfologik annotatsiyasi oynada o‘z ifodasini topadi

4) so‘z birikmasi bo‘yicha qidirish – [W + W + W_n+ Grammatik kategoriya]. Ushbu qidiruvda turg‘un birikmalarni ajratishda qo‘l keladi. Masalan, o‘rmon qiroli, dala malikasi kabi perifrazalarda gab bo‘lagi vazifasida grammatic qo‘sishchalar bilan kelish imkoniyati hisobga olinadi.

Ushbu tizim ham latin, ham kirill grafikasi bo‘yicha qidirish imkoniyatiga ega.

O‘zbek tili elektron korpusining ichki korpusi parallel korpus bo‘lib, unda o‘zbekcha-inglizcha va inglizcha-o‘zbekcha parallel matnlarning tarjimon xotirada segmentlangan birliklar bo‘yicha toppish imkoniyati mavjud.



The screenshot shows the O'zbek Tili Korpusi website. The sidebar includes links for BOSH SAHIFA, LOYIHA A'ZOLARI, IZLASH, MORFOANALIZATOR, LINGVISTIK ANALIZATOR, PARALLEL KORPUS, TA'LIMIY KORPUS, and MUALLIFLIK KORPUSI. The main content area features a search bar with tabs for kitob (book), IZLASH, TOZALASH, and O'zbek. It also includes language selection buttons for USLUB (Uzbek), English, Russian, and O'zbek Yangi versiya. A detailed view of a text snippet from a document by Prezident bu ishlanmalarni qo'llab-quvvatlab, ularni muhandislik jihatidan takomillashtirish, xomashyo zaxiralari va zarur texnikalarni aniq hisob-kitob qilib, eng ma'qulini tanlash zarurligini ta'kidladi. The English translation below states: "The President supported these developments and emphasized the need to fine-tune them from an engineering point of view, select the best option based on an accurate calculation of raw materials reserves and required equipment."

Korpus qidiruv tizimi sifatida so‘zning bexato ekanligini tasdiqlash yoki uning xatoligini unga yaqin so‘zlar bazasidan qidirish korpus texnologiyisi uchun muhim sanaladi[Abdurakhmonova, N. 2019,2021,2022].

Korpus lingvistikasining rivojlanish evolyutsiyasi shuni ko‘rsatadiki, ushbu fan kompyuter lingvistikasi bilan uzviy bog‘liq. Chunki u yoki bunda yaratilgan ilmiy yutuqlar bir-birini



to‘ldirishga, imkoniyatlardan foydalanishning optimal usullaridan foydalanishga omil bo‘lib xizmat qiladi. Korpuslar uchun berilgan ta’riflar ichida Webster lug‘atida tilga oid ta’rifida “bilimlar yoki dalillar to‘plami; tilni tavsiflovchi tahlil uchun foydalanilgan so‘zlar to‘plami” izohi berilgan, boshqa manbalarda korpus (ko‘pligi - corpora) yozma matnlar yoki yozib olingan og‘zaki nutqning transkriptsiyasidan tuzilgan lingvistik ma’lumotlar to‘plami bo‘lib, asosiy maqsadi tilda mavjud farazlarni tasdiqlash deya, ta’riflanadi. Korpus haqidagi ta’riflar va farazlar biroz farq qilsa-da, ammo, barcha fikrlar birlashib, korpus til birliklarining xususiyatlarini aniqlash maqsadida qidiruv dasturiga bo‘ysundirilgan matnlar majmui, tabiiy tildagi elektron shaklda saqlanadigan yozma yoki og‘zaki, kompyuterlashtirilgan qidiruv tizimiga dasturiy ta’midot asosida joylashtirilgan matnlar jamlanmasi ta’rifi shakllantiradi.

Foydalanilgan adabiyotlar ro‘yxati:

1. Мирзиёев Ш. Миллий тараққиёт йўлимизни қатъият билан давом эттириб, янги босқичга кўтарамиз. – Тошкент: Ўзбекистон, 2017. – Б.168.
2. Abdurahmonova N. O‘zbek tili elektron korpusining kompyuter modellari. Monografiya. – Toshkent, 2021.
3. Abdurakhmonova, N., Tulihev, U., Ismailov, A., & Abduvahobo, G. (2022). Uzbek electronic corpus as a tool for linguistic analysis. In Компьютерная обработка тюркских языков. TURKLANG 2022 (pp. 231-240).
4. Abduraxmonova, N. Z. Q., & Urazaliyeva, M. Y. (2022). O ‘zbek tili elektron korpusida (<http://uzbekcorpus.uz/>) og ‘zaki matnlar korpusini yaratishning nazariy va amaliy masalalari. Academic research in educational sciences, 3(3), 644-650.
5. Mengliev, D., Barakhnin, V., & Abdurakhmonova, N. (2021). Development of intellectual web system for morph analyzing of uzbek words. Applied Sciences, 11(19), 9117.
6. Abdurakhmonova, N. (2019). Dependency parsing based on Uzbek Corpus. In of the International Conference on Language Technologies for All (LT4All).
7. Agostini, A., Usmanov, T., Khamdamov, U., Abdurakhmonova, N., & Mamasaidov, M. (2021, January). Uzwordnet: A lexical-semantic database for the uzbek language. In Proceedings of the 11th Global Wordnet conference (pp. 8-19).
8. <https://www.merriam-webster.com/dictionary/corpus>
9. Crystal D. An Encyclopedic Dictionary of Language and Languages. - Oxford, 1992.