

MILLIY KORPUS VA TIL KORPUSLARINI YARATISH MASALALARI

O‘ZBEK TILI KORPUSINING DASTURIY TA’MINOTI INTERFEYSI VA QIDIRUV TIZIMIDAN FOYDALANISH

Qarshiyev Abduvali Berkinovich

karshiyev@umail.uz

Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti
Samarqand filiali professori
Fizika-matematika fanlari nomzodi

Karimov Suyun Amirovich

suyun1950@mail.ru

Samarqand davlat universiteti professori
Filologiya fanlari doktori

Tursunov Muhammadsolih Sa’din o’g’li

muhammadsolih927@gmail.com

Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti
Samarqand filiali assistenti

Xolmuxamedov Baxtiyor Farxodovich

bxolmuxamedov@mail.ru

Samarqand davlat universiteti katta o’qituvchisi
Filologiya fanlari bo’yicha falsafa doktori

Annotasiya: So’ngi vaqtlarda tillarni o’qitish, til ustida turli xil statistik tahlillar olish va tadqiq qilishda ma’lumotlarni avtomatlashitirish, katta hajmdagi ma’lumotlar bilan ishlashga doir dasturiy ta’minotlar yaratilmoqda. Shunday vaziyatda til korpuslarini yaratish juda muhim hisoblanadi.

Ushbu maqolada o’zbek tilik korpusiga mo’ljallangan uzbekcorpora.uz va so’z va iboralarni korpus bo’ylab qidirish (konkordans) tizimi tavsiflangan. uzbekcorpora.uz dasturiy ta’minoti online rejimda ishlaydigan bepul platforma hisoblanadi. Bu esa tilni tadqiq qilishda o’rganuvchiga istalgan joyda, istalgan kompyuterda ishlashiga imkon beradi.

Annotation: Recently, the software has been developed to teach languages, automate data in obtaining and researching various statistical analyzes on languages, and working with large amounts of data. In such a situation, the creation of language corps is very important.

This article describes the uzbekcorpora.uz corpus system for words and phrases for the Uzbek language corpus. uzbekcorpora.uz software is a free online platform. This allows the learner to work anywhere, on any computer, while studying a language.

Аннотация: Недавно было разработано программное обеспечение для обучения языкам, автоматизации данных при получении и исследовании различных статистических анализов языков и работы с большими объемами данных. В такой ситуации очень важно создание языкового корпуса.

В данной статье описывается система корпусов uzbekcorpora.uz слов и словосочетаний для корпуса узбекского языка. Программа uzbekcorpora.uz - это бесплатная онлайн-платформа. Это позволяет учащемуся работать где угодно, на любом компьютере, изучая язык.

Kalit so'zlar: uzbekcorpora.uz, o'zbek tili, korpus, konkordans, qidiruv tizimi.

Tizimning asosiy oynasi.

O'zbek tili milliy korpusi bilan ishlash uchun brauzer yordamida <http://uzbekcorpora.uz/> saytiga kiriladi. Saytga kirgan vaqtda 1-rasmdagi interfeys hosil bo'ladi. Bu interfeysning chap qismi menyular panelidan iborat bo'lib, korpusning umumiy ma'lumotlari, loyiha to'g'risida ma'lumotlar va dasturiy ta'minotlar keltirilgan[4].

1-rasm. Asosiy interfeys

O'ZBEK TILI KORPUSI	
Asosiy	Korpus nima?
Konkordans	Korpus - muayyan tilda matnlar to'plamiga asoslangan elektron shakldagi axborot-ma'lumot tizimidir. Milliy korpus bu tili o'zining mavjudligining muayyan bosqichida (yoki bosqichida) turli xil janrlari, uslubi, hududiy va ijtimoiy variantlari bilan birgalikda boshqa ko'rinishlarda ifodalaydi.
Tokenayzer	Milliy korpus ilmiy tadqiqot va tillarni o'qitish uchun tilshunolar (korpus tilshunolari, jadal rivojlanayotgan zamonaviy tilshunoslik sohasi mutaxassislari) tomonidan yaratiladi. Dunyoning eng muhim tillarining aksariyatida o'zlarining milliy korpuslari (to'liqligi va ilmiy so'zlarni qayta ishlash darajasidan farqli) mavjud. Britaniya Milliy korpusi (BMK) umumiy namuna sifatida olinadi: boshqa ko'plab zamonaviy korpuslar unga yo'naltirilgan. Pragadagi Karlov universiteti tomonidan yaratilgan Chexiya milliy korpusi slavyan tillari korpusi orasida alohida o'rin tutadi.
Lemmatayzer	Ikkinchidan, korpus o'zining tarkibidagi matnlar (xususan, razmetka yoki annotatsiya deb ataladigan) xususiyati haqida qo'shimcha ma'lumot beradi. Razmetka - Korpusning asosiy xarakteristikasi, korpusni zamonaviy intermedda, shu jumladan rus tilda keng namoyish etilgan matnlarning oddiy to'plamlaridan (yoki "kutubxonalaridan") ajratib turadi (Masalan, Maksim Moshkovning eng mashhur "kutubxonasi" yoki "Rossiya Virtual kutubxonasi" kabi). Bugungi kunda mutaxassislar rus klassik adabiyotining "Asosiy elektron kutubxonasi" ni tuzishdi va ular matnlarni taqdim etishning ilmiy rejasiga, nufuzli bosma nashrlarning eng mukammal reproduksiyasiga asos soldi. Biroq, bunday kutubxonalar ilmiy tadqiqotlar uchun ishlov berilmagan tilda juda cheklangan. Bundan tashqari, kutubxonalar matnlarning mazmuniga qiziqish bildirgan kishilarning til qobiliyatidan ko'ra ko'proq yaratilganligini unutmamak kerak. Milliy korpus tuzuvchilari uchun kitobning qiziqiligi yoki foydasi, uning yuksak badiiy yoki ilmiy jihatdan muhimligi kabi omillar ahamiyatga ega emas. Milliy korpus elektron kutubxonadan farqli o'laroq - bu "qiziq" yoki "foydal" matnlar to'plami emas. Bu matnlar to'plami tili o'rganishda qiziq va foydali bo'lishi mumkin. Kichik yozuvchining romani, odatdagi telefon subhatini yozish, odatiy ijara shartnomasi va h.k. bo'lishi mumkin. Albatta, klassik san'at asarlari adabiyotlari bilan birgalikda.
Razmetkalash	Korpus ilmiy va o'quv qiymati bilan baholangan va razmetka qilingan. Hozirda Rossiya milliy korpusida besh turdagi razmetkadan qo'llanilgan: metateks, morfologik (so'z o'zgaruvchan), sintaktik, aksent va semantik. Yaqin kelajakda, asosiy korpusda ta'limning razmetkasi va soddalashtirilgan sintaktik yozuvni joriy qilish rejalashtirilgan (sintaktik. Muloqaza izohli to'plami taqdim etilganidan boshqacha ko'rinishdan farq qiladi). Razmetkalash tizimi doimiy ravishda takomillashtirilmoqda.
Chastotali lug'at	Nima uchun milliy korpusga muhtojmiz? Milliy korpus, birinchi navbatda, tilning lug'aviy va grammatikasiga oid ilmiy tadqiqotlar uchun mo'ljallangan. Bir ikki asrda tilning o'zgarish jarayoni kichik qiymatda amalga oshadi. Korpusning yana bir vazifasi - ko'rsatilgan joylarga tegishli barcha turdagi murojaatlarni taqdim etishdir (leksika, grammatika, aksentologiya, til tarixi). Zamonaviy kompyuter texnologiyalari katta hajmli matnlarni lingvistik ishlav berish jarayonlarini bir necha bor soddalashtiradi va tezlatadi.
Korpus nima?	Ilgari, tadqiqotchi faqatgina matnlarni ko'rib chiqishi va kerakli tahlillarni qo'lda tayyorlashi mumkin edi; bu dastlabki (ammo mutlaqo muqarrar) faoliyat juda qiyin edi va katta hajmdagi matnlarni qayta ishlashda qiyinchilik yuzaga keldi. Hozirda tahlil qilinayotgan material hajmi va axborot olish tezligi cheklavlar mavjud emas, ya'ni tadqiqotchilarning turli xil hajmdagi matnlari mavjud. Bu til haqidagi bilimimizni rivojlantirishni sekinlashtirmaydi: ilgari mavjud bo'lmagan statistika - so'zlarni qayta ishlash, shu jumladan, tilning tuzilishi va rivojlanishini, tarkibini, tili ilmiy yechimlarini to'g'ri taqin qilinishi yoki shubha qoldirmasligi aniqlaydi ammo qat'iy asoslanmagan. Hozirgi kunda tillarning grammatik strukturasi ilmiy ta'riflari, shuningdek, nufuzli akademik lug'atlar - deyarli barchasi istisnosiz - bu tillarning korpuslari asosida tuzilishi kerak. Korpus ma'lumotlarini hisobga olish juda ko'p ixtisoslashgan ilmiy tadqiqotlar uchun juda zarur (majburiy emas).
Tarkib va tuzilish	Milliy korpuslarning asosiy iste'molchilari, albatta, juda ko'p turi yo'nalishdagi tilshunos tadqiqotchilardir. Lekin, korpus foydalanuvchilarining diapazoni professional tilshunolar bilan chegaralanib qolmaydi. Muayyan davr yoki muayyan muallifning tiliga oid ishonchli statistikalar adabiyotshunolar, tarixchilar va boshqa ko'plab soha vakillari uchun qiziqarli bo'lishi mumkin. Milliy korpus, ona tili yoki chet til sifatida o'qitish uchun muhimdir. Chet ellik, maktab o'qituvchisi, o'qituvchi, jurnalist, muharrir va yozuvchi tanqit mualliflar orasida noma'lum so'z yoki grammatik shaklni qo'lash xususiyatlarini tez va samarali tekshirib ko'rishlarida yordam beradi. Shunday qilib, milliy korpus kasb bog'liq hamma uchun, zarurat yoki oddiy qiziqish bilan, tilning tuzilishi va faoliyati, xususan, ushbu tiling murakkab spikerlari va uni chet til sifatida o'qiganlarning barchasiga til haqidagi savollarga javob berishga harakat qiladi.
Korpus statistikasi	
Morfologiya	
Loyiha haqida	
Loyiha ishtirokchilari	
Ishlanmalar	
Korpusdan foydalanish	
Boshqa korpuslar	

Asosiy – ushbu menyuda o'zbek tili milliy korpusi to'g'risidagi ma'lumotlar keltirilgan;

Konkordans – ushbu menyu orqali korpusning qidiruv tizimiga o'tiladi. Bu tizim konkordans nomli paragrafda to'liq tavsiflangan;

Tokenayzer – ushbu menyu orqali matnlarni tokenlash dasturiga o'tiladi. Bu dastur tokenayzer nomli paragrafda to'liq tavsiflangan;

Lemmatayzer – ushbu menyu orqali matnlarni lemmalash dasturiga o'tiladi. Bu dastur lemmatayzer nomli paragrafda to'liq tavsiflangan;

Razmetkalash – ushbu menyu orqali so'zlarning razmetkasini ko'rish mumkin bo'lgan dasturga o'tiladi. Bu dastur razmetkalash nomli paragrafda to'liq tavsiflangan;

Chastotali lug'at – ushbu menyu orqali matnlarni statistik tahlil qilish dasturiga o'tiladi. Bu dastur chastotali lug'at nomli paragrafda to'liq tavsiflangan;

Korpus nima? – ushbu menyuda til korpusiga ta'riflar, korpus to'g'risidagi ma'lumotlar keltirilgan;

Tarkib va tuzilish – ushbu menyuda o'zbek tili milliy korpusining tuzilish strukturasi keltirilgan;

Korpus statistikasi – ushbu menyuda til korpusi bazasiga kiritilgan matnlar soni, gaplar soni, so'zlar soni to'g'risidagi ma'lumotlar keltirilgan;

Morfologiya – ushbu menyuda til korpusida morfologik razmetkalash va morfologik razmetkalashning standartlari to'g'risidagi ma'lumotlar keltirilgan;

Loyiha haqida – ushbu menyuda o'zbek tili milliy korpusini yaratish

to'g'risidagi ma'lumotlar keltirilgan;

Loyiha ishtirokchilari – ushbu menyuda o'zbek tili milliy korpusini yaratishda ishtirok etayotgan xodimlar to'g'risidagi ma'lumotlar keltirilgan;

Ishlanmalar – ushbu menyuda o'zbek tili milliy korpusini yaratish jarayonida chop etilgan maqolalar, dasturiy guvohnomalar to'g'risidagi ma'lumotlar keltirilgan;

Korpusdan foydalanish – ushbu menyuda o'zbek tili milliy korpusidan foydalanish to'g'risidagi ma'lumotlar keltirilgan;

Boshqa korpuslar – ushbu menyuda boshqa davlatlarning milliy korpuslari to'g'risidagi ma'lumotlar keltirilgan [1];

Konkordans

Boshqa korpuslar singari uzbekcorpora.uz tizimining asosiy vazifalaridan biri bo'lgan korpus bo'ylab qidiruv tizimi amalga oshiriladi. Bunda qidirilayotgan so'zning xususiyatlari to'g'risida ma'lumotlar beradi. Asosiy oynaning menyular qatoridan (1-rasmga qarang) konkordans tugmasi bosilgan vaqtda 2-rasmdagi interfeys ochiladi. Foydalanuvchilarga tizimdan foydalanishlariga qulay bo'lishi uchun sodda interfeys ishlab chiqilgan va hamma funsiyalari bitta oynada aks ettirilgan[3].

2-расм. Конкорданс интерфейси



2-rasmda ko‘rinib turgan konkordanser vositasi bir qancha funksiyalarga ega bo‘lib, ular yordamida uzbekcorpora.uz o‘zbek tilini o‘rganuvchilar va tadqiqotchilar uchun samarali vositaga aylanadi.

Bu xususiyatlarni quyidagicha jamlab aytish mumkin:

1. *Admin* – bu tugma yordamida tizim administratori yoki administrator ruxsat bergan foydalanuvchilar ma’lumotlar bazasiga kirishi mumkin. Bunda ma’lumotlar bazasini boshqarish (ma’lumotlarni qo‘shish, tahrirlash, o‘chirish) imkoniyatlari mavjud.

2. *Qidirilayotgan so‘z* – bu maydon korpus qidiruv tizimidan foydalanishdagi dastlabki vazifa bo‘lib, bu maydonga o‘rganilayotgan so‘z kiritiladi va “natijani ko‘rish” tugmasi bosiladi. “o‘chirish” tugmasi esa maydonni qidirilayotgan so‘zdan tozalaydi.

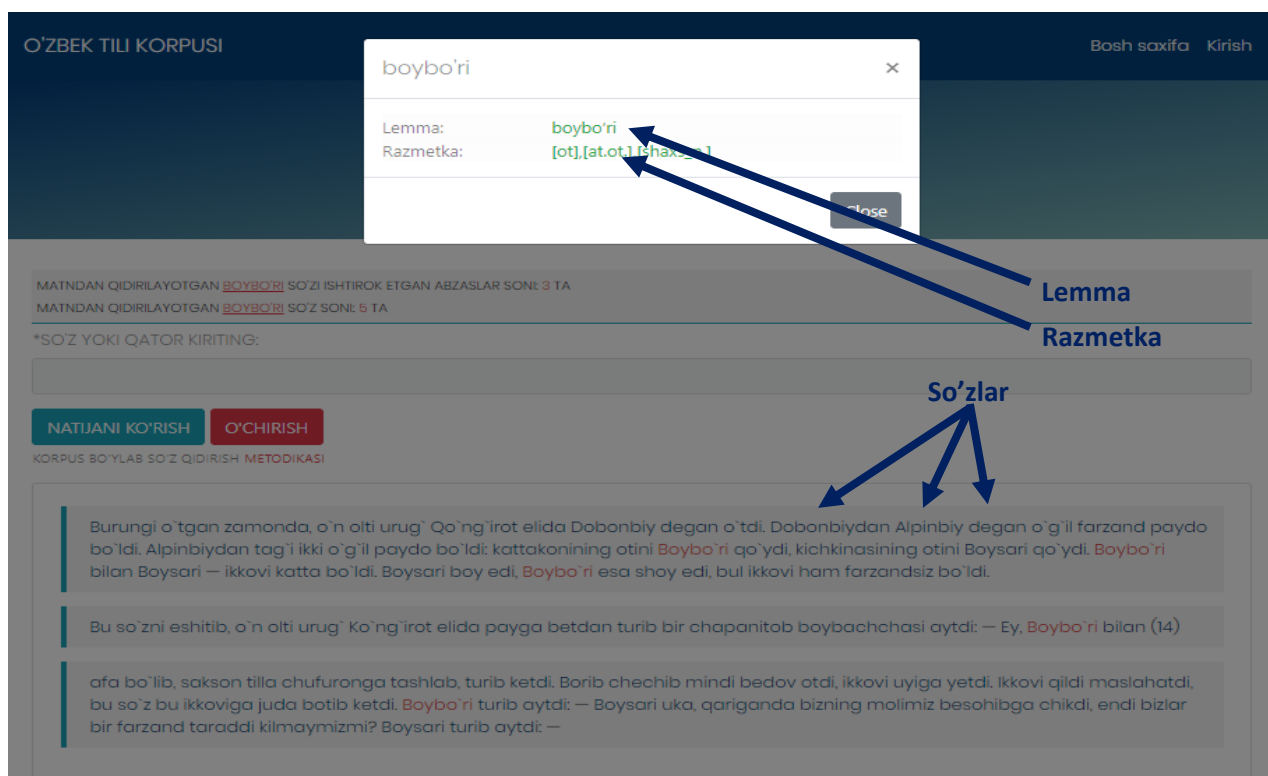
3. *Abzaslar* – ma’lumot kiritish maydoniga so‘z kiritilib, “natijani ko‘rish” tugmasi bosilgan vaqtda natija sifatida o‘rganilayotgan so‘z ishtirok etgan abzaslar ko‘rinadi. Har bir abzas bir-biri bilan ajratilib ko‘rsatilgan.

4. *So‘zlar* – Natija sifatida topilgan abzaslardagi so‘zlar ustida sichqoncha yurgizilganda, ya’ni sichqoncha qaysi so‘zning ustida bo‘lsa shu so‘zning tagida chiziqliq paydo bo‘ladi. Bunda so‘zning ustiga sichqoncha kelgan vaqtda so‘z teglanadi va so‘z ustidan sichqoncha olib kelib bosilsa 3-rasmda ko‘rsatilganidek so‘zning xususiyatlari (so‘z lemmasi va razmetkasi) hosil bo‘ladi. O‘rganilayotgan so‘z esa boshqa so‘zlardan farqli ravishda boshqa rangda ko‘rinadi.

5. *Statistika* – o‘rganilayotgan so‘z maydonga kiritilib, “natijani ko‘rish” tugmasi bosilgan vaqtda tahlil maydonida ikkita statistik tahlil paydo bo‘ladi:

- Qidiruv natijasi sifatida topilgan, o‘rganilayotgan so‘z ishtirok etgan abzaslar soni;
- O‘rganilayotgan so‘z, ma’lumotlar bazasida mavjud bo‘lgan matnlarda ishtirok etgan miqdori[1].

3-rasm. So‘zning lemmasi va razmetkasi.



Dastur ish natijasi sifatida 2-rasmda ko'rsatilgan. So'z kiritish maydoniga “boybo'ri” so'zi kiritilgan. “natijani ko'rish” tugmasi bosilganda ma'lumotlar bazasidan o'rganilayotgan so'z (“boybo'ri”) ishtirok etgan abzaslar natija sifatida hosil bo'lgan. “boybo'ri” so'zining ustiga sichqonchani olib kelib so'z ustidan bosilganda 3-rasmdagi ma'lumot hosil bo'ladi. 3-rasmda “boybo'ri” so'zining lemmasi va razmetkasi hosil bo'lgan [2].

1-жадвал

Qiririlayotgan so'z	So'zning lemmasi	So'zning razmetkasi
boybo'ri	Boybo'ri	[ot],[at.ot],[shaxs_n.]

O'rganilayotgan so'z (“boybo'ri”)ning razmetkasi 3ta tegdan iborat bo'lib ular quyidagicha nomlanadi:

[ot] – bu teg o'rganilayotgan so'zning ot so'z turkumi tarkibiga kiradi;

[at.ot] – bu teg o'rganilayotgan so'zning ot so'z turkumidagi atoqli otlar guruhiga mansubligini bildiradi;

[shaxs_n.] – bu teg o'rganilayotgan so'zning ot so'z turkumidagi atoqli otlar guruhida mavjud bo'lgan belgilardan biri shaxs nomi ekanligi ko'rsatilgan[2].



Foydalanilgan adabiyotlar:

1. Tursunov M.S., Qarshiyev A.B., Karimov S.A., Development of a modern corpus of computational linguistics,
<https://www.scopus.com/record/display.uri?eid=2-s2.0-85102147010&origin=AuthorNamesList&txGid=13070f5e69ea0839e04674499810efe1>, DOI: 10.1109/ICISCT50599.2020.9351376
2. Баранов А.Н., Корпусная лингвистика // Баранов А. Н. Введение в прикладную лингвистику. М., 2001. С. 112–137.
3. Гвишиани Н.Б., Практикум по корпусной лингвистике: Учеб. пособие по английскому языку. М.: 2008.
4. McEnery T., Hardie A. Corpus linguistics: method, theory and practice. Cambridge University Press, 2012.