



The IX International Conference on Computer Processing of Turkic Languages “TurkLang 2021”

UZBEKCORPORA.UZ: UZBEK LANGUAGE CORPUS SOFTWARE



Lecturer: Abduvali KARSHIEV, Mukhammadsolikh TURSUNOV, Bakhtiyor KHOLMUKHAMEDOV (Uzbekistan)

Abstract. This article describes the system uzbekcorpora.uz for the Uzbek language national corpus. The software consists of components such as body search for words and phrases (concordance), markup (word characteristics), lemmas, tokens, and frequency dictionaries. There is also an admin section for editing the corpus database. Uzbekcorpora.uz software is a free online platform. This allows the learner to work anywhere, on any computer, while researching a language.

Tokenizer

Lemmatizer

Marking

Concordance

Corpus control system

Asosiy	<h2>Korpus nima?</h2> <p>Korpus - muayyan tilda matnlar to'plamiga asoslangan elektron shakldagi axborot-ma'lumot tizimidir. Milliy korpus bu tilni o'zining mavjudligining muayyan bosqichida (yoki bosqichida) turli xil janrlari, uslubi, hududiy va ijtimoiy variantlari bilan birgalikda boshqa ko'rinishlarda ifodalaydi.</p> <p>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.</p> <p>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 internetda, shu jumladan rus tilida 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 reproduksiya 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 unutmash 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 "foydali" matnlar to'plami emas; Bu matnlar to'plami tilni o'rganishda qiziq va foydali bo'lishi mumkin. Kichik yozuvchining romani, odatdagi telefon suhbatini yozish, odatiy ijara shartnomasi va h.k. bo'lishi mumkin. Albatta, klassik san'at asarlari adabiyotlari bilan birgalikda.</p> <p>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 Mulohaza izohli to'plami taqdim etilganidan boshqacha ko'rinishdan farq qiladi). Razmetkalash tizimi doimiy ravishda takomillashtirilmoqda.</p> <h3>Nima uchun milliy korpusga muhtojmiz?</h3> <p>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 tarxi). Zamonaviy kompyuter texnologiyalari katta hajmli matnlarni lingvistik ishlav berish jarayonlarini bir necha bor soddalashtiradi va tezlatadi.</p> <p>Ilgari, tadqiqotchilari 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 kelardi. Hozirda tahlil qilinayotgan material hajmi va axborot olish tezligi cheklavl mavjud emas, ya'ni tadqiqotchilarning turli xil hajmdagi matnlari mavjud. Bu til haqidagi bilimimizni rivojlanishini sekinlashtirmaydi: ilgari mavjud bo'lmagan statistika - so'zlarni qayta ishlash, shu jumladan, tilning tuzilishi va rivojlanishini, tarkibini, tilni ilmiy yechimlarini to'g'ri talqin qilinishi yoki shubha qoldirmasligi aniqlaydi ammo qat'i 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).</p> <p>Milliy korpuslarning asosiy iste'molchilari, albatta, juda ko'p turli yo'nalishdagi tilshunos tadqiqotchilardir. Lekin, korpus foydalanuvchilarining diapazoni professional tilshunoslar 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 tilni yoki chet til sifatida o'rgatish uchun muhimdir. Chet ellik, maktab o'qituvchisi, o'qituvchi, jurnalist, muharrir va yozuvchi taniqli mualliflar orasida noma'lum so'z yoki grammatik shaklni qo'llash 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 tilning murakkab spikerlari va uni chet til sifatida o'qiganlarning barchasiga til haqidagi savollarga javob berishga harakat qiladi.</p>
Konkordans	
Tokenayzer	
Lemmatayzer	
Razmetkalash	
Chastotali lug'at	
Korpus nima?	
Tarkib va tuzilish	
Korpus statistikasi	
Marfologiya	
Loyiha haqida	
Loyiha ishtirokchilari	
Ishlanmalar	
Korpusdan foydalanish	
Boshqa korpuslar	

Fig. 1. Main interface

O`ZBEK TILI KORPUSI

Statistics Tokenlar soni: 163

Asosiy
Burungi o'tgan zamonda, o'n olti urug' Qo'ng'iro't elida Dobonbiy degan o'tdi. Dobonbiydan Alpinbiy degan o'g'il farzand paydo bo'ldi. Alpinbiydan tag'i ikki o'g'il paydo bo'ldi: kattakonining otini Boybo'ri qo'ydi, kichkinasining otini Boysari qo'ydi. Boybo'ri bilan Boysari — ikkovi katta bo'ldi. Boysari boy edi, Boybo'ri esa shoy edi, bu ikkovi ham farzandsiz bo'ldi.

Konkordans

Tokenayzer (highlighted)

Lemmaayzer

Razmetkalash

Chastotali lug'at

Korpus nima?

Tarkib va tuzilish

Korpus statistikasi

Morfologiya

Loyiha haqida

Loyiha ishtirokchilari

Ishlanmalar

Korpusdan foydalanish

Boshqa korpuslar

Text

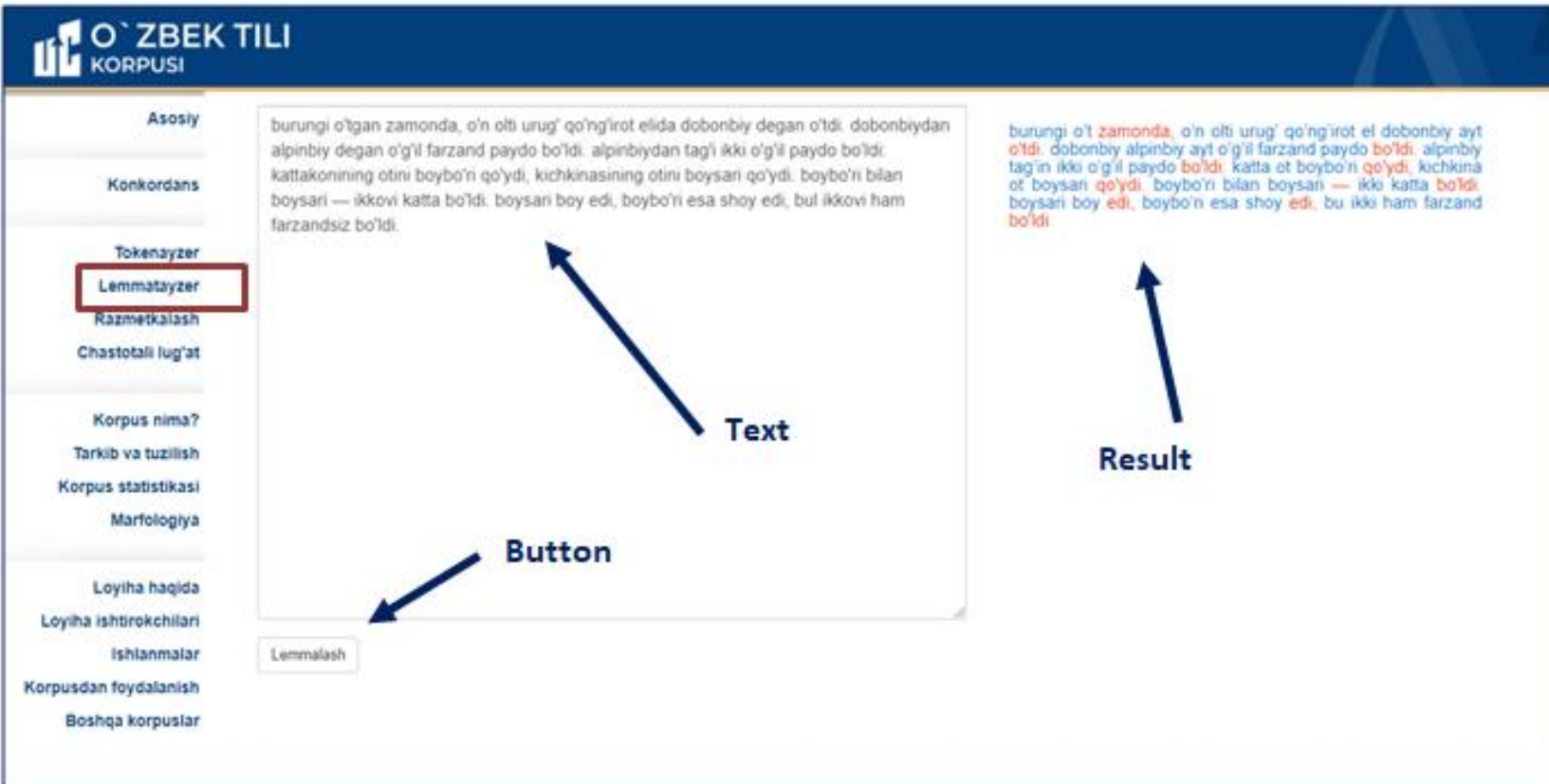
Ana endi o'n olti urug' Qo'ng'iro't elida bir chufuron to'y bo'ldi. Xaloyiqlarni, elatyalarni to'yga xabar qildi. Shu to'yga barcha xaloyiqlar yig'ildi. Biylar ham to'yga keldi. To'ydagi kattalar ilgariqiday izzat qilib, qadimgiday otini ushlamadi. Biylar: "Mazmuni bu odamlar bizning kelganimizdan bexabar qoldi", — deb otini o'zi boyfab, ma'raka-maylsga kelib o'tira berdi. Biylarning ko'nglini xushlamadi, otini ushlamadi, ostiga libos tashlamadi; osh tortdi, so'zgan tovoqni choshlamadi; osh tortganda, oshning ketini-ko'tini tortdi. Bu qilgan xizmatni biylar ko'rib, ilgari izzat ko'rib yurgan odamlar — biylar aytdi: — Bizlar o'n olti urug' Qo'ng'iro'tning boyi ham shoyi bo'lsak, bizlar kelsak, otimizni ushlar edinglar, ko'nglimizni xushlar edinglar, ostimizga libosni tashlar edinglar, bu daf'a bizdan nima ko'tohlik o'tdi, bizni bunday behurmat qildinglar.

Button

Tokens

Burungi
o'tgan
zamonda
o'n
olti
urug'
Qo'ng'iro't
elida
Dobonbiy
degan
o'tdi.
Dobonbiydan
Alpinbiy
degan
o'g'il
farzand
paydo
bo'ldi.

Fig. 2. Tokenizer interface



O`ZBEK TILI KORPUSI

Asosiy

Konkordans

Tokenayzer

Lemmatayzer

Razmetkalash

Chastotali lug'at

Korpus nima?

Tarkib va tuzilish

Korpus statistikasi

Marfologiya

Loyiha haqida

Loyiha ishtirokchilari

ishlanmalar

Korpusdan foydalanish

Boshqa korpuslar

burungi o'lgan zamonda, o'n olti urug' qo'ng'iroq elida dobonbiy degan o'ldi. dobonbiydan alpinbiy degan o'g'il farzand paydo bo'ldi. alpinbiydan tag'i ikki o'g'il paydo bo'ldi. kattakonining otini boybo'ni qo'ydi, kichkinasining otini boysari qo'ydi. boybo'ni bilan boysari — ikkovi katta bo'ldi. boysari boy edi, boybo'ni esa shoy edi, bul ikkovi ham farzandsiz bo'ldi.

Text

Button

Lemmalash

burungi o'l zamonda, o'n olti urug' qo'ng'iroq el dobonbiy ayt o'ldi. dobonbiy alpinbiy ayt o'g'il farzand paydo bo'ldi. alpinbiy tag'in ikki o'g'il paydo bo'ldi. katta ot boybo'ni qo'ydi, kichkina ot boysari qo'ydi. boybo'ni bilan boysari — ikki katta bo'ldi. boysari boy edi, boybo'ni esa shoy edi, bu ikki ham farzand bo'ldi.

Result

Fig.3. Lemmatizer interface

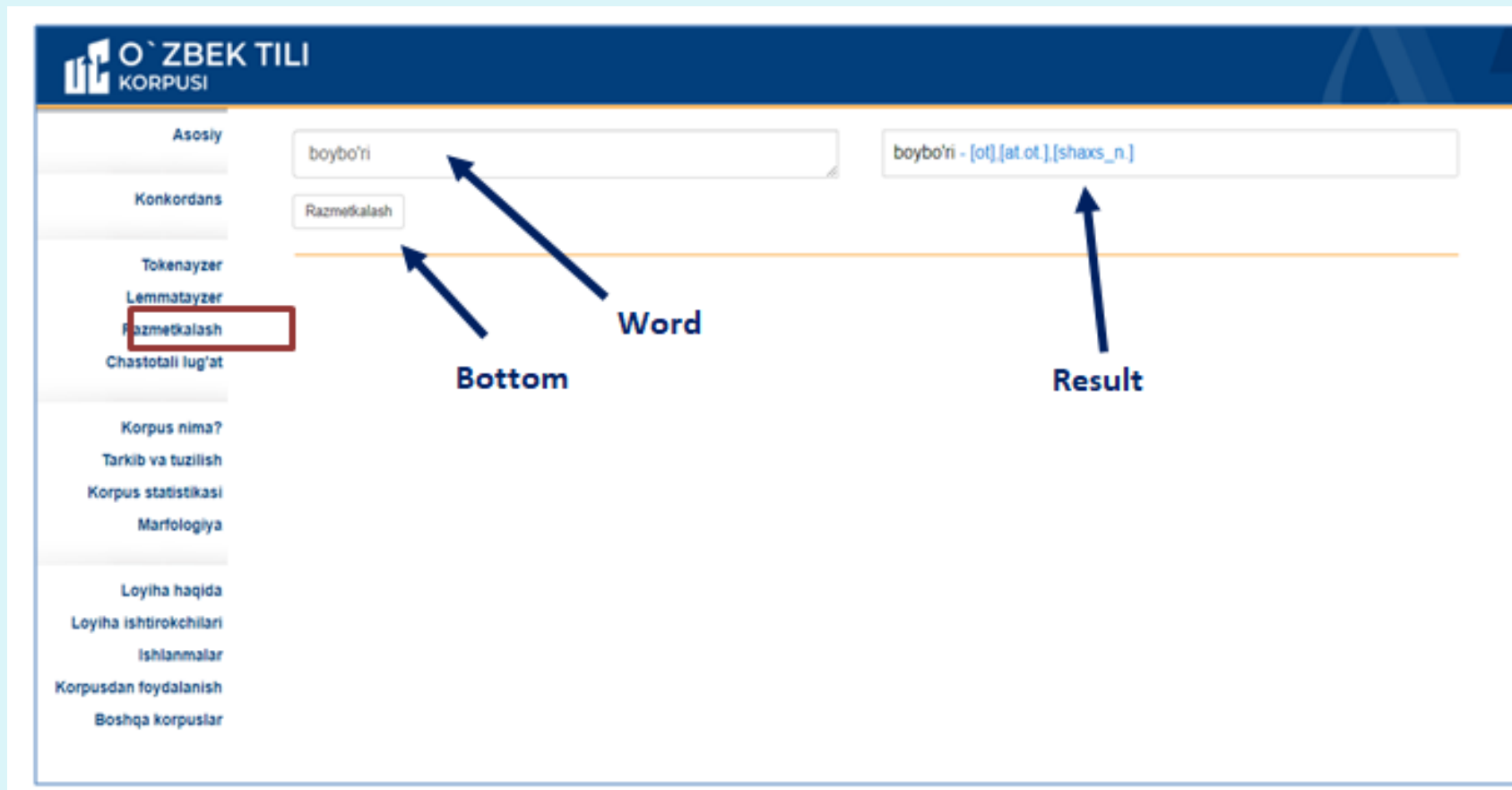


Fig. 4. marking interface

O`ZBEK TILI KORPUSI Bosh sahifa Kirish

KORPUS BO`YLAB SO`Z QIDIRISH
METODIKASI Admin

MATNDAN QDIRILAYOTGAN **BOYBO`RI** SO`ZI ISHTIROK ETGAN ABZASLAR SONI: 3 TA Statistics
 MATNDAN QDIRILAYOTGAN **BOYBO`RI** SO`Z SONI: 5 TA

*SO`Z YOKI QATOR KIRITING:

boybo`ri Searched words

NATUANI KO`RISH **O`CHIRISH**

KORPUS BO`YLAB SO`Z QIDIRISH METODIKASI

Words **Paragra**

Burungi o`tgan zamonda, o`n olti urug` Ko`ng`irot elida Dobonbiy degan o`tdi. Dobonbiydan Alpinbiy degan o`g`il farzand paydo bo`ldi. Alpinbiydan tag`i ikki o`g`il paydo bo`ldi: kattakanining otini **Boybo`ri** qo`ydi, kichkinasining otini Boysari qo`ydi. **Boybo`ri** bilan Boysari — ikkovi katta bo`ldi. Boysari boy edi, **Boybo`ri** esa shoy edi, bu ikkovi ham farzandsiz bo`ldi.

Bu so`zni eshitib, o`n olti urug` Ko`ng`irot elida payga betdan turib bir chapanitab boybachchasi aytdi: — Ey, **Boybo`ri** bilan (14)

afa bo`lib, sakson tilla chufuronga tashlab, turib ketdi. Borib chechib mindi bedov otdi, ikkovi uyiga yetdi, ikkovi qildi maslahatdi, bu so`z bu ikkoviga juda botib ketdi. **Boybo`ri** turib aytdi: — Boysari uko, qariganda bizning molimiz besohibga chikdi, endi bizlar bir farzand taraddi kilmaymizmi? Boysari turib aytdi: —

Fig. 5. Concordance interface

O'ZBEK TILI KORPUSI Bosh sahifa | Kirish

boybo'ri x

Lemma: boybo'ri

Razmetka: [ot],[at.ot],[shaxs_n.]

Close

MATNDAN QIDIRILAYOTGAN **BOYBO'RI** SO'ZI ISHTIROK ETGAN ABZASLAR SONI: 3 TA
 MATNDAN QIDIRILAYOTGAN **BOYBO'RI** SO'Z SONI: 5 TA

*SO'Z YOKI QATOR KIRTING:

Words

NATIJANI KO'RISH O'CHIRISH

KORPUS BO'YLAB SO'Z QIDIRISH METODIKASI

Burungi o'tgan zamonda, o'n olti urug' Qo'ng'iro't elida Dobonbiy degan o'tdi. Dobonbiydan Alpinbiy degan o'g'il farzand paydo bo'ldi. Alpinbiydan tag'i ikki o'g'il paydo bo'ldi: kattakonining otini **Boybo'ri** qo'ydi, kichkinasining otini Boysari qo'ydi. **Boybo'ri** bilan Boysari — ikkovi katta bo'ldi. Boysari boy edi, **Boybo'ri** esa shoy edi, bul ikkovi ham farzandsiz bo'ldi.

Bu so'zni eshitib, o'n olti urug' Ko'ng'iro't elida payga betdan turib bir chapanitob boybachchasi aytdi: — Ey, **Boybo'ri** bilan (14)

afa bo'lib, sakson tilla chufuronga tashlab, turib ketdi. Borib chechib mindi bedov otdi, ikkovi uyiga yetdi. Ikkovi qildi maslahatdi, bu so'z bu ikkoviga juda botib ketdi. **Boybo'ri** turib aytdi: — Boysari uka, qariganda bizning molimiz besohibga chikdi, endi bizlar bir farzand taraddi kilmaymizmi? Boysari turib aytdi: —

Fig. 6. The lemma and mark of the word.