KORPUSDAN QIDIRUV TIZIMI SIFATIDA FOYDALANISH

O‘roqova Dilnoza Odil qizi
Mirzo Ulug‘bek nomidagi O‘zbekiston
Milliy universiteti 2-bosqich magistranti.
E-mail: adilovnadilnoza09@gmail.com

Annotatsiya: Mazkur maqola korpus tushunchasi, uning tildagi o‘rni, korpusning qidiruv turlari, ishlash prinspi 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 komyuter 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 mavjud bo‘lgan zamonaviy matnlardan foydalaniladi.

Yuqorida belgilangan matn turlaridan korpus shakllantirilganda quyidagilami qidirish mumkin:

1) kontekstdagi so‘zlaming 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 ishlash prinsipi 3 turdagi 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 foydali 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 tahlilni 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 teglari 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‘pchiligi 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 tahlilning ikki muhim tomoni hisoblanadi, ya’ni sifat va miqdoriy jihati 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 algaritmi 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’limiy, parallel korpus)
5) Foydalanuvchilar yo‘riqnomasi
6) Elektron lug‘atlar
7) Tezaurus

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

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

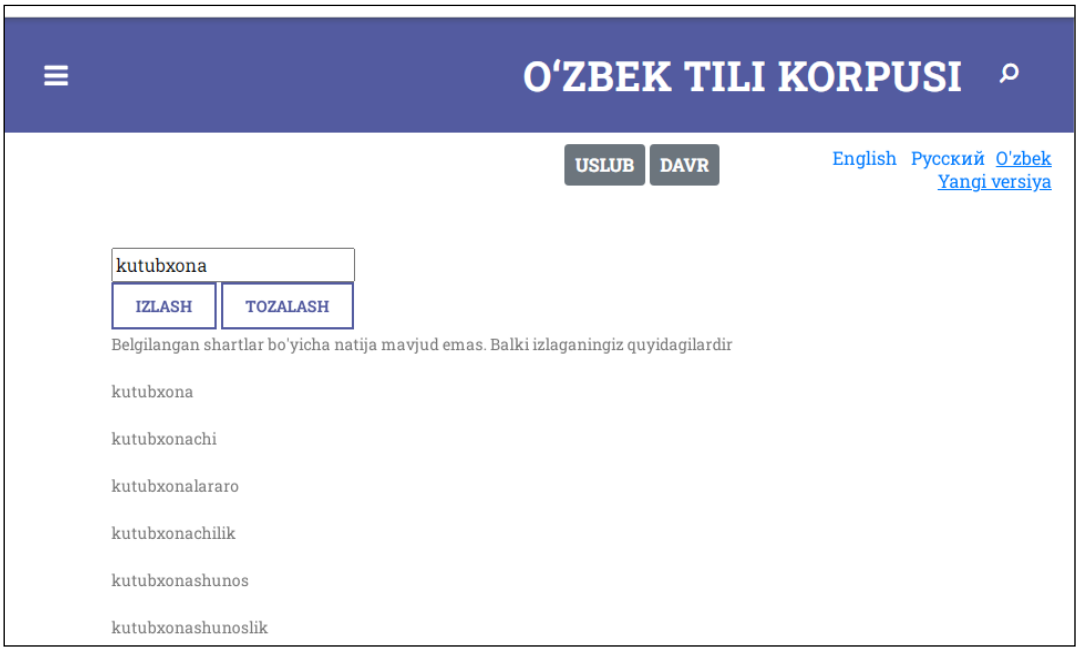
Ushbu electron korpus statistic ma’lumotlarni quyidagi jihatlarga ko‘ra chiqarish imkoniyatiga ega:

1) lemma bo‘yicha qidirish – masalan, kitob so‘zi qidirilganda, ushbu so‘z o‘zak holatda matn tarkibida aniqlanadi

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].

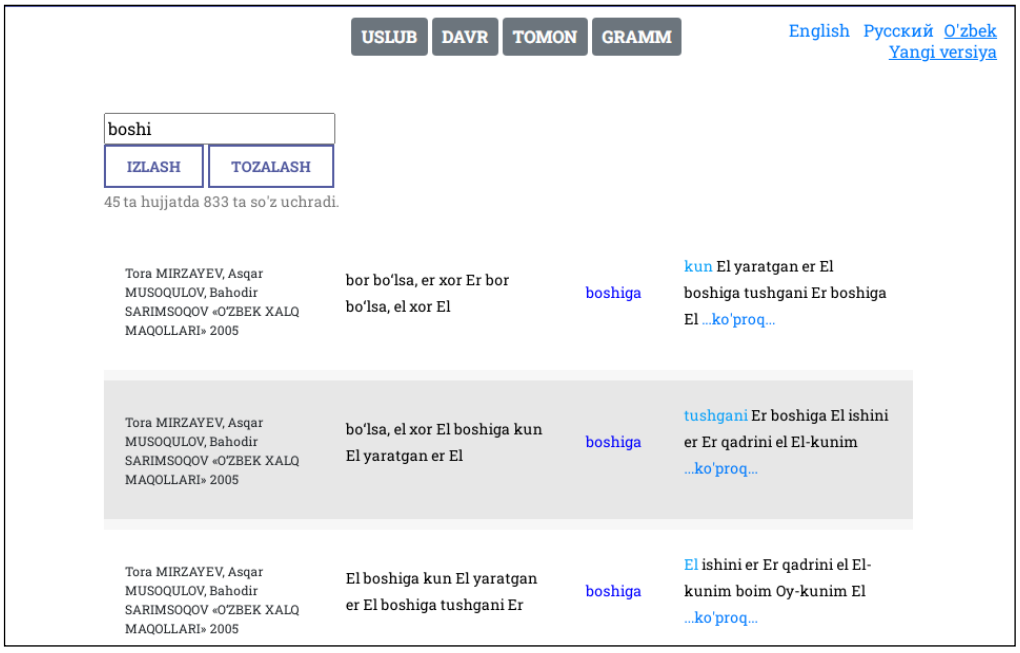
Lemma korpusdan ma’lumot qidirishning asosiy birligi sanaladi. Morfologik tahlil natijasida mustaqil so‘z turkumining statistik natijasi olinadi. Morfologik shakllangan so‘zlar kombinatsion jihatdan turli shakllarga ega bo‘ladi.

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



(asosan, Grammatik belgilari) 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]



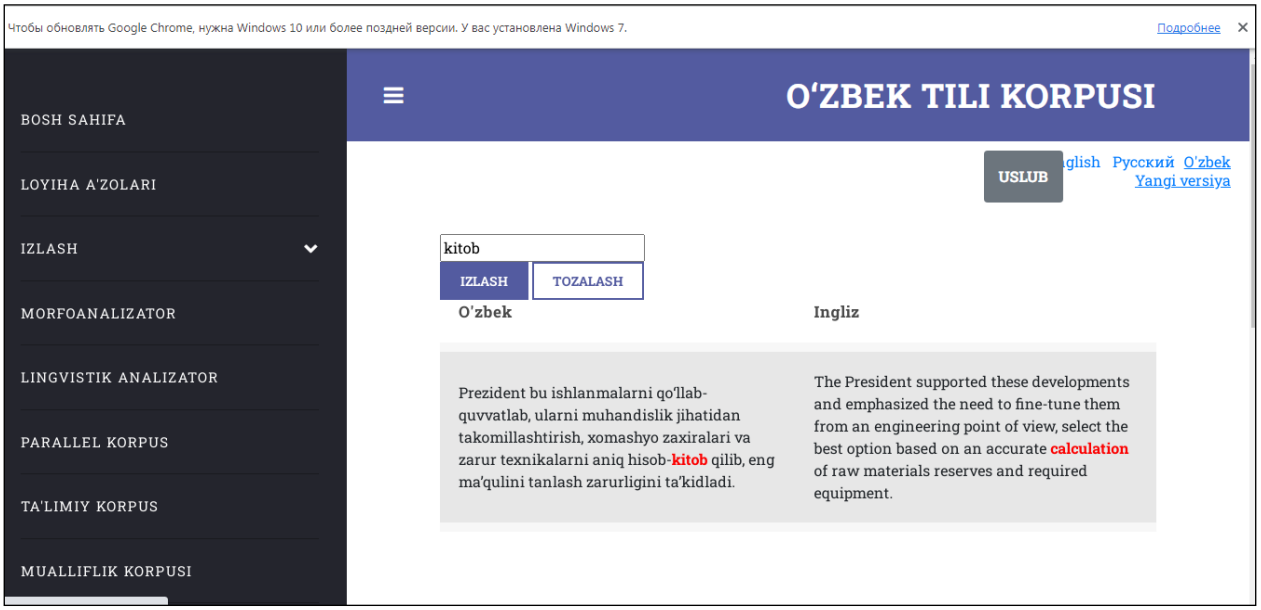
Konkordans qidirishda qidiruv tizimidagi so‘zning o‘zidan oldingi yoki keying keluvchi birliklar bilan birikish holati n-gram bo‘yicha ko‘rsatiladi.

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 grammatik qo‘shimchalar bilan kelish imkoniyati hisobga olinadi.

Ushbu tizim ham lotin, 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.



Korpus qidiruv tizimi sifatida so‘zning bexato ekanligini tasdiqlash yoki uning xatoligini unga yaqin so‘zlar bazasidan qidirish korpus texnologiysi 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 transkripsiyasidan 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’minot asosida joylashtirilgan matnlar jamlanmasi ta’rifini 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., Tuliyev, 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.