O‘ZBEK TILI MILLIY KORPUSI TUZILISHI: ASOSIY KOMPONENTLAR VA AXBOROT TURLARI

Primova Mastura Hakim qizi,
O‘qituvchi
primovamastura@navoiy-uni.uz
ToshDO‘TAU

Annotatsiya. O‘zbek tili milliy korpusi tilshunoslik, adabiyotshunoslik va sun’iy intellekt sohalarida qo‘llaniladigan murakkab tizim bo‘lib, uning tuzilishi matnlar to‘plami, annotatsiya qatlamlari va ma’lumotlar bazasi arxitekturasi asosida shakllantirilgan. Korpusning asosiy komponentlariga turli davrlar (zamonaviy, tarixiy), janrlar (badiiy, ilmiy, rasmiy) va dialektlardagi matnlar, shuningdek, ularga qo‘shilgan morfologik, sintaktik va semantik annotatsiyalar kiradi. Ushbu maqolada korpusning tuzilishi, asosiy komponentlari va undagi axborot turlari atroflicha o‘rganiladi. Tadqiqot korpus lingvistikasi usullariga asoslanib, matnlarning turli janrlari va manbalarini tahlil qiladi. Korpus matnlar to‘plami, morfologik, sintaktik va semantik annotatsiya, metama’lumotlar va qidiruv tizimi kabi asosiy qismlardan iborat ekani aniqlanadi. Natijalar shuni ko‘rsatadiki, korpus turli davrlar va janrlarga oid matnlarni qamrab, tilshunoslik tadqiqotlari va sun’iy intellekt loyihalari uchun katta imkoniyatlar beradi. Korpusning foydalanuvchilar uchun qulay interfeysi va qidiruv vositalari uning qo‘llanilishini yanada samarali qiladi. Maqolada korpusning kelajakda rivojlanishi, xususan, annotatsiya sifatini oshirish va yangi texnologiyalar bilan integratsiya qilish bo‘yicha takliflar beriladi. Ushbu tadqiqot o‘zbek tili milliy korpusining ahamiyatini ochib beradi va tilshunoslar hamda texnologlar uchun qimmatli resurs bo‘lib xizmat qiladi.

Kalit so‘zlar: *O‘zbek tili milliy korpusi, korpus lingvistikasi, leksik-grammatik teglash, korpus platformalari, kollokatsiya, KWIC (Key Word in Context), lingvistik teglash, metama’lumotlar (metadata).*

Kirish

Til korpusi – bu ma’lum bir tilga oid yozma (va nutq) matnlarning elektron shaklda jamlangan ulkan to‘plami bo‘lib, u foydalanuvchi uchun maxsus qidiruv tizimi vazifasini ham o‘taydi[1]. Ya’ni, korpus – raqamlashtirilgan matnlarni lingvistik jihatdan teglangan (annotatsiyalangan) holda saqlovchi va ulardan turli xil axborot olish imkonini beruvchi kompleks tizimdir. Milliy korpuslar odatda o‘z ichiga millionlab so‘zlarni qamrab olgan matnlar majmuini oladi; masalan, Britaniya milliy korpusi (BNC) taxminan 100 million so‘zli ingliz tilining turli janrdagi matnlarini o‘z ichiga oladi, Rus milliy korpusi (RNC) esa 1 milliarddan ortiq so‘z shakllarini jamlagan holda doimiy ravishda kengayib borayapti. Korpus **tarkibidagi matnlar turli manbalardan**, jumladan **badiiy adabiyotlar, ilmiy maqolalar,**

ommaviy axborot vositalari materiallari va **og‘zaki nutq yozuvlari**dan tanlab to‘planadi [2]. Elektron korpus oddiy matn arxivigina emas, balki undagi har bir so‘zni va gapni lingvistik jihatdan tahlillab belgilovchi hamda foydalanuvchi so‘rovlariga tezkor javob beruvchi axborot tizimidir.

Korpus tuzilmasining darajalari. Har qanday zamonaviy til korpusi bir necha darajadagi asosiy komponentlar yig‘indisidan tashkil topadi:

1) **matn bazasi va lingvistik teglash** – korpusdagi barcha matnlarning struktura va til birligidagi belgilarini o‘z ichiga oladi;
2) **texnik arxitektura** – ma’lumotlarni saqlash, qayta ishlash va qidiruvni ta’minlovchi dasturiy-apparat tizimlarini qamrab oladi;
3) **foydalanuvchi interfeysi** – korpus bilan ishlash uchun qulay grafikali yoki veb-ishchi muhit bo‘lib, unda qidiruv, filtr, statistik chiqish kabi funksiyalar mujassam.

Quyida ushbu har bir komponent batafsil yoritiladi, shuningdek, korpuslardan olinadigan axborot turlari – lingvistik belgilash, metama’lumotlar va statistik ma’lumotlar alohida tahlil qilinadi.

Matn bazasi va lingvistik komponentlar

Matnlar va ularning tuzilishi. Korpusning asosini turli manbalardan to‘plangan matnlar bazasi tashkil qiladi. Korpusga kiritilayotgan har bir matn odatda oldindan raqamli ko‘rinishga keltiriladi va maxsus formatga solinadi (masalan, UTF-8 kodida matn). Matn ichidagi birliklar eng kichik **tokenlarga ajratiladi** – token bu so‘z, son, tinish belgisi kabi ajratib olinuvchi minimal elementdir. Tokenlarga ajratish (tokenizatsiya, tokenlash) natijasida **matn gaplarga** va **gaplar so‘zlarga** bo‘linadi. Har bir so‘z tokeniga lingvistik xususiyatlar bog‘lanadi: **uning lemmasi** (lug‘aviy asosi yoki boshlang‘ich shakli) **aniqlanadi** va **so‘z turkumi, grammatik kategoriya** kabi teglar bilan belgilab chiqiladi. Masalan, Rus milliy korpusida 1 milliarddan ziyod so‘z shakllari avtomatik lemma va morfologik teglar bilan belgilangan bo‘lib, har bir so‘zning barcha mumkin bo‘lgan grammatik tahlillari unga biriktirilgan. Shu tariqa, korpus matnlari ichida qidiruv faqat xom tekst bo‘yicha emas, balki lemmatlangan va teglangan shaklda amalga oshiriladi – foydalanuvchi so‘zning aynan qaysi grammatik turkumda kelishini ham tanlab izlashi mumkin.

Leksik-grammatik teglash. Korpus tarkibidagi so‘z tokenlariga biriktiriladigan lingvistik teglash (annotatsiya) bir necha qatlamdan iborat bo‘lishi mumkin. Eng asosiy qatlam – **morfologik teglash**, ya’ni har bir so‘zning qaysi so‘z turkumi (masalan, **ot, fe’l, sifatl– yoki part-of-speech, POS**) ekanini ko‘rsatish va uning grammatik grammemalarini (kelishik, zamon kabi kategoriyalarni) belgilashdir. Korpus yaratuvchilari til strukturasi murakkabliklarini e’tiborga olgan

holda bunday teglar tizimini ishlab chiqadilar. Masalan, RNC misolida har bir so‘z shakli uchun uning lemmasi va so‘z turkumi, so‘ngra so‘z turkumiga xos doimiy grammatik xususiyatlar va ayni o‘sha kontekstdagi o‘zgaruvchi grammatik shakllari ko‘rsatiladi [3]. Korpusning teglangan qismida shu to‘liq morfologik tahlil foydalanuvchiga ko‘rsatiladi, qolgan qismida esa faqat lemma va so‘z turkumi aks ettiriladi.

Sintaktik struktura. Lingvistik teglashning keyingi darajasi – sintaktik annotatsiya bo‘lib, ba’zi korpuslarda har bir **gapning tuzilishi (parsing)** ham saqlanadi. Bu sintaktik bog‘lanishlar ko‘rinishida (odatda, daraxt struktura shaklida) ifodalanadi: gap tarkibidagi so‘zlar orasidagi sintaktik munosabatlar (boshlovchi, to‘ldiruvchi, aniqlovchi va hokazo) **maxsus teglar** bilan bog‘langan graflar tarzida beriladi. Masalan, Rus milliy korpusining **SynTagRus** deb nomlangan maxsus sintaktik subkorpusi mavjud bo‘lib, u 1 million atrofidagi gaplarga to‘liq sintaktik daraxt tuzilmalarini biriktirgan. Bunday daraxtbank (treebank) korpuslar til sintaksisini o‘rganish va mashinada tushunishni rivojlantirishda muhim manba hisoblanadi. Ingliz tilida **Penn Treebank**, nemis tilida **TIGER Corpus** kabi mashhur sintaktik teglangan korpuslar mavjud. Sintaktik teglash tufayli foydalanuvchi korpusdan nafaqat alohida so‘zlarni, balki ma’lum bir **sintaktik konstruksiyalarni** ham izlab topishi mumkin bo‘ladi (masalan, **“ot+fe’l+ot”** tuzilmasidagi gaplar va hokazo).

Agglyutinativ tillarda belgilash muammosi. Ta’kidlash joizki, turkiy agglyutinativ tillarda (so‘zga ko‘plab qo‘shimchalar ketma-ket qo‘shiluvchi tillarda) korpus uchun morfologik analiz va so‘zni asos (lemma)ga keltirish masalasi murakkabroq bo‘ladi. O‘zbek tilini ham qamrab oluvchi tadqiqotlarda aynan POS-teglash va stemming (so‘zdan qo‘shimchalarni ajratib, negiz shaklini topish) muammolari dolzarb ekani qayd etilgan. Jumladan, B. Elov, Sh. Hamroyeva va boshqalar o‘zlarining ilmiy ishlarida agglyutinativ tillarda so‘zning grammatik kategoriyasini avtomatik aniqlash qiyinligi, ba’zan qo‘shimchalar tufayli ko‘p ma’noli holatlar paydo bo‘lishi va ularni bartaraf etish uchun maxsus algoritmlar zarurligini ta’kidlaydilar [4]. Shu bois, zamonaviy o‘zbek korpuslarini yaratishda **morfologik analizatorlar** va **lemmatizatorlar** ishlab chiqish alohida e’tibor talab qilmoqda. Masalan, o‘zbek tili uchun yaratilgan **https://uznatcorpara.uz/** kabi onlayn morfologik analizatorlar korpusdagi so‘zlarni tahlil qilishga xizmat qilmoqda.

Texnik arxitektura va platforma tizimi

Arxitektura ahamiyati. Korpusning muvaffaqiyatli ishlashi uchun uning ichki texnik arxitekturasi juda muhimdir. Faqatgina matn hajmini oshirish bilan cheklanib bo‘lmaydi – *agar korpus arxitekturasi zaif bo‘lsa* yoki u *lingvistik jihatdan teglanmagan* bo‘lsa, foydasi cheklangan bo‘lib qoladi. Mark Deyvis

(COCA korpusi asoschisi) ta’kidlaganidek, korpus hajmi **“qancha katta bo‘lsa shuncha yaxshi”** degan qoidaga faqat to‘g‘ri arxitektura va to‘liq annotatsiya bilan erishish mumkin; aks holda 100 million so‘zli yaxshi belgilangan korpus 1 milliard so‘zli ammo belgilanishi sust korpusdan afzalroq natija berishi mumkin [5]. **Optimal arxitektura** matnlarni samarali indekslash, tezkor qidiruv va katta hajmda ham barqaror ishlashni ta’minlaydi.

Ma’lumotlar bazasi tuzilishi. Korpus ichki tizimi ko‘pincha maxsus ma’lumotlar bazasi yoki indekslash usullariga tayangan holda quriladi. Korpusdagi har bir so‘z alohida yozuv (record) sifatida saqlanishi, unga lemma va teglar bog‘lanishi mumkin. Zamonaviy yirik korpuslarda tezkor qidiruv uchun ko‘pincha **relatsion ma’lumotlar bazasi arxitekturasi qo‘llaniladi** (masalan, COCA va boshqa English-Corpora.org korpuslari shu usulda). Bunda har bir so‘z maxsus sonli identifikator (ID) bilan bazaga kiritiladi, so‘zlarning matndagi ketma-ketligi indekslar orqali saqlanadi. Masalan, COCA korpusida so‘zlar raqamlar ko‘rinishida *“corpus”* jadvalida saqlanib, alohida *“lexicon”* jadvalida IDga mos so‘zning o‘zi, lemmasi va POS-tegi saqlanadi; shu tariqa massiv matnlarda ham ma’lumotga murojaat juda tez bajariladi [6]. Bunday arxitektura yordamida foydalanuvchi milliardlab so‘zli korpuslarda ham bir necha soniya ichida murakkab so‘rovlarni bajarishi mumkin – masalan, berilgan so‘zdan oldin keluvchi eng ko‘p 100 sifatni butun korpus bo‘ylab sanash kabi og‘ir so‘rovlar 1-2 soniyada natija beradi. Shuningdek, **klasterlangan indekslar** va **ma’lumotlarni ixtiyoriy kontekst bo‘yicha saqlash** kabi usullar qidiruvni optimallashtiradi. Korpus arxitekturasi to‘g‘ri tashkil etilgan bo‘lsa, hatto maxsus korpus dasturlari (masalan, Sketch Engine)dan ham tezroq natija olish mumkinligi tajribada ko‘rsatilgan.

Korpus platformalari. Bugungi kunda korpus tuzish uchun tayyor dasturiy platformalar va tizimlar mavjud. Masalan, jahon tajribasida **Sketch Engine, CQPweb** kabi platformalar ko‘plab korpuslar uchun universal yechim sifatida qo‘llanadi – ularda **matnlarni kiritish, avtomatik teglash** va **qidiruv interfeysi** birgalikda ta’minlanadi. Ayrim milliy korpuslar esa o‘ziga mos maxsus platformada yaratilgan: masalan, RNC uchun Maxsus qidiruv dasturi (Yandex bilan hamkorlikda) ishlab chiqilgan bo‘lib, u korpusning morfologik va sintaktik qidiruvlarini ta’minlaydi [4]. O‘zbek tilining milliy korpusi (https://uzschoolcorpara.uz/) ham maxsus veb-platforma sifatida ishga tushirilgan: unda matn bazasi serverda SQLServer ma’lumotlar bazasida saqlanadi, foydalanuvchi so‘rovlari Python yordamida qayta ishlanib, natijalar sahifada ko‘rsatiladi (loyiha tafsilotlarida shu keltirilgan). Korpus platformasining muhim qismi – bu lingvistik ishlov berish modullaridir: masalan, so‘zlarni avtomatik teglash uchun morfologik analizator yoki maxsus modellar ulanadi. RNC misolida Mystem va Dialing morfologik dasturlari integratsiya qilingan bo‘lib, ular matnlardagi so‘zlarning formasini aniqlashda ishlatiladi. O‘zbek tilida ham shunga o‘xshash

tarzda tayyor dasturiy modullar (https://uznatcorpara.uz/) korpusga ulanmoqda. Umuman, texnik arxitektura jozibadorligi shundaki, yaxshi optimizatsiya qilingan tizim katta hajmdagi korpuslarda ham foydalanuvchiga bir zumda javob qaytara oladi va yangi matnlarni qo‘shish ham oson kechadi.

Foydalanuvchi interfeysi va qidiruv imkoniyatlari

Interfeys dizayni va funksional. Korpusdan samarali foydalanish uchun uning foydalanuvchi interfeysi qulay va funksional bo‘lishi lozim. Interfeys foydalanuvchiga oson tushunarli bo‘lishi, turli ehtiyojlarga mos qidiruv turlarini taklif qilishi zarur. Odatda korpusning veb-interfeysi orqali quyidagi asosiy imkoniyatlar taqdim etiladi:

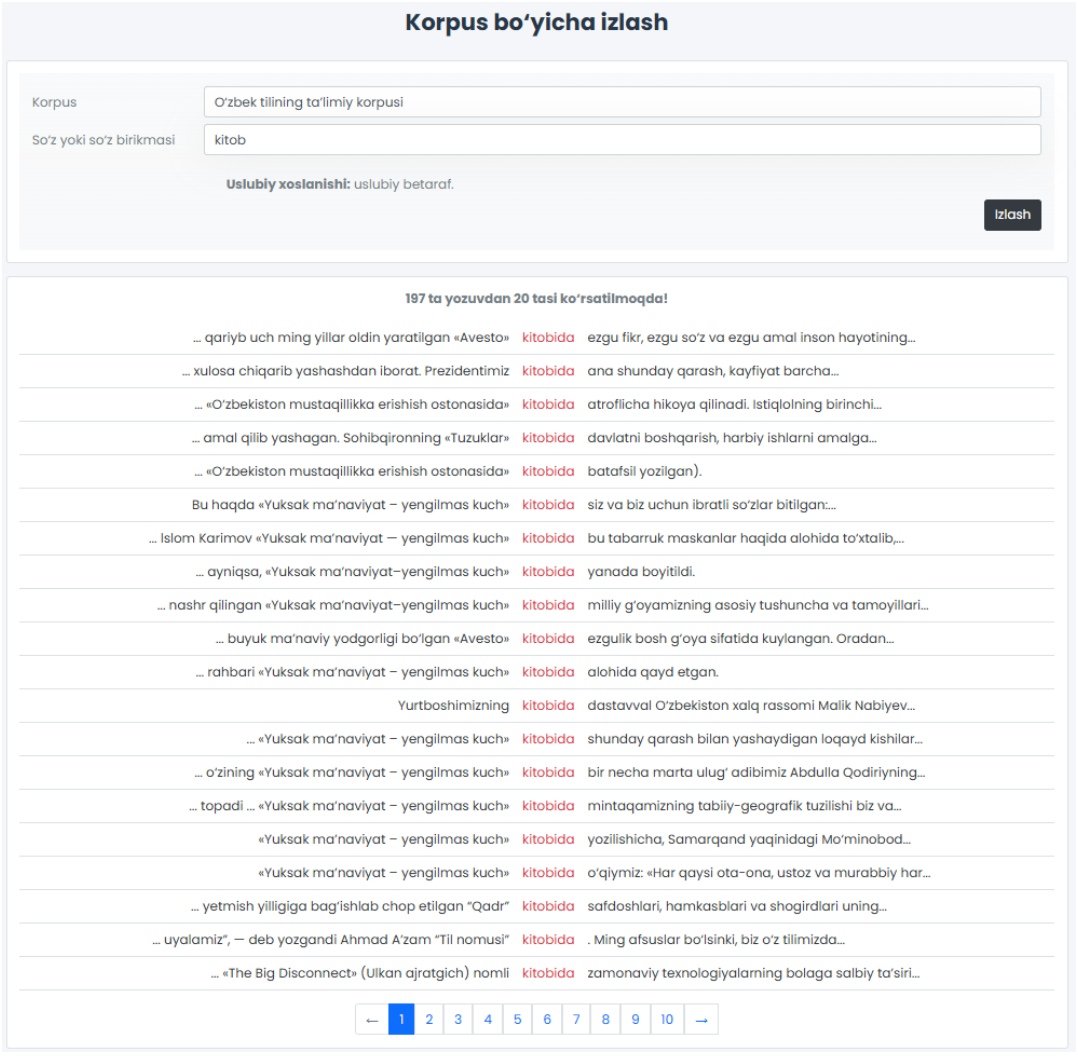
1) **Oddiy qidiruv** – matndan ma’lum bir so‘z yoki iborani izlash (bu Google qidiruviga o‘xshash, lekin korpus ichida amalga oshiriladi). Foydalanuvchi qidiruv satriga so‘z kiritadi va korpusdan ushbu so‘z uchragan barcha kontekstlar olinadi.

2) **Kengaytirilgan qidiruv** – foydalanuvchi lemma bo‘yicha (ya’ni so‘zning barcha shakllarini qamrab) yoki ma’lum so‘z turkumi bo‘yicha so‘rov berishi mumkin. Masalan, faqat fe’l lemmasini qidirish, yoki *“kitob_Noun”* kabi so‘z turini ko‘rsatib izlash imkoniyati. Bunday qidiruv POS-teglar yordamida amalga oshadi. RNC interfeysida maxsus *“Лексико-грамматический поиск”* degan bo‘limda foydalanuvchi so‘z yonida grammatik xususiyatlarni belgilashi mumkin – masalan, *“OT, birlik, Qaratqich kelishigi”* kabi.

3) **Murakkab qidiruv** – bir nechta so‘zlarning o‘zaro yaqinligini yoki bir gap ichida birga kelishini qidirish. Masalan, *“A va B so‘zlari bir gapda masofasi 5 tokendan oshmaydigan holda uchrasin”* kabi murakkab shartlar qo‘yilishi mumkin. Bunday qidiruvlar **CQL (Corpus Query Language)** yoki grafik forma orqali qo‘llab-quvvatlanadi.

4) **Filtrlash** – bu qidiruv natijalarini yoki qidiruvning o‘zini ma’lum metama’lumotlarga ko‘ra cheklashdir. Masalan, foydalanuvchi *faqat badiiy adabiyot janridagi matnlardan qidirmoqchi* bo‘lsa, kerakli filtrni tanlaydi. Yoki *ma’lum yillar oralig‘idagi matnlardan izlash*, yoki *muallifi ayol bo‘lgan matnlar* doirasidagina qidirish kabi. Jahon korpuslarida bunday filtrlash oddiy *“bir necha bosqichda”* amalga oshadi: masalan, COCA interfeysida foydalanuvchi **janr, yil, manba bo‘yicha saralab olish** orqali o‘ziga xos virtual kichik korpus yaratishi mumkin. Inglizcorpora saytida bir nechta belgi tanlash bilan, masalan, *“2000-yillarning jurnallari ichida dasturlash mavzusidagi matnlar”* kabi tanlangan bo‘limni bir necha soniyada shakllantirish imkoni borligi ko‘rsatilgan. O‘zbek tilining milliy korpusida ham matnlar **uslub (janr) bo‘yicha (ilmiy, badiiy, publitsistik, rasmiy, so‘zlashuv), yaratilgan vaqti bo‘yicha (sovet davri, mustaqillik davri kabi)** filtrlanishi nazarda tutilgan.

Natijalarni chiqarish shakllari. Korpus interfeysi qidiruv natijalarini foydalanuvchiga qulay tarzda namoyish qilishi ham muhimdir. Odatda **natijalar konkordans ko‘rinishida taqdim etiladi** [7]. **Konkordans** – bu kalit so‘zning atrofidagi kontekst bilan bir qatorda bir necha qatorda ko‘rsatilishidir. Masalan, foydalanuvchi “ilm” so‘zini izlasa, korpus dasturi ushbu so‘z uchragan har bir gapni alohida qator qilib, so‘zning o‘zi markazda ko‘rinadigan qilib chiqaradi. Bu usul **KWIC (Key Word in Context)** deb ham ataladi. Konkordans foydalanuvchiga qidirilgan so‘zning turli kontekstlarda qanday ishlatilishini tezda ko‘zdan kechirishga yordam beradi – muhim lug‘aviy va grammatik namunaklarni aniqlash osonlashadi. Misol uchun, 1.2.1-rasmda keltirilgan konkordansda *“day by day”* iborasi har xil misollarda keltirilgan bo‘lib, har bir qator korpusning boshqa bir hujjatidan olingan (bir qator ikkinchi qator bilan bevosita bog‘liq matn emas).

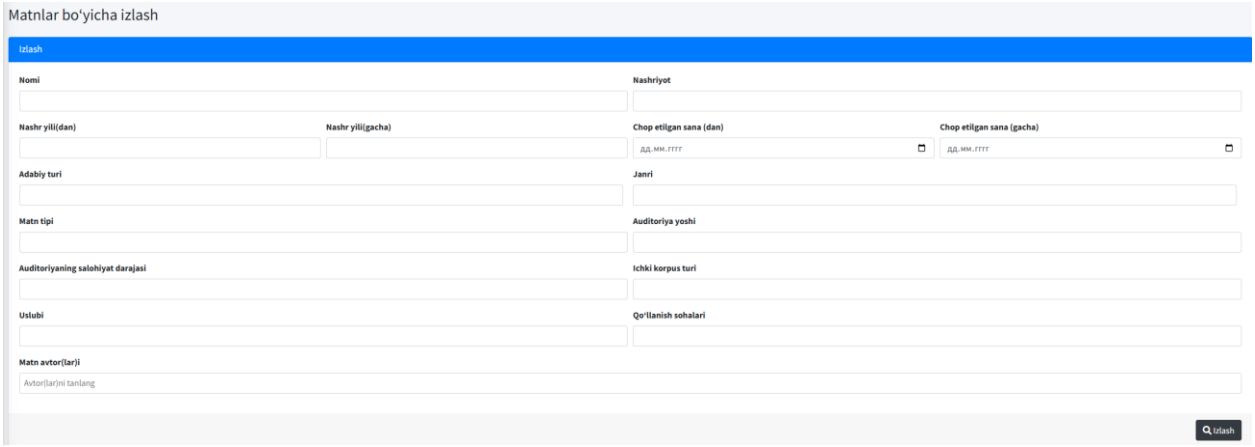


1-rasm. KWIC (Key Word in Context)

Korpus interfeysi odatda bir vaqtning o‘zida **30–50 ta** konkordans qatordan natijani ko‘rsatadi, foydalanuvchi esa keyingi sahifalarga o‘tib barcha topilmalarni

ko‘rib chiqishi mumkin. Bundan tashqari, har bir konkordans qator yonida matn manbasi ko‘rsatiladi (masalan, “Xalq so‘zi gazetasi, 2018 yil”) va u ustiga bosib to‘liq matnni ko‘rish imkoni ham bo‘lishi mumkin.

Statistik funksiyalar. Korpus foydalanuvchilari ko‘pincha individual misollar bilan cheklanmasdan, matnlar bo‘yicha statistik ma’lumotlarni olishga ham qiziqadilar. Shuning uchun zamonaviy korpus interfeyslari **so‘z chastotasi, kollokatsiyalar, ngramlar** kabi vositalarni ham taqdim etadi. Misol uchun, COCA korpusida foydalanuvchi istalgan so‘zni qidirganda, uning umumiy chastotasi (korpusda necha marta uchragani) hamda janrlar kesimidagi chastotasi (masalan, og‘zaki nutqda 500 marta, ilmiy matnlarda 200 marta kabi) grafigi bilan ko‘rsatiladi.



2-rasm. Korpus qidiruv tizimi

Bu bilan so‘zning qaysi uslubda ko‘proq ishlatilishi yoki vaqt o‘tishi bilan chastotadagi o‘zgarishi tahlil qilinadi. **Kollokatsiya** – tanlangan so‘zning yon-atrofida eng ko‘p uchraydigan so‘z birikmalaridir; masalan, korpus tahlili natijasida *“ilm”* so‘zi bilan eng ko‘p birga keluvchi so‘zlar *“fani”, “sohasida”, “arbobi”* kabi kollokatsiyalar ekani aniqlanishi mumkin. Korpus dasturlari buni aniqlash uchun statistik o‘lchovlardan (MI – o‘zaro axborot miqdori yoki t-skori kabi) foydalanadi va foydalanuvchiga Top-20 eng muhim kollokatsiyalar ro‘yxatini bera oladi. Masalan, COCA korpusida biror so‘z uchun Collocates bo‘limi mavjud bo‘lib, u so‘zning chap yoki o‘ng tomonida berilgan masofa ichida eng ko‘p uchragan so‘zlarni va ularning ko‘rsatkichlarini chiqaradi. Yuqorida aytib o‘tilganidek, maxsus **so‘rov orqali ham kollokatsiyani topish mumkin:** masalan, SQL so‘rov tili misolida “kitob” so‘zidan oldin keluvchi sifatlar topilib, ular orasidan 100 tasi saralanishi misoli keltirilgan edi – amalda foydalanuvchi buni oddiy interfeys orqali bir necha belgi tanlab bajaradi, tizim esa ortda bunday kodni ishlatadi.

N-grammalar esa korpusdagi so‘zlarning eng ko‘p uchraydigan ketma-ketliklaridir (N so‘zdan iborat frazalar). Korpusdan, masalan, eng ko‘p qo‘llanilgan **ikki so‘zli birikmalar (bigram)** yoki **uch so‘zli birikmalar (trigram)** avtomatik

chiqarib olinishi mumkin. Bu tilning *iboralari, maqol* va *iqtiboslarini* aniqlashda foydali. Masalan, o‘zbek tilining ilk milliy korpusi matnlari asosida n-gram usuli orqali til modelini qurish tajribasi amalga oshirilgan bo‘lib, bunda korpusdagi eng ko‘p uchraydigan so‘z ketma-ketliklari aniqlangan. Natijada, masalan, o‘zbek tilida eng ko‘p takrorlanadigan uch so‘zli iboralar (*“aloqa vositalari orqali”*, *“Prezidenti matbuot xizmati ma’lum qildi” singari*) ajratib olinishi mumkin – bu ma’lumotlar til xususiyatlarini o‘rganishda yoki mashina tarjimasida qo‘l keladi[8].

Yuqoridagi funksiyalar barchasi korpus interfeysi orqali birlashtirilgan holda taqdim etiladi [9]. Yaxshi interfeys oddiy foydalanuvchilar uchun sodda qidiruvni, lingvistlar uchun esa murakkab tahliliy so‘rovlarni amalga oshirishga imkon yaratadi. Misol tariqasida, Britaniya milliy korpusining avvalgi BNCweb interfeysi ham, hozirgi Sketch Engine dagi versiyasi ham foydalanuvchiga yuqorida sanalgan barcha imkoniyatlarni beradi – **so‘zlarni regex (muntazam ifodalar) bo‘yicha qidirishdan tortib**, matnlarni metama’lumotlar bo‘yicha ajratish va hatto tezavror (thesaurus) usulida o‘xshash so‘zlarni topishgacha. Rus milliy korpusining onlayn tizimi esa rus tilining o‘ziga xos grammatikalarini (masalan, fe’lning vid kategoriyasi, otning irglik/nisbat kategoriyasi) ham qidiruv shartlari sifatida qo‘shishga ixtisoslashgan. O‘zbek tilining elektron korpusi uchun ham shunday imkoniyatlarni bosqichma-bosqich joriy etish rejalangan – xususan, UzbCorpus.uz saytining 2020 yilgi ilk versiyasida token, lemma bo‘yicha oddiy qidiruvlar ishga tushirilgan bo‘lsa, keyinchalik kengaytirilgan qidiruv turlari ham qo‘shib borilmoqda.

Korpusdagi axborot turlari

Korpus tizimida turli axborot turlari qatlamlari mavjud bo‘lib, ularni uch asosiy guruhga ajratish mumkin:

1) **lingvistik axborot** – matnlarning lingvistik tavsifi va teglanishi (morfologik, sintaktik va boshqalar);

2) **metama’lumotlar** – har bir matnning tavsiflovchi ma’lumotlari (yaratilgan vaqti, janri, muallifi, manbasi, tili va hokazo);

3) **statistik axborot** – korpusdan foydalanuvchi so‘rov natijasida olinadigan chastotaviy yoki taqqoslovchi ko‘rsatkichlar (so‘zlar chastotasi, konkordans natijalari, kollokatsiyalar, n-gramma ro‘yxatlari va boshqalar).

Yuqorida biz aslida bu turdagi axborotlarning barchasiga to‘xtaldik – quyida ularni yana tizimli ravishda sharhlaymiz.

Lingvistik teglash qatlami

Lingvistik axborot korpus matnlarining annotatsiya qilingan qatlamidir. Uning tarkibiga yuqorida ta’riflangan **morfologik teglar, sintaktik strukturalar,**

shuningdek, ba’zan **semantik teglash** ham kiradi. Lingvistik teglash tufayli korpus *“oddiy matnlar to‘plami”*dan haqiqiy **til modellari bazasiga** aylanadi – chunki har bir so‘z va gap haqida mashina o‘qiy oladigan boy ma’lumotlar biriktirilgan bo‘ladi[10]. Morfologik teglashning ahamiyati shundaki, u *korpusdan ma’lum grammatik shakllarni izlash*, *grammatik hodisalar tezligini aniqlash* kabi ilmiy ishlarni amalga oshirishga imkon beradi. Masalan, lingvist olim ingliz tilida fe’lning zamon shakllari qo‘llanish dinamikasini o‘rganmoqchi bo‘lsa, to‘g‘ridan-to‘g‘ri korpusdan *“Past Perfect”* bo‘lib kelgan fe’llarni yillar kesimida sanab bera oladi – buning uchun korpusdagi har bir fe’lning zamon kategoriyasi belgilangan bo‘lishi kifoya. Rus milliy korpusida hatto so‘zlarning **semantik xususiyatlari** ham qisman belgilangan: masalan, unda *“inson”* yoki *“hayvon”* semantik sinfiga mansub otlar alohida belgiga ega, yoki **onomastik** (*shaxs ismi, joy nomi*) otlar alohida ajratilgan. Bunday chuqur belgilash qatlami keyinchalik korpusdan, masalan, *“hayvonlarni ifodalovchi otlar ishtirok etgan frazeologizmlar”* kabi nozik so‘rovlarni amalga oshirishda asqotadi.

O‘zbek tilining elektron korpuslari hozircha to‘liq lingvistik teglashga ega boshlang‘ich bosqichda. B.Elov va boshq. (2024) ishlanmalarida o‘zbek matnlarini avtomatik teglash va tahlil qilish ustida ishlar olib borilmoqda [11]. Jumladan, yaratilgan korpus imkoniyatlarini kengaytirish uchun xorijiy tajribadan foydalanilayotgani, xususan, og‘zaki matnlar korpusini ham yaratish va uni teglash masalalari o‘rganilayotgani qayd etiladi. Bu shuni ko‘rsatadiki, yaqin yillarda o‘zbek korpuslarida ham morfologik belgilash bilan birga sintaktik va semantik annotatsiya qatlamlari paydo bo‘lishi kutilmoqda. Shu tariqa, lingvistik axborot qatlami korpusning *“yuragi”* bo‘lib, u orqali korpus ilmiy-tadqiqotlarda qo‘llanadi va tilning raqamli modelini yaratadi. Aynan lingvistik belgilash tufayli korpus **milliy tilni chuqur o‘rganishda beqiyos manba bo‘la oladi** – bu haqda mahalliy tadqiqotchilar (B. Elov, Sh. Hamroyeva va boshq.) ham o‘z ishlarida ta’kidlaydi.

Metama’lumotlar (metadata)

Har bir korpus matni qandaydir manbadan olingan bo‘ladi va uning o‘ziga xos metama’lumotlari mavjud. **Metama’lumot** – bu matn haqidagi ma’lumot. Korpus dizaynida matnlarni tanlash va tasniflash juda muhim bo‘lgani sababli, har bir kiritilayotgan matnga tegishli metama’lumotlar biriktiriladi: **yaratilgan sanasi, janri, uslubi, mavzusi, muallifi, manbasi (masalan, qaysi jurnal yoki sayt), tili (agar original matn boshqa tilda bo‘lsa yoki o‘zbekcha lotin yozuvida yoxud kirill yozuvida bo‘lishi)** kabi ko‘plab tavsiflar ko‘rsatiladi. Bu ma’lumotlar korpus tuzishda ikki sababga ko‘ra muhim: birinchidan, **balanslangan korpus** tuzish uchun – ya’ni korpus tarkibida turli davr va turli janrlarning ulushi muvozanatli bo‘lishini nazorat qilish; ikkinchidan, keyinchalik foydalanuvchi bu metama’lumotlar orqali saralash va filtrlashni amalga oshirishi uchun. Misol tariqasida, BNC korpusi maxsus balanslash rejasi asosida tuzilgan bo‘lib, unda matnlar 1960–1990 yillardagi

ingliz tilini qamrab oluvchi 100 million so‘z hajmida yig‘ilgan, shundan 10% og‘zaki (nutq transkriptlari) va 90% yozma matnlar; yozma matnlar esa o‘z navbatida ilmiy, publitsistik, badiiy adabiyot, rasmiy hujjatlar kabi toifalarga bo‘lingan. Shu tarzda BNC ichida har bir matnga uning manbasi (masalan, “The Times gazetasi, 1991 yil, yangilik xabari” kabi) haqida batafsil metama’lumot bog‘langan. Rus milliy korpusida ham har bir matn uchun **bibliografik ma’lumotlar** berilgan: **asarning nomi, muallifi, yozilgan yili, janri, uslubi (badiiy, ilmiy va h.k.), matn turi (og‘zaki/yozma), hatto og‘zaki nutq bo‘lsa, gapiruvchining jinsi, yoshi, ijtimoiy holati** kabi qo‘shimcha ma’lumotlar kiritilgan. O‘zbek tilining milliy korpusi uchun B.Elov va boshq. (2023) tadqiqotlarida o‘zbek adabiy tilining besh uslubga mansub manbalari alohida jamlangani ta’kidlanadi (*so‘zlashuv uslubi, badiiy, ilmiy, rasmiy, publitsistik*) [12]. Bu shuni anglatadiki, har bir matnga qaysi uslubga tegishli ekanini ko‘rsatuvchi belgi qo‘yilgan. Kelgusida foydalanuvchi, masalan, faqat publitsistik uslubdagi matnlardan misollar izlashni xohlasa, *“janr: publitsistika”* filtrini tanlab qidirishi mumkin bo‘ladi.

Metama’lumotlarning yana bir foydasi – **korpus tarkibini tahlil qilish** va **taqqoslash** imkonini berishi. Masalan, korpusdan avtomatik ravishda *“badiiy adabiyot”* va *“ilmiy uslub”* bo‘yicha ikki alohida kichik korpus ajratib olib, ularning lug‘at tarkibini taqqoslash mumkin. Yoki vaqt bo‘yicha dinamik o‘zgarishni kuzatish uchun korpusni yillarga ajratish mumkin: masalan, COCA korpusida matnlar yillar va besh yilliklar bo‘yicha belgilangan, shuning uchun foydalanuvchi so‘zning 1990-yillardagi va 2010-yillardagi tezligidagi farqni oson ko‘ra oladi (grafigi orqali). Korpus interfeyslari metama’lumotlar bilan ishlashni intuitiv oson qilishga intiladi – misol uchun, English-Corpora (COCA) tizimida *“Virtual Corpus”* degan funksiya mavjud bo‘lib, foydalanuvchi bir necha soniya ichida tanlagan metama’lumotlar asosida shaxsiy kichik korpus tuzishi va unga tegishli so‘zlar statistikasini ajratib olishi mumkin. Xuddi shunday, Sketch Engine dasturida ham *“Subcorpus”* funksiyasi bor: foydalanuvchi masalan, muallifi Oybek bo‘lgan matnlar to‘plamini yoki 2018 yildan keyin chiqqan internet maqolalari to‘plamini ajratishi mumkin – buning uchun muallif va yil metama’lumotlarini filtrlaydi xolos, tizim uning uchun alohida korpus shakllantirib beradi.

Umuman olganda, **metama’lumotlar** korpusning *tasnifiy axboroti* bo‘lib, u korpus tuzilishini yanada boy qiladi. Chunki oddiy matnlar to‘plamidan farqli o‘laroq, korpusda matnlarning kelib chiqishi va xususiyatlari to‘g‘risidagi ma’lumotlar ham bazaga kiritiladi. Bu ma’lumotlar ilmiy tadqiqotlarda masalan, *stilistika, sotsiolingvistika, diatopik tilshunoslik (lahja va shevalar bo‘yicha)* tadqiqotlarni o‘tkazishga imkon yaratadi. Korpus tuzishda shu sababli metama’lumotlarni to‘plash va to‘g‘ri formatda kiritish alohida bosqich sifatida qaraladi.

Statistik axborot va foydalanuvchi ma’lumotlari

Statistik axborot – korpusdan foydalanuvchi olishi mumkin bo‘lgan turli tahliliy ma’lumotlar. Bular korpusga oldindan kiritilib qo‘yilgan emas, balki foydalanuvchi so‘roviga binoan **hisoblab chiqariladigan ko‘rsatkichlardir**. Ular qatoriga yuqorida tilga olingan **so‘zlar chastotasi, kollokatsion lug‘atlar, ngramlar** kiradi. Bundan tashqari, foydalanuvchi korpusdan o‘zi istalgan ikki holatni solishtirish natijasida statistik xulosa ham olishi mumkin – masalan, *“A matnlar korpusida so‘z X chastotasi B korpusidagidan 2 barobar ko‘p”* degan kabi ma’lumot.

Chastotali lug‘atlar – korpus asosida tuzilgan so‘zlarning ro‘yxatidir. Korpus hajmi katta bo‘lgani tufayli bunday ro‘yxatlar tilning eng keng qo‘llaniladigan so‘zlarini aniqlashda muhim manba bo‘ladi. Masalan, BNC korpusi asosida tuzilgan BNC1000 degan ro‘yxatda ingliz tilidagi eng ko‘p ishlatiladigan 1000 ta so‘z keltirilgan va u til o‘rganuvchilar uchun qo‘llanma sifatida mashhur. Korpusdan chastotani olish odatda dasturiy ravishda “so‘zlar ustida oddiy sanaq” ishidir, lekin katta hajmda bu oson emas – masalan, 1 milliard so‘zli korpusda turli so‘z shakllari soni (types) yuz minglarni tashkil qiladi. Shunga qaramay, Leipzig korpus to‘plami kabi loyihalar dunyoning 20 dan ortiq tili uchun avtomatik ravishda 1 million so‘zdan iborat korpuslar tuzib, ularning **chastotali lug‘atlarini** ham e’lon qilgan [12]. Ularning maqsadi resursi kam tillar uchun ham asosiy korpus statistikalarini ochiq taqdim etishdir. O‘zbek tilida ham ilk bor shunday chastotali lug‘at B.Elov tomonidan tuzilgan bo‘lib, https://uzschoolcorpara.uz/ uchun jami so‘zlar chastotasi hisoblab chiqilgan (natijada o‘zbek tilida eng ko‘p qo‘llaniladigan so‘z **“va”** ekanligi kabi qiziqarli faktlar aniqlandi). Chastota axboroti lingvistika tadqiqotlarida muhim: masalan, **Zipf qonunini tekshirish**, **so‘z boyligini baholash**, **leksikografik baza tuzish** kabi ishlarda korpusdan olingan chastotalar asos bo‘ladi.

Kollokatsiyalar va n-gramlar – bular korpus matnlaridan hosil qilinadigan **birikmalar statistikasidir**. Kollokatsiya ko‘pincha ikki yoki undan ortiq so‘zning bir-biriga yaqin yoki bir gap ichida kutilganidan ko‘ra ko‘proq birga uchrashini bildiradi. Korpus buni aniqlash uchun eng yaxshi manbadir: masalan, korpus tahlili shuni ko‘rsatishi mumkinki, *“ijtimoiy”* so‘zi eng ko‘p *“tarmoqlar”* so‘zi bilan yonma-yon qo‘llanilar ekan (natijada “ijtimoiy tarmoqlar” birikmasi kollokatsion bog‘lanish kuchiga ega fraza deyish mumkin). Bunday kollokatsiyalarni korpusdan chiqarib olish uchun dastur har bir so‘z uchun unga eng ko‘p yaqin turgan so‘zlarni sanaydi va **MI o‘lchovi** orqali umumiy uchrashuv ehtimoli bilan solishtiradi. Natijada esa foydalanuvchiga, masalan, tanlangan **X so‘zining 20 ta eng yaqin kollokati ro‘yxati** va **ularning MI skorlari** taklif etiladi. Ushbu axborot tilshunoslikda **valentlik va frazeologizmlarni** o‘rganishda, avtomatik tarjimada sinonimik birikmalarni tanlashda va boshqa sohalarda qo‘l keladi.

N-grammalar esa kengroq tushuncha bo‘lib, u korpusdagi eng ko`p uchrayotgan N uzunlikdagi so‘z zanjirlarini ifodalaydi. Masalan, **2 so‘zli zanjirlar – bigrammalar**; **3 so‘zli – trigrammalar** va hokazo. Korpusdan avtomatik ravishda chastotasi yuqori n-grammalarni chiqarish orqali tilga xos iboralar aniqlanadi. Masalan, ingliz tilida *“at the end of the day”* 5 so‘zli birikmasi juda ko‘p qo‘llaniladi – buni katta korpuslarsiz aniqlash qiyin, lekin korpus statistikasi darhol ko‘rsatadi. O‘zbek tilida korpus yordamida *“degan edi”, “qayd etib o‘tdi”, “shuni ham ta’kidlash joiz”* kabi og‘zaki nutq va rasmiy uslubda keng uchraydigan birikmalarni aniqlash mumkin.

Foydalanuvchi statistikasi. Ayrim korpus tizimlari foydalanuvchilarning qaysi so‘zlarni ko‘p qidirayotgani, eng ko‘p so‘rovlar qaysi tillar uchun berilayotgani kabi meta-statistik ma’lumotlarni ham yuritadi. Bu, albatta, korpusning ichki til ma’lumotidan tashqari, foydalanishiga oid axborotdir. Masalan, milliy korpus administratorlari qaysi davrda qidiruvlar soni oshganini, ko‘proq qaysi turkumdagi so‘zlar izlanayotganini tahlil qilib borishi mumkin. Bu tur ma’lumotlar korpusni yanada takomillashtirish (masalan, eng ko‘p so‘ralayotgan funksiyani qo‘shish) uchun xizmat qiladi.

O‘zbek va jahon korpuslarini qurish tajribasi

Yuqorida keltirilgan komponent va axborot turlari zamonaviy korpus tuzishning **umumiy tamoyillaridir**. Korpus tuzishda maqsad tilning imkon qadar to‘liq va balansli ko‘rinishini yaratish, uni ishonchli lingvistik belgilash bilan ta’minlash va foydalanuvchiga qulay foydalanish imkoniyatini berishdir. Jahonda milliy korpuslar soni yildan yilga ortib bormoqda: ma’lumotlarga ko‘ra, hatto 1990-yilga kelib dunyo tillari uchun 600 ga yaqin korpus mavjud bo‘lgan, hozirda esa minglab korpus loyihalari borligi taxmin qilinadi (bunga turli ixtisoslashgan korpuslar ham kiradi) [13]. Korpuslarning hajmi ham oshib bormoqda – masalan, ingliz tilining Amerikaner korpusi COCA yiliga 20 million so‘zga kengayib, hozirda 1 milliarddan oshdi, Britaniya milliy korpusi esa hali ham 100 million so‘z hajmida **“statik”** holda qolmoqda. Bu shuni ko‘rsatadiki, **monitor korpuslar (doimiy yangilanib boruvchi)** va **statik korpuslar** farqlanadi. COCA singari monitor korpus zamonaviy til o‘zgarishlarini muntazam aks ettirib boradi, BNC kabi statik korpus esa muayyan davrning til hosilasi sifatida tarixiy ahamiyatga ega bo‘ladi. O‘zbek tilida ham istiqbolda monitor korpus yaratish masalasi ko‘ndalang turibdi – chunki tilimiz juda tez yangi so‘zlar bilan boyib, ommaviy axborot til uslubi o‘zgarib bormoqda.

Milliy korpus loyihasi (UzbCorpus.uz). O‘zbek tilining birinchi elektron korpusi sifatida UzbCorpus.uz platformasi 2021 yilda ishga tushirildi. Bu korpus O‘zbekistonda davlat miqyosidagi ilmiy loyiha doirasida boshlangan bo‘lib, dastlab badiiy asarlar matnlarini jamlashdan boshladi. Keyinchalik, B.Elov boshchiligida

ToshDO‘TAU tadqiqotchilar jamoasi ushbu korpusni rivojlantirish ishlari davom ettirildi – unga publitsistik va ilmiy matnlar qo‘shildi, hajmi oshirildi. Hozirda **UzbCorpus.uz** tarkibida 10 milliondan ortiq so‘z mavjud (aniq statistik ma’lumotlar e’lon qilinmagan, lekin ilmiy maqolalarda shu atrofida ko‘rsatilgan). Korpus tarkibi besh funksional uslubni qamrab olgan matnlardan iborat ekani ilmiy adabiyotda qayd etilgan. Korpusning interfeysi uch tilda (o‘zbek, ingliz, rus) taqdim qilingan bo‘lib, unda token va lemma bo‘yicha qidiruv, konkordans natijalarini ko‘rish imkoniyati bor. Mazkur korpus haligacha rivojlanishda davom etmoqda. Bu shuni ko‘rsatadiki, o‘zbek korpusini yanada multimediya (matn + audio) shaklida boyitish rejalangan, bu esa jahon korpusshunosligidagi zamonaviy tendensiyalarga mos keladi (masalan, rus tilida ham audio va sheva korpuslari alohida subkorpuslar sifatida kiritilganligi ma’lum).

Jahon korpuslari bilan qiyos. O‘zbek tilini korpuslash tajribasini jahon yutuqlari bilan qiyoslash foydali. Britaniya va Amerika ingliz tilining korpuslari (BNC va COCA) o‘zbek korpusiga o‘rnak bo‘la oladigan bir nechta xususiyatga ega: ularda matnlar juda keng ko‘lamda jamlangan va aniq balanslangan; har bir so‘z to‘liq morfologik tegangan (BNC uchun CLAWS POS-teglar, COCA uchun Penn Treebank tagset asosida); foydalanuvchi interfeysi esa nihoyatda boy funksionalga ega (lemma bo‘yicha izlash, sinonim bo‘yicha izlash, hatto WordNet ma’lumotlariga bog‘langan holda ma’nodoshlik bo‘yicha qidiruv va h.k. – COCA misolida). Rus milliy korpusi esa o‘zbek tiliga ko‘proq strukturaviy jihatdan yaqin bo‘lgani uchun qiziqarli tajriba beradi: unda agglyutinativ xususiyatlar bo‘lmasa-da, boy flektiv morfologiyaga ega rus tilini teglashda katta mehnat qilingan va ayni paytda ushbu korpus ochiq tarzda internetda foydalanishga imkon beradi. RNC tarkibida, yuqorida aytilganidek, sintaktik daraxtlar subkorpusi, she’riy matnlar subkorpusi, parallel korpuslar kabi turli komponentlar borligi e’tiborga loyiq – o‘zbek tilida ham kelgusida shunga o‘xshash parallel korpuslar (masalan, o‘zbek-ingliz, o‘zbek-rus tarjima korpusi) hamda ixtisoslashgan soha matnlari korpuslarini yaratish rejasi bor.

Yana bir jahon loyihasi – **Leipzig Corpora Collection (LCC)** – o‘zbek tili korpusini rivojlantirishda foydali bo‘lishi mumkin. LCC doirasida 250 dan ortiq til uchun bir xil formatda korpus ma’lumotlari va ularning asosiy statistik ko‘rsatkichlari (wordlist, bigram list va hokazo) bepul tarqatiladi. Masalan, LCCda o‘zbek tilidan ham 1 million so‘zli matnlar to‘plami va uning so‘z chastotasi ro‘yxati mavjud. Bu kichik hajmli bo‘lsa-da, qiyosiy tadqiqotlarda as qotadi. Kelgusida o‘zbek milliy korpusini ham Leipzig kabi ochiq formatda, xalqaro standartlarga mos holda eksport qilish va xalqaro korpuslar qatoriga qo‘shish maqsad qilinmoqda – bu o‘zbek tilining dunyo miqyosida o‘rganilishiga, turkiy tillarni solishtirishda ishtirok etishiga xizmat qiladi.

Xulosa. Xulosa o‘rnida, korpus tuzilishi va axborot turlari bo‘yicha yuqoridagi tahlillar shuni ko‘rsatadiki, zamonaviy elektron korpus – bu ko‘p qavatli

murakkab tizimdir. Unda matnlar lingvistik jihatdan boy belgilangan, texnik bazasi kuchli arxitekturaga tayangan va foydalanuvchi uchun keng qidiruv hamda tahlil imkoniyatlarini taqdim etadi. Optimal korpus yaratish uchun ushbu uch asos – **matn bazasi, texnik arxitektura va interfeys –** birdek mustahkam bo‘lishi lozim [14]. Shu bilan birga, korpus bilan bog‘liq **barcha axborot turlari** – lingvistik, metama’lumot va statistik – to‘liq va izchil ravishda o‘zaro bog‘langan holda ishlaydi. O‘zbekiston tilshunoslari tomonidan boshlangan milliy korpus loyihasi ham aynan shu tamoyillarga intilayotgani ko‘zga tashlanadi.

B.Elov, N.Abdurahmonova kabi olimlar korpus lingvistikasi sohasida dastlabki ishlarni amalga oshirib, o‘zbek tilini raqamli korpuslash yo‘lida muhim qadamlar qo‘yishgan – bu ishlarda jahon korpus tajribasiga tayanilgan va milliy til xususiyatlariga mos yechimlar taklif qilingan. Kelayotgan yillarda o‘zbek tili korpusi yanada kengayib, to‘liq belgilangan, milliardlab so‘zli zamonaviy korpusga aylanishi va jahon korpuslaridan kam bo‘lmagan ilmiy ahamiyat kasb etishi kutilmoqda. Bu borada olib borilayotgan ishlar – lingvistik teglashning avtomatlashtirilishi, arxitekturaning optimallashtirilishi, yangi matnlar bilan boyitish – korpus tuzilishini takomillashtirib, foydalanuvchilar va tadqiqotchilar uchun yanada qulay muhit yaratib beradi. Korpus tuzilishi va undagi axborot qatlamlari qanchalik boy bo‘lsa, tilshunoslikda shunchalik chuqur va ishonchli tahlillar qilish imkoni tug‘iladi. Bu esa ona tilimizni ilmiy jihatdan teranroq anglash va kelgusida turli sun’iy intellekt tizimlarida qo‘llash uchun mustahkam asos yaratadi.

Foydalanilgan adabiyotlar:

11. Primova, M. (2023). Til korpuslarida matnlarni annotatsiyalash: afzallik va kamchiliklari. *Uzbekistan: Language and Culture*, *4*(4).
12. Копотев, М. В., & Янда, Л. (2006). Национальный корпус русского языка. *Вопросы языкознания*, (5), 149-155.
13. Avgustinova, T., & Zhang, Y. (2009, September). Exploiting the Russian national corpus in the development of a Russian Resource Grammar. In *Proceedings of the workshop on adaptation of language resources and technology to new domains* (pp. 1-11).
14. Elov, B. B., Khamroeva, S. M., Alayev, R. H., Khusainova, Z. Y., & Yodgorov, U. S. (2023). Methods of processing the uzbek language corpus texts. *International Journal of Open Information Technologies*, *11*(12), 143-151.
15. Suhr, C., Nevalainen, T., & Taavitsainen, I. (Eds.). (2019). *From data to evidence in English language research*. Leiden: Brill.
16. Zeldes, A. (2021). Corpus architecture. In *A practical handbook of corpus linguistics* (pp. 49-73). Cham: Springer International Publishing.

17. Bolgün, M. A. (2013). The significance of data-driven descriptions of forms in explicit grammar instruction. *Procedia-Social and Behavioral Sciences*, *70*, 430-446.
18. https://tsuull.uz/sites/default/files/b_elov_a_abdullayev_n_xudaybergano v_ozbek_tili_korpusi_matnlari_0.pdf
19. Elov, B., Alayev, R., & Abdullayev, A. (2024). Tabiiy tilning statistik modellari. *digital transformation and artificial intelligence*, *2*(6), 178-189.
20. https://www.academia.edu/124652038/Korpus_lingvistikasi
21. Elov, B., & Xudayberganov, N. (2024). O‘zbek tili korpusi matnlarini pos teglash usullari. *Computer Linguistics: problems, solutions, prospects*, *1*(1).
22. Elov, B., Hamroyeva, S., Alayev, R., Xusainova, Z., & Yodgorov, U. (2023). O ‘zbek tili korpusi matnlarini qayta ishlash usullari. *Digital transformation and artificial intelligence*, *1*(3), 117-129.
23. Biemann, C., Heyer, G., Quasthoff, U., & Richter, M. (2007). The Leipzig corpora collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistic*, *2007*.
24. 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), B. 644-650.
25. Davies, M. (2015). Corpora: an introduction. *The Cambridge handbook of English corpus linguistics*, 11-31.