O‘ZBEK TILI KORPUSINI YARATISH VA UNI YURITISHNING AHAMIYATI

Yo‘ldashaliyeva Rayyona
Andijon davlat chet tillari instituti talabasi
https://doi.org/10.5281/zenodo.24218711

Annotatsiya: Mazkur maqolada o‘zbek tili korpusini yaratish va uni muntazam yuritishning ilmiy, ta’limiy, texnologik hamda ijtimoiy-siyosiy ahamiyati tahlil qilinadi. Korpus lingvistikasining nazariy asoslari, jahon tajribasi va o‘zbek tilshunosligidagi mavjud korpuslar (milliy, ta’limiy, dialektal, parallel) qiyosiy yoritiladi. Shuningdek, o‘zbek tilining agglutinativ xususiyati, ikki yozuv tizimining (kirill va lotin) mavjudligi hamda teglangan ochiq ma’lumotlar yetishmasligi bilan bog‘liq muammolar ko‘rsatiladi. Korpusni yuritish bo‘yicha amaliy tamoyillar (yangilash, sifat nazorati, yozuvlarni uyg‘unlashtirish, formatlarni standartlash, ochiqlik) asoslab beriladi. Maqola natijalari korpus lingvistikasi, lug‘atchilik, til ta’limi va kompyuter lingvistikasi sohalarida tadqiqot olib borayotganlar uchun foydali bo‘lishi mumkin.
Kalit so‘zlar: korpus lingvistikasi, o‘zbek tili korpusi, milliy korpus, ta’limiy korpus, dialektal korpus, teglash, kompyuter lingvistikasi, sun’iy intellekt, agglutinativ til.

Abstract: The article examines the scientific, educational, technological and socio-political significance of creating and maintaining a corpus of the Uzbek language. It reviews the theoretical foundations of corpus linguistics, international experience and existing Uzbek corpora, and discusses problems specific to Uzbek: its agglutinative morphology, the coexistence of Cyrillic and Latin scripts, and the shortage of openly available annotated data. Principles of corpus maintenance, including regular updating, quality control, script harmonisation, format standardisation and openness, are substantiated.
Keywords: corpus linguistics, Uzbek corpus, national corpus, educational corpus, annotation, computational linguistics, artificial intelligence.

Аннотация: В статье анализируется научная, образовательная, технологическая и общественно-политическая значимость создания и систематического ведения корпуса узбекского языка. Представлен сравнительный обзор теоретических основ корпусной лингвистики, международного опыта и существующих в узбекском языкознании корпусов (национального, учебного, диалектного, параллельных). Кроме того, рассматриваются проблемы, обусловленные агглютинативным строем узбекского языка, сосуществованием двух систем письма (кириллицы и латиницы), а также нехваткой размеченных открытых данных. Обосновываются практические принципы ведения корпусов: обновление, контроль качества, унификация графики, стандартизация форматов и обеспечение открытости. Результаты исследования могут представлять интерес для специалистов в области корпусной лингвистики, лексикографии, лингводидактики и компьютерной лингвистики.

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

Ключевые слова: корпусная лингвистика, корпус узбекского языка, национальный корпус, учебный корпус, диалектный корпус, разметка, компьютерная лингвистика, искусственный интеллект, агглютинативный язык.

Kirish

XXI asrda til taraqqiyoti ko‘p jihatdan uning raqamli muhitdagi o‘rni bilan belgilanmoqda. Tilning elektron resurslari, qidiruv tizimlari, mashina tarjimasi va sun’iy intellekt vositalarida qanchalik to‘liq aks etishi uning kelajakdagi nufuziga bevosita ta’sir ko‘rsatadi. Bu jarayonda tilning haqiqiy qo‘llanishini aks ettiruvchi, kompyuterda qayta ishlash mumkin bo‘lgan matnlar to‘plami, ya’ni korpus, markaziy o‘rin tutadi.

O‘zbek tiliga davlat tili maqomi 1989-yil 21-oktabrda berilgan [4]. Shundan buyon tilga e’tibor ortib borayotgan bo‘lsa-da, uning elektron resurslari rivojlangan tillarga nisbatan sezilarli darajada kam. Mutaxassislar o‘zbek tilining ham milliy korpusi yaratilishi lozimligini va bu sohada milliy kompyuter lingvistikasi tizimli rivojlanishi kerakligini ta’kidlab kelmoqda [14]. 2025-yil 21-oktabrda, davlat tili maqomiga 36 yil to‘lgan kunda, davlat rahbari sun’iy intellekt o‘zbek tilini mukammal darajada tushunishi zarurligini aytdi va Raqamli texnologiyalar vazirligiga “O‘zbek tili korpusi”ni yaratish vazifasi topshirildi [15]. Bu masalaning ilmiy bo‘lib qolmay, davlat darajasidagi ustuvor vazifaga aylanganini ko‘rsatadi.

Maqolaning maqsadi o‘zbek tili korpusini yaratish va uni yuritishning ahamiyatini nazariy hamda amaliy jihatdan asoslash, mavjud muammolarni aniqlash va ularni hal etish yo‘llarini ko‘rsatishdir. Tadqiqotda tahliliy, qiyosiy va tavsifiy metodlardan foydalanildi; manba sifatida ilmiy maqolalar, konferensiya materiallari va ochiq elektron resurslar o‘rganildi.

1. Korpus tushunchasi va turlari

Korpus lingvistikasi nazariyasida korpus deganda muayyan til yoki uning bir qismini namoyish etuvchi, kompyuterda qayta ishlashga yaroqli, ma’lum tamoyillar asosida tanlangan va odatda izohlangan (teglangan) matnlar to‘plami tushuniladi [1; 2]. Oddiy matnlar yig‘indisidan farqli o‘laroq, korpus vakillik (reprezentativlik), muvozanat va qidiruv imkoniyati kabi talablarga javob berishi kerak.

Korpuslar vazifasiga ko‘ra turlicha bo‘ladi. Milliy korpus ilmiy va rasmiy matnlar, og‘zaki va badiiy matnlar, matbuot hamda dialektlar kabi ichki korpuslarni o‘z ichiga olishi bilan ajralib turadi [3]. Ta’limiy korpus esa o‘quv jarayoni uchun saralangan matnlardan tashkil topadi va o‘quv lug‘atlari, darsliklar hamda test materiallarini tayyorlashda foydalaniladi [8; 9]. Bulardan tashqari muallif korpusi, poetik matnlar korpusi, parallel (ikki yoki undan ortiq tildagi), diaxronik va dialektal korpuslar mavjud [3; 5; 17].

2. Jahon tajribasi

Korpus lingvistikasining zamonaviy bosqichi Braun korpusi bilan bog‘lanadi. AQSh bosma nasridan olingan taxminan million so‘zlik bu korpus keyinchalik yaratilgan boshqa korpuslar uchun o‘ziga xos andoza bo‘lib xizmat qilgan [4]. Keyinchalik Britaniya milliy korpusi kabi yirik loyihalar paydo bo‘lgan [10].

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

Hajm jihatidan tillar o‘rtasidagi tafovut katta. O‘zbek tilshunosligidagi ayrim ishlarda ingliz tili milliy korpusi 2,4 milliarddan ortiq, rus tilining milliy korpusi taxminan 200 million, turk tilining milliy korpusi esa 50 milliondan ortiq so‘z va so‘z shakli hajmiga ega ekani qayd etiladi [3]. Bu ko‘rsatkichlar aniq loyiha va yilga qarab o‘zgarishi mumkinligi sababli, ularni tadqiqotda keltirishda manba ko‘rsatilishi lozim.

Shuni ham ta’kidlash kerakki, nafaqat yirik jahon tillari, balki ko‘plab o‘rta va kichik hajmli tillar uchun ham milliy korpuslar yaratilgan, hatto qadimgi tillarning korpuslari ham ishlab chiqilgan [14]. Demak, korpusga ega bo‘lish tilning kattaligiga emas, uni rivojlantirishga bo‘lgan ehtiyojga bog‘liq.

3. O‘zbek tilshunosligida korpus yaratish holati

O‘zbek tili uchun milliy korpus yaratish bo‘yicha tadqiqotlar A. Po‘latov, B. Mengliyev, N. Abdurahmonova kabi olimlar tomonidan boshlab berilgan [11; 14]. B. Mengliyev milliy til korpusini til birliklarining xususiyatlarini aniqlash maqsadida elektron qidiruv imkoniyatiga ega matnlar to‘plami sifatida ta’riflaydi [11].

Hozirgi kunda quyidagi yo‘nalishlarda ishlar olib borilmoqda:

Milliy va ta’limiy korpuslar. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universitetida o‘zbek tilining milliy, ta’limiy va elektron korpuslarini yaratish bo‘yicha ishlar amalga oshirilmoqda [5]. Ta’limiy korpus va uning imkoniyatlari bo‘yicha tadqiqotlar e’lon qilingan [8; 9]. Elektron ta’limiy resurslar sifatida uzbekcorpus.uz va uzschoolcorpara.uz platformalari qayd etiladi [9].

Parallel korpuslar. O‘zbek-turk tillari parallel korpusi (uzturkfraz.uz) yaratilgan bo‘lib, uning ma’lumotlar bazasini shakllantirish texnologiyasi tadqiq qilingan [9].

Veb-korpuslar. Sketch Engine platformasida o‘zbek tilining UzWaC korpusi mavjud [9].

Diaxronik korpus. Buxoro davlat universitetida o‘zbekcha mediamatnlar diaxronik korpusini yaratish orqali til rivojini kuzatish tajribasi o‘rganilmoqda; o‘zbek tilining diaxronik korpusini shakllantirish zarurati ilmiy asoslanmoqda [5].

Dialektal korpus. Dialektal birliklar korpusi milliy korpusning tarkibiy qismi sifatida qaraladi va dialekt so‘zlar bazasini elektronlashtirish o‘zbek tilining to‘laqonli korpusini yaratishga xizmat qiladi [17].

Matnlarni formatlash. Matnlarni korpusga kiritishda formatlarning xilma-xilligi va nomuvofiqligi umumiy format zarurligini keltirib chiqaradi. Bu borada JSON formatidan foydalanish qidiruv tezligini oshirishi va kengayuvchanlik muammolarini kamaytirishi mumkinligi ko‘rsatilgan; «Alpomish» dostoni matnlarini kiritish tajribasi tavsiflangan [6].

4. Korpusni yaratish va yuritishning ahamiyati

4.1. Ilmiy ahamiyati

Korpus til hodisalarini tasodifiy misollarga emas, real matnlar va ularning statistikasiga tayanib o‘rganish imkonini beradi. U leksikografiya, grammatika, so‘z birikmalari (kollokatsiyalar), so‘z ma’nosining kengayishi va torayishi, yangi iboralarning paydo bo‘lishini kuzatishda qo‘l keladi [3]. Korpus metodlari til evolyutsiyasi, diaxronik o‘zgarishlar, madaniyatlararo interferensiya va til ta’limidagi samaradorligi bilan ham asoslanadi [5]. O‘zbek lug‘atchiligida chastotali lug‘atlar va

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

dostonlar matnini tahlil qilish tajribalari mavjud bo‘lsa-da, ular har doim ham aynan korpus tahliliga asoslanmaydi [5]. Bu korpusga asoslangan lug‘atchilikni rivojlantirish ehtiyojini ko‘rsatadi.

4.2. Ta’limiy ahamiyati

Elektron korpus o‘quv lug‘atchilikni rivojlantirishda muhim resurs bo‘lib, yangi avlod lug‘atlarini kompyuter texnologiyalari yordamida yaratish imkonini beradi [7]. O‘qituvchi va o‘quvchi so‘zning real kontekstdagi qo‘llanishini ko‘ra oladi, o‘quv materiallari esa tilning haqiqiy chastotasiga asoslanib tanlanadi. Chet tillarni o‘qitish metodikasida ham korpus metodlari qo‘llanadi [5]; o‘zbek tili uchun parallel korpuslar tarjima va qiyosiy tadqiqotlarga xizmat qiladi.

4.3. Texnologik ahamiyati

Mashina tarjimasi, avtomatik teglash, nomli obyektlarni aniqlash va til modellarini o‘qitish katta hajmli va sifatli ma’lumotlarni talab qiladi. O‘zbek tili bu borada kam resursli til hisoblanadi: ochiq ma’lumotlar to‘plami kam, Vikipediyaning 2021-yil iyul nusxasida atigi 12,2 million so‘z bo‘lgan va maqolalar sifati ham bir xil emas [12]. Teglangan korpus yo‘qligi tabiiy tilni qayta ishlash masalalarini hal etishni qiyinlashtiradi [10]. Shunga qaramay, so‘nggi yillarda ijobiy natijalar ham bor: BERT asosidagi POS-teglash modellari o‘rtacha 91% aniqlikka erishgan [13]. Biroq shu tadqiqotda foydalanilgan katta hajmli teglangan ma’lumotlar to‘plami hamma uchun ochiq emasligi ham qayd etilgan [13]. Demak, ochiq va sifatli korpus bo‘lmasa, texnologik taraqqiyot cheklanib qoladi.

4.4. Ijtimoiy-siyosiy ahamiyati

Sun’iy intellekt o‘zbek madaniyati va qadriyatlarini to‘g‘ri aks ettirishi uchun uni o‘zbek tilidagi ishonchli matnlarda o‘qitish zarur. 2025-yilgi topshiriqda ham o‘zbek tilidagi adabiyot, maqola va hisobotlarning aksariyati raqamlashmagani, elektron ko‘rinishdagilari esa turli joylarda saqlanayotgani muammo sifatida ko‘rsatilgan [15]. Bu korpus faqat til resursi emas, milliy axborot infratuzilmasining bir qismi ekanini bildiradi.

5. Korpusni yuritishning zaruriyati va asosiy muammolar

Korpusni bir marta yaratib qo‘yish yetarli emas. Til doimiy o‘zgaradi: yangi so‘zlar, o‘zlashmalar, imlo va uslub o‘zgarishlari paydo bo‘ladi. Yangilanmagan korpus tilning bugungi holatini aks ettirmaydi va uning ilmiy hamda amaliy qiymati kamayadi. Shu sababli quyidagi yo‘nalishlar muhim:

1. Yozuvlar uyg‘unligi. O‘zbek tilida kirill va lotin yozuvi birga qo‘llanadi; kirill 1956-yilda joriy etilgan, lotin yozuvi 1995-yilda qayta joriy qilingan [11]. Yig‘ilgan matnlarning bir qismi kirillda, qolgani lotinda bo‘lgani uchun ularni bir yozuvga o‘tkazish kerak bo‘ladi [12]. Lotin yozuvidagi o‘ va g‘ harflaridagi modifikatsiya belgilari va tutuq belgisi ayrim modellarda noto‘g‘ri ajratilishi kuzatilgan, shuning uchun o‘zbek lotin yozuviga xos normallashtirish qoidalari zarur [13].

2. Morfologik murakkablik. O‘zbek tili agglutinativ til bo‘lib, so‘zlar ko‘plab qo‘shimchalar yordamida hosil bo‘ladi va o‘zak ajratish qiyinlashadi. Turkiy tillarning agglutinativligi mashina tarjimasida alohida muammo tug‘dirishi ko‘rsatilgan [12]. Shu bois morfologik teglash (masalan, FST texnologiyasi asosida) korpusning ajralmas qismi bo‘lishi kerak [10; 16].

3. Teglash tizimi va standartlar. Teglash uchun yagona tegset va annotatsiya tamoyillari ishlab chiqilishi kerak. Mavjud tegsetlarda so‘z turkumlariga deyarli bittadan teg berilgani, ammo har bir turkum o‘zbek tilida ichki bo‘linishga ega ekani tanqid qilingan [10].

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

4. Formatlar va texnik moslik. Matnlarni yagona formatda saqlash qidiruv tezligi va kengayuvchanlikni ta’minlaydi [6].

5. Sifat nazorati. OCR xatolari, takrorlar va noto‘g‘ri bo‘linishlar ma’lumot sifatini pasaytiradi. Ayrim modellar o‘qitilgan matnlar skanerlangan kitoblardagi xatolar tufayli shovqinli ekani qayd etilgan [13].

6. Vakillik va muvozanat. Korpus janrlar (badiiy, ilmiy, matbuot, rasmiy, og‘zaki nutq), davrlar va hududlar bo‘yicha muvozanatli bo‘lishi kerak. Faqat internet materiallariga tayanish tanlovni cheklaydi [5].

7. Ochiqlik va huquqiy jihatlar. Ilmiy hamjamiyat foydalana oladigan ochiq resurslar yaratish, shu bilan birga mualliflik huquqini hurmat qilish muhim. Turkiy tillar uchun ma’lumot sifati va tadqiqot imkoniyatlari cheklanganligi umumiy muammo sifatida e’tirof etiladi [12].

8. Ilmiy sifat. O‘zbek NLP bo‘yicha chop etilgan ishlar soni ko‘p bo‘lsa-da, resurs va vositalar sifati har doim ham shu miqdorga mos kelmasligi ta’kidlangan [10]. Shuning uchun tayyor resurslar qayta foydalanish mumkin bo‘lgan, hujjatlashtirilgan va baholangan bo‘lishi kerak.

6. Takliflar

Tahlil asosida quyidagi takliflarni bildirish mumkin: milliy korpusni yagona platformada jamlash va uni davlat idoralari, universitetlar va kutubxonalar ma’lumotlari bilan boyitib borish;

kirill va lotin yozuvlari uchun yagona transliteratsiya va normallashtirish standartini joriy etish;

morfologik va sintaktik teglangan, ochiq litsenziyali kichik, lekin sifatli “oltin standart” korpuslar yaratish;

diaxronik, dialektal, og‘zaki nutq va parallel korpuslarni milliy korpusning subkorpuslari sifatida rivojlantirish;

korpusni muntazam yangilash tartibi (davriylik, mas’ul tashkilot, versiyalash) va sifat mezonlarini belgilash;

ta’lim jarayonida korpusdan foydalanish ko‘nikmalarini o‘qitish (o‘qituvchi va talabalar uchun qo‘llanmalar).

Xulosa

O‘zbek tili korpusi tilshunoslik tadqiqotlari, lug‘atchilik, ta’lim va sun’iy intellekt uchun asosiy resurs hisoblanadi. Mavjud ishlar (milliy, ta’limiy, parallel, dialektal va diaxronik korpuslar) muhim poydevor yaratgan bo‘lsa-da, teglangan ochiq ma’lumotlar yetishmasligi, ikki yozuv tizimi, morfologik murakkablik va yagona standartlar yo‘qligi kabi muammolar saqlanib qolmoqda. Korpusni yaratish bilan birga uni muntazam yuritish, yangilash, sifat nazoratidan o‘tkazish va ochiqligini ta’minlash uning uzoq muddatli qiymatini belgilaydi. 2025-yilda davlat darajasida qo‘yilgan vazifa bu sohadagi ishlarni jadallashtirish uchun imkoniyat yaratadi. Kelgusidagi tadqiqotlar korpus hajmi, tarkibi va ochiqligini aniq ko‘rsatkichlar asosida baholashga qaratilishi maqsadga muvofiq.

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.

Foydalanilgan adabiyotlar

1. **Po‘latov A.** Kompyuter lingvistikasi: O‘quv qo‘llanma. – Toshkent: Akademnashr, 2011. – 520 b.
2. **Mengliyev B.** O‘zbek tili milliy korpusi: muammo va yechimlar // O‘zbek tili va adabiyoti. – Toshkent, 2019. – № 3. – B. 3–14.
3. **Abdurahmonova N.** Korpus lingvistikasi asoslari. – Toshkent: Universitet, 2020. – 160 b.
4. **O‘zbekiston Respublikasining “Davlat tili haqida”gi Qonuni** (yangi tahriri). – Toshkent: O‘zbekiston, 1995.
5. **Yo‘ldoshev B.** O‘zbek leksikografiyasi va korpus lingvistikasi. – Samarqand: SamDU nashri, 2021. – 192 b.
6. **Mamatov A.** O‘zbek tilining izohli lug‘ati va kompyuter bazasi muammolari // Tilshunoslik va o‘zbek tili masalalari. – Toshkent, 2018. – B. 45–52.
7. **Nurmonov A.** Tilshunoslik nazariyasi. – Toshkent: Fan, 2002. – 180 b.
8. **Qodirov B.** Ta’limiy korpuslar va ularning lingvodidaktik xususiyatlari // Ta’lim va fan. – Toshkent, 2022. – № 2. – B. 88–94.

© Ushbu material Creative Commons Attribution 4.0 International (CC BY 4.0) litsenziyasi asosida tarqatiladi.