TIL KORPUSI LINGVISTIKASI: SHAKLLANISH TARIXI, RIVOJLANISH BOSQICHLARI VA ZAMONAVIY TENDENSIYALAR (O‘ZBEK TILI MISOLIDA)

Alimova Ma’mura
Andijon davlat chet tillari instituti stajor tadqiqotchisi
DOI:https://doi.org/10.5281/zenodo.23275714

An’anaviy tilshunoslik ko‘pincha tilni olimning shaxsiy intuitsiyasi, subyektiv qarashlari yoki cheklangan nazariy “ideal gapiruvchi” modeli orqali o‘rganish bilan cheklanib qolgan bo‘lsa, zamonaviy korpus lingvistikasi to‘liq empirik, ya’ni mutlaqo dalillarga va real faktlarga asoslangan ilmiy yondashuvga tayanadi va bu yo‘nalish kundalik hayotda, badiiy adabiyotda, ommaviy axborot vositalarida hamda turli funksional uslub va vaziyatlarda bevosita ishlatilgan minglab va millionlab so‘zlardan iborat tabiiy matnlar elektron bazasini chuqur tahlil qilish orqali tilning jonli qonuniyatlari, ichki rivojlanish tendensiyalari va amaliy qo‘llanilish xususiyatlarini ochib beradi.

Ilmiy jihatdan puxta va talablarga to‘liq javob beradigan matn korpusi oddiygina tartibsiz matnlar yig‘indisi emas, balki qat'iy ilmiy-metodologik mezonlar asosida puxta shakllantirilgan murakkab tizim bo‘lib, unga badiiy, ilmiy-texnik, publistik, rasmiy-ish yuritish matnlari hamda jonli og‘zaki so‘zlashuv nutqi namunalari muvozanatli tarzda qamrab olinadi, shuningdek, matnlardagi har bir leksema va so‘z shaklining turkumiy mansubligi, grammatik kategoriyalari, morfologik xususiyatlari va sintaktik vazifasi maxsus teglash tizimi orqali chuqur annotatsiya qilinib, tilning barcha lahja, sheva va zamonaviy qatlamlarini to‘laqonli aks ettiruvchi yetarli hajm hamda yuqori vakillik darajasi ta'minlanadi.

O‘zbek tili kabi boy tarixiy ildizlarga ega bo‘lgan, biroq elektron resurs bazasi va raqamli infratuzilmasi jadal shakllanish bosqichida turgan tillar uchun sifatli milliy korpuslarning yaratilishi o‘ta muhim strategik ahamiyat kasb etib, u sun'iy intellekt texnologiyalari, neyron tarmoqlarga asoslangan mashina tarjimasi tizimlari, ovozli ma'lumotlarni matnga o‘giruvchi nutqni tanish dasturlari va intellektual avtomatik matn tahrirchilarini yaratishda tabiiy tilni qayta ishlash jarayonining asosiy poydevori bo‘lib xizmat qiladi, tilshunoslarga so‘zlarning haqiqiy qo‘llanish chastotasi, semantik valentligi, uslubiy bo‘yoqdorligi va sintaktik birikuvchanligini matematik aniqlikda belgilash imkonini beradi hamda o‘zbek tilini ona tili yoki xorijiy til sifatida o‘qitish jarayonida real nutq namunalariga tayanadigan zamonaviy o‘quv qo‘llanmalari, lug‘atlar va multimedia darsliklarini yaratishga mustahkam zamin yaratadi.

Jahon miqyosida o‘zining yuqori samaradorligini amalda to‘la ko‘rsatgan ilg‘or xalqaro korpus lingvistikasi tajribalarini o‘zbek tilining o‘ziga xos grammatik tuzilishi, agglyutinativ tabiati va leksik xususiyatlaridan kelib chiqqan holda ijodiy moslashtirish bugungi kunda o‘ta dolzarb vazifa bo‘lib qolmoqda, chunki bunda milliy o‘ziga xoslikni va tilning boy imkoniyatlarini to‘liq saqlagan holda O‘zbek milliy korpusining semantik-sintaktik imkoniyatlarini, qidiruv mexanizmlari va ko‘p funksiyali tahliliy bazasini kengaytirish ilmiy-amaliy tadqiqotlar samaradorligini keskin oshiribgina qolmay, balki tilimizning jahon axborot makonidagi raqobatbardoshligini ham ta'minlaydi.

Korpus lingvistikasining zamonaviy bosqichi asosan o‘tgan asrning o‘rtalariga, aniqrog‘i 1960-yillarda AQShda dunyodagi ilk elektron matn bazasi sanalgan Brown Corpus (1961–1964) yaratilishi bilan boshlangan bo‘lib, bu ilmiy yutuq tilshunoslikda eksperimental va statistik tadqiqotlarga yo‘l ochdi, keyinchalik esa bu an’ana Buyuk Britaniyada LOB korpusi, jahon miqyosida esa tilshunoslikning oltin standartiga aylangan British National Corpus (BNC, 1994) hamda bugungi kunda AQShda millionlab so‘zlarni qamrab olgan ulkan COCA (Corpus of Contemporary American English) kabi yirik va ko‘p

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

qirrali korpuslarning paydo bo‘lishi bilan izchil davom etdi va nihoyat 2010-yillardan boshlab raqamli texnologiyalarning shiddatli rivojlanishi tufayli zamonaviy sun’iy intellekt, chuqur neyron tarmoqlar, katta til modellari (LLM) hamda turli og‘zaki va vizual ma’lumotlarni o‘z ichiga olgan multimodal korpuslar davri boshlandiki, bu tub burilishlar tabiatan o‘zbek tilshunosligiga ham o‘zining sezilarli ta’sirini ko‘rsatib, milliy tilimizni raqamli muhitda tadqiq etishga bo‘lgan yondashuvlarni butunlay yangi bosqichga ko‘tardi.

O‘zbek tilida korpus lingvistikasining shakllanishi va rivojlanish jarayoni asosan 2010-yillardan boshlab faol tus ola boshladi va bugungi kunda bu soha bir nechta muhim yo‘nalishlarda bosqichma-bosqich rivojlanmoqda: birinchidan, o‘zbek tilining milliy xususiyatlarini aks ettiruvchi turli hajmdagi milliy korpusi loyihalari amalga oshirilmoqda; ikkinchidan, o‘quv jarayoni va metodik ta’minotga mo‘ljallangan ta’limiy korpuslar yaratilmoqda; uchinchidan, mumtoz adabiyotimiz durdonalari, xususan, mutafakkir shoir Alisher Navoiy asarlari tili va leksik boyligini chuqur o‘rganishga qaratilgan maxsus mualliflik korpuslari shakllantirilmoqda; to‘rtinchidan, tarjima nazariyasi va amaliyotini takomillashtirishga xizmat qiluvchi o‘zbek-ingliz va o‘zbek-rus tillararo parallel korpuslari faoliyat yuritmoqda; beshinchidan esa, zamonaviy internet madaniyati va jonli til jarayonlarini kuzatish imkonini beruvchi ijtimoiy tarmoqlar matnlari asosidagi norasmiy va dinamik korpuslar maydonga kelmoqda.

Shunga qaramay, o‘zbek tilidagi mavjud korpuslar hozirgi kunda ham umumiy hajmi, leksik va statistik qamrovi, lingvistik annotatsiya sifati hamda janrlar xilma-xilligi jihatidan jahon korpus lingvistikasining ilg‘or standartlaridan ma’lum darajada orqada qolmoqda, xususan, matnlardagi so‘zlarning morfologik va sintaktik jihatdan to‘liq avtomatlashtirilmagani, ya’ni sun’iy intellekt yordamida teglash tizimining yetarlicha takomillashmagani bugungi kundagi asosiy ilmiy-texnik muammolardan biri bo‘lib qolmoqda va bu holat o‘zbek tilshunosligida raqamli resurslarni yanada takomillashtirish hamda xalqaro dasturiy ta’minotlar bilan integratsiya qilishni taqozo etadi.

Hozirgi davrda jahon miqyosida korpus lingvistikasining eng muhim va ustuvor tendensiyalari qatoriga sun’iy intellekt va neyron tarmoqlar imkoniyatlaridan foydalangan holda matnlarni yuqori aniqlikda avtomatik annotatsiya qilish, matn bilan bir qatorda audio, video va vizual ma’lumotlarni o‘z ichiga olgan zamonaviy multimodal korpuslarni barpo etish, raqamli resurslarga yetarli darajada ega bo‘lmagan kam resursli tillar uchun maxsus samarali lingvistik metodikalarni ishlab chiqish, foydalanuvchilar va tadqiqotchilar uchun ochiq manbali bo‘lgan hamda real vaqt rejimida doimiy ravishda yangilanib turuvchi dinamik korpuslarni shakllantirish, shuningdek, raqamli bazalarni tuzishda ma’lumotlar tarafkashligi va mualliflik etikasi muammolarini qat’iy hisobga olish kabi fundamental yo‘nalishlar kiradi.

Mazkur ilg‘or tendensiyalar o‘zbek tili lingvistikasi uchun ham ulkan ilmiy imkoniyatlar eshigini ochib beradi, xususan, bugungi kunda jadal rivojlanib borayotgan zamonaviy katta til modellari (LLM) va transfer learning (bilimlarni ko‘chirish) usullari yordamida o‘zbek tili korpuslarini yaratish jarayonini bir necha barobar tezlashtirish hamda uning sifat ko‘rsatkichlarini jahon standartlariga olib chiqish to‘la mumkin bo‘ladi.

Umuman olganda, til korpusi lingvistikasi o‘tgan asrning o‘rtalaridagi oddiy Brown Corpus davridan boshlab bugungi kundagi murakkab sun’iy intellekt davrigacha bo‘lgan shonli va mashaqqatli tarixiy yo‘lni bosib o‘tdi, shu bois o‘zbek tili uchun milliy, yuqori sifatli, chuqur morfologik va sintaktik annotatsiyalangan mukammal korpusni yaratish bugungi kunda milliy tilshunosligimiz oldida turgan eng dolzarb ilmiy-amaliy vazifalardan biri hisoblanadi. Ilg‘or jahon tajribasini o‘zbek tilining o‘ziga xos grammatik va leksik xususiyatlariga to‘la moslashtirish hamda sun’iy intellekt vositalaridan oqilona foydalanish orqali bu ulkan vazifani muvaffaqiyatli hal etish mumkin bo‘lib, kelgusida olib boriladigan

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

bunday fundamental tadqiqot natijalari o‘zbek tilshunosligi, kompyuter lingvistikasi va zamonaviy til texnologiyalarining kelgusidagi ravnaqiga xizmat qiladi.

Foydalanilgan adabiyotlar:

1. Abjalova, M. A. (turli yillardagi ishlari). O‘zbek tili milliy korpusi va kompyuter lingvistikasi masalalari bo‘yicha maqolalar.
2. Elov, B. B., & boshqalar. (2020–2025). O‘zbek tili korpuslari, morfologik analizatorlar va annotatsiya masalalari bo‘yicha ilmiy maqolalar.
3. O‘zbekiston Respublikasi Fanlar akademiyasi va Toshkent davlat o‘zbek tili va adabiyoti universiteti materiallari. O‘zbek tilining elektron korpuslari bo‘yicha hisobotlar va maqolalar.
4. Sharipov, M., & boshqalar. (2022–2025). O‘zbek tili uchun POS-tagging va morfologik annotatsiya bo‘yicha tadqiqotlar.
5. Hamroyeva, Sh. (turli yillar). Milliy til korpusi va uning lingvistik xususiyatlari bo‘yicha maqolalar.

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