O‘ZBEK TILINING MILLIY KORPUSINI YARATISHDA SUN’IY INTELLEKT TEXNOLOGIYALARIDAN FOYDALANISH

Abdumalikova Mushtariybegim Mashxurbek qizi
O‘zbekiston davlat jahon tillari universiteti
Xorijiy til va adabiyot fakulteti 1-bosqich talabasi.

Annotatsiya. *Mazkur maqolada o‘zbek tilining milliy korpusini yaratish jarayonida sun’iy intellekt texnologiyalaridan foydalanishning nazariy va amaliy jihatlari keng qamrovli tahlil qilinadi. Milliy korpus tushunchasi, uning lingvistik tadqiqotlar va ta’lim tizimidagi ahamiyati yoritiladi. Shuningdek, tabiiy tilni qayta ishlash (NLP), mashinaviy o‘qitish, chuqur o‘rganish va avtomatik annotatsiya kabi sun’iy intellekt texnologiyalarining korpus yaratishdagi roli ilmiy asosda tahlil qilinadi. Maqolada o‘zbek tilining morfologik murakkabligi, dialektal xilma-xilligi va kichik resursli til sifatidagi xususiyatlari hisobga olingan holda amaliy takliflar ilgari suriladi.*

Kalit so‘zlar: *milliy korpus, o‘zbek tili, sun’iy intellekt, tabiiy tilni qayta ishlash, avtomatik annotatsiya, mashinaviy o‘qitish.*

Annotation. *This article provides a comprehensive analysis of the use of artificial intelligence technologies in the development of the National Corpus of the Uzbek language. It examines the concept of a national corpus and its significance for linguistic research and education. The role of natural language processing (NLP), machine learning, deep learning, and automatic annotation in corpus construction is discussed. Considering the morphological complexity, dialect diversity, and low-resource status of Uzbek, the study proposes practical recommendations for corpus development.*

Keywords: *national corpus, Uzbek language, artificial intelligence, natural language processing, automatic annotation, machine learning.*

Аннотация. *В статье рассматриваются теоретические и практические аспекты использования технологий искусственного интеллекта при создании Национального корпуса узбекского языка. Анализируется значение национального корпуса для лингвистических исследований и образовательной системы. Особое внимание уделяется применению технологий обработки естественного языка, машинного обучения и автоматической аннотации. С учетом морфологической сложности и статуса узбекского языка как языка с ограниченными ресурсами предлагаются практические рекомендации.*

Ключевые слова: *национальный корпус, узбекский язык, искусственный интеллект, обработка естественного языка, автоматическая аннотация.*

Globallashuv va raqamlashtirish jarayonlari tilshunoslik sohasida yangi metodologik yondashuvlarni shakllantirdi. Zamonaviy lingvistik tadqiqotlar endilikda katta hajmdagi real matnlar bazasiga, ya’ni korpuslarga tayanadi. Milliy korpus — bu muayyan tilning yozma va og‘zaki namunalari tizimli ravishda jamlangan, elektron

shaklda saqlanadigan va lingvistik belgilashdan o‘tgan ma’lumotlar bazasidir. Bunday korpus tilning leksik, grammatik, uslubiy va pragmatik xususiyatlarini ilmiy asosda o‘rganish imkonini beradi.

O‘zbek tili turkiy tillar oilasiga mansub bo‘lib, boy morfologik tizimga ega. Agglutinativ tuzilma, so‘z yasashning keng imkoniyatlari va dialektal xilma-xillik milliy korpus yaratish jarayonini murakkablashtiradi. An’anaviy usullar asosida qo‘lda annotatsiya qilish katta vaqt va resurs talab qiladi. Shu sababli sun’iy intellekt texnologiyalarini jalb etish dolzarb vazifaga aylanmoqda.

Milliy korpus tilning real qo‘llanilish namunasini aks ettiradi. U nazariy grammatik qoidalarni emas, balki amaliy nutq jarayonini ko‘rsatadi. Korpus lingvistik tadqiqotlar, lug‘at tuzish, tarjima, til o‘qitish metodikasi va sun’iy intellekt tizimlarini o‘qitishda asosiy manba vazifasini bajaradi. Dunyo tajribasida ingliz tilining *“British National Corpus”,* rus tilining *“Национальный корпус русского языка”* kabi yirik loyihalari mavjud. Ular millionlab matn birliklarini o‘z ichiga oladi va avtomatik qidiruv, morfologik tahlil hamda semantik belgilash imkoniyatini beradi. O‘zbek tilida ham milliy korpus yaratish tashabbuslari mavjud bo‘lsa-da, ularning hajmi va funksional imkoniyatlari hali to‘liq rivojlanmagan.

O’zbek tilining milliy korpusini yaratishda sun’iy intellekt texnologiyalaridan ham keng foydalanilmoqda. Sun’iy intellekt, xususan tabiiy tilni qayta ishlash texnologiyalari korpus yaratish jarayonini sezilarli darajada tezlashtiradi. Avtomatik tokenizatsiya, lemmatizatsiya va morfologik tahlil vositalari katta hajmdagi matnlarni qisqa vaqt ichida qayta ishlash imkonini beradi. Mashinaviy o‘qitish algoritmlari matnlarni janr, uslub va mavzu bo‘yicha tasniflashda qo‘llaniladi. Chuqur o‘rganish modellariga asoslangan neyron tarmoqlar esa kontekstual semantik belgilashni amalga oshiradi. Bunday yondashuvlar qo‘lda bajariladigan annotatsiya jarayonini avtomatlashtirish imkonini beradi.

O‘zbek tilining agglutinativ tuzilishi sababli bir so‘z tarkibida bir nechta grammatik affikslar mavjud bo‘lishi mumkin. Sun’iy intellekt asosidagi morfologik analizatorlar so‘zlarni ildiz va qo‘shimchalarga ajratib, grammatik kategoriyalarni aniqlaydi. Bu jarayon korpusning lingvistik aniqligini oshiradi.

Korpus yaratishda eng muhim bosqichlardan biri — annotatsiya jarayonidir. Annotatsiya deganda matn birliklariga lingvistik belgi qo‘yish tushuniladi. Bunga so‘z turkumini aniqlash, morfologik xususiyatlarni belgilash, sintaktik bog‘lanishlarni ko‘rsatish kiradi. Sun’iy intellekt yordamida avtomatik annotatsiya qilish jarayoni tezkor bo‘lsa-da, ayrim muammolarni yuzaga keltiradi. Masalan, o‘zbek tilida omonim shakllar ko‘p uchraydi. “Olma” so‘zi meva yoki buyruq shakli bo‘lishi mumkin. Model kontekstni to‘g‘ri aniqlamasa, noto‘g‘ri belgilash yuz beradi. Shuningdek, dialektal variantlar va so‘zlashuv shakllari ham avtomatik tizimlar uchun qiyinchilik tug‘diradi. Shu sababli sun’iy intellekt va lingvist mutaxassislarning hamkorligi muhim hisoblanadi.

Kichik resursli til muammosi

O‘zbek tili global miqyosda kichik resursli til hisoblanadi. Annotatsiyalangan korpuslar hajmi cheklanganligi sababli sun’iy intellekt modellarini o‘qitish uchun yetarli ma’lumot mavjud emas. Bu esa model aniqligini pasaytiradi. Transfer learning va ko‘p tilli modellar yordamida ushbu muammoni qisman hal qilish mumkin. Masalan, turkiy tillar orasidagi o‘xshashliklardan foydalanib, modelni moslashtirish samarali natija berishi mumkin.

O‘zbek tilining milliy korpusini yaratishda bosqichma-bosqich strategiya ishlab chiqish zarur. Avvalo, turli janr va uslublardagi matnlar to‘planishi lozim. Keyinchalik sun’iy intellekt asosida dastlabki avtomatik annotatsiya amalga oshiriladi. So‘ngra lingvist mutaxassislar tomonidan verifikatsiya jarayoni o‘tkaziladi. Shuningdek, korpusni ochiq platforma sifatida rivojlantirish, tadqiqotchilar va dasturchilar uchun API interfeys yaratish muhimdir. Bu ilmiy hamkorlikni kuchaytiradi va til texnologiyalarini rivojlantirishga xizmat qiladi.

Xulosa qilib aytganda, o‘zbek tilining milliy korpusini yaratish zamonaviy lingvistika va raqamli texnologiyalar integratsiyasini talab etadigan murakkab jarayondir. Sun’iy intellekt texnologiyalari ushbu jarayonni tezlashtirish, aniqlikni oshirish va resurslarni tejash imkonini beradi. Biroq to‘liq avtomatlashtirish emas, balki inson va sun’iy intellekt hamkorligi eng samarali yondashuv hisoblanadi. Kelajakda milliy korpus o‘zbek tilining ilmiy rivojlanishi va global axborot makonida munosib o‘rin egallashiga xizmat qiladi.

FOYDALANILGAN ADABIYOTLAR:

1.O’zbek tili uchun sun’iy intellekt asosida intellektual platforma yaratish to’g’risida ma’lumot. Vaqt axborot portali, 2025.

2.O’zbekiston sun’iy intellektning milliy til modelini ishlab chiqmoqda. Gazeta.uz, 2025.

3.Istiqlol davri o’zbek tili terminologiyasining rivoji va raqamli texnologiyalar ta’siri. O’zbekiston tilshunoslik materiallari, 2023.

4. H. Dadaboev. O„zbek terminologiyasining zamonaviy rivojlanish yo„nalishlari va raqamli resurslar. Monografiya, 2024.