# SUN'IY INTELLEKT ASOSIDA TIL KORPUSLARINI YARATISHDA AVTOMATIK TAHLIL QILISHNING LINGVISTIK MODELLARI

Andijon davlat chet tillari instituti stajor tadqiqotchisi
**Alimova Ma'mura**
**DOI:**https://doi.org/10.5281/zenodo.23249258

**Annotatsiya:** Maqolada zamonaviy sun'iy intellekt texnologiyalari va chuqur neyron tarmoq modellarini o'zbek tili korpuslarini yaratish jarayoniga tatbiq etish masalalari tadqiq qilingan. Shuningdek, an'anaviy qo'lda annotatsiya qilish usullarining cheklovlari tahlil qilinib, transfer learning va katta til modellari yordamida o'zbek tilining agglyutinativ tabiatidan kelib chiqadigan omonimiya hamda sintaktik noaniqliklarni bartaraf etish yo'llari ochib berilgan. Morfologik va sintaktik tahlilni avtomatlashtirish orqali mashina tarjimasi, avtomatik tahrirchilar va nutqni tanish tizimlarini rivojlantirish istiqbollari yoritilgan.

**Kalit so'zlar:** korpus lingvistikasi, sun'iy intellekt, neyron tarmoqlar, avtomatik tahlil, agglyutinativ til, o'zbek tili, transfer learning, morfologik teglash.

**Аннотация:** В статье исследуются вопросы применения современных технологий искусственного интеллекта и моделей глубоких нейронных сетей в процессе создания корпусов узбекского языка. Проанализированы ограничения традиционных методов ручной аннотации, а также раскрыты пути преодоления омонимии и синтаксических неопределенностей, обусловленных агглютинативным характером узбекского языка, с помощью transfer learning и больших языковых моделей. Освещены перспективы развития машинного перевода, автоматических редакторов и систем распознавания речи посредством автоматизации морфологического и синтаксического анализа.

**Ключевые слова:** корпус лингвистика, искусственный интеллект, нейронные сети, автоматический анализ, агглютинативный язык, узбекский язык, transfer learning, морфологическая разметка.

**Annotation:** The article examines the application of modern artificial intelligence technologies and deep neural network models to the creation of Uzbek language corpora. It analyzes the limitations of traditional manual annotation methods and explores ways to overcome homonymy and syntactic ambiguities—arising from the agglutinative nature of the Uzbek language—using transfer learning and large language models. The prospects of developing machine translation, automated editors, and speech recognition systems through the automation of morphological and syntactic analysis are also highlighted.

**Keywords:** corpus linguistics, artificial intelligence, neural networks, automatic analysis, agglutinative language, Uzbek language, transfer learning, morphological tagging.

Zamonaviy ilm-fan taraqqiyoti, xususan, kompyuter lingvistikasi va sun'iy intellekt texnologiyalarining jadal rivojlanishi an'anaviy tilshunoslik tadqiqot usullarini tubdan o'zgartirib



*Mualliflik huquqi egasi tomonidan ushbu material Creative Commons Attribution 4.0 (CC BY 4.0) litsenziyasi asosida tarqatiladi.*

yubordi. Agar o'tmishda til hodisalari asosan olimning shaxsiy intuitsiyasi, cheklangan nazariy modellar yoki qo'lda yig'ilgan kichik hajmli kartotekalar asosida o'rganilgan bo'lsa, bugungi kunda empirik dalillarga tayanuvchi korpus lingvistikasi jahon ilm-fanida yetakchi o'rinni egallamoqda. Til korpusi — real hayotda ishlatiladigan minglab va millionlab matnlarning maxsus tizimlashtirilgan, lingvistik jihatdan annotatsiya qilingan elektron bazasi bo'lib, u tilni tabiiy holatida har tomonlama tadqiq etish imkonini beradi. O'zbek tili kabi boy tarixiy ildizlarga ega, biroq raqamli resurslar bazasi jadal shakllanish bosqichida bo'lgan tillar uchun sifatli va keng qamrovli milliy korpularni yaratish strategik ahamiyat kasb etadi.

Jahon tajribasi shuni ko'rsatmoqdaki, korpus lingvistikasining dastlabki bosqichlarida matnlarni yig'ish va ularga ishlov berish asosan qo'lda (manual) amalga oshirilgan bo'lib, bu jarayon ko'p vaqt va ulkan inson resurslarini talab qilgan. Bugungi kunga kelib esa raqamli texnologiyalar va sun'iy intellekt (Artificial Intelligence) vositalarining kirib kelishi bu jarayonni avtomatlashtirish, katta til modellari (Large Language Models) hamda chuqur neyron tarmoqlar yordamida matnlarni yuqori aniqlikda tahlil qilish imkonini bermoqda. Xususan, tabiiy tilni qayta ishlash (NLP) sohasidagi so'nggi yutuqlar, so'zlarning vektorli taqdimoti (Word Embedding) va transfer learning usullari til korpuslarini yaratishda avtomatik annotatsiya qilishning yangi lingvistik modellarini ishlab chiqishni taqozo etmoqda.

Shu bilan birga, o'zbek tilining o'ziga xos agglyutinativ tabiati, boy so'z yasalish imkoniyatlari, omonimiya hodisalari hamda sintaktik qurilishdagi murakkabliklar matnlarni avtomatik qayta ishlash va morfologik-sintaktik teglash jarayonida o'ziga xos ilmiy-metodologik muammolarni keltirib chiqarmoqda. Mavjud o'zbek tili korpuslarida annotatsiya sifatini oshirish, xalqaro standartlarga mos keladigan avtomatik tahlil qilishning lingvistik modellarini yaratish va ularni amaliyotga joriy etish bugungi kunda zamonaviy o'zbek tilshunosligi oldida turgan eng dolzarb vazifalardan biridir. Ushbu maqolaning asosiy maqsadi sun'iy intellekt bazasida til korpuslarini yaratishda avtomatik tahlil qilishning nazariy-metodologik asoslarini ochib berish hamda o'zbek tili materiallari asosida samarali lingvistik modellarni ishlab chiqish yo'llarini ilmiy jihatdan asoslashdan iboratdir.

Korpus lingvistikasining shakllanish va rivojlanish tarixi shuni ko'rsatadiki, sifatli va hajmli matn bazasini yaratish eng murakkab hamda mas'uliyatli bosqichlardan biri hisoblanib, matnlarga lingvistik ishlov berish va ularni annotatsiya qilishda dastlabki davrlarda asosan qo'lda bajariladigan manual usullarga tayanilgan va bu yondashuv o'zining jiddiy cheklovlari hamda metodologik qiyinchiliklarini namoyon etgan, chunki minglab va millionlab so'zlardan iborat matnlarni inson kuchi yordamida tahlil qilish, ularning turkumiy mansubligini belgilash va sintaktik bog'liqliklarini tarmoqlash ulkan vaqt resurslarini, ortiqcha moliyaviy xarajatlarni hamda ko'p sonli mutaxassislarning tinimsiz mehnatini talab qilgan, shuningdek, qo'lda annotatsiya qilish jarayonida subyektiv omillar, ya'ni tadqiqotchining shaxsiy yondashuvi, bilim darajasi yoki o'ziga xos intuitsiyasi tufayli yo'l qo'yiladigan xatolar natijalarning obyektivligiga va bir xilligiga putur yetkazgan. Ana shu jiddiy cheklovlarni bartaraf etish maqsadida zamonaviy lingvistikada matnlarni avtomatik qayta ishlashning statistik va lingvistik yondashuvlarini o'zaro uyg'unlashtirgan innovatsion metodologiyalar ishlab chiqildi, chunki ilk bosqichlarda statistik modellar matndagi so'zlarning uchrash chastotasiga tayanib ish ko'rgan bo'lsa, sof lingvistik yondashuvlar tilning qat'iy qoidalari va morfemik tahlilga asoslangan edi, bugungi kunga kelib esa bu ikki yondashuvning o'zaro integratsiyalashuvi natijasida gibrid modellar yaratildi va bu uyg'unlik kompyuterga nafaqat so'zlarning statistik taqsimotini, balki ularning ma'noviy va strukturaviy tabiatini ham birdek hisobga olgan holda yuqori aniqlikda ishlash imkonini bermoqda. Avtomatik tahlil qilish prinsiplari tilshunoslikning barcha asosiy sathlarini


Mualliflik huquqi egasi tomonidan ushbu material Creative Commons Attribution 4.0 (CC BY 4.0) litsenziyasi asosida tarqatiladi.

bosqichma-bosqich qamrab olib, avvalo leksik darajada matn tarkibidagi so‘zlarning asli, qo‘llanish chastotasi va omonimlik xususiyatlari aniqlanadi, so‘ngra morfologik darajada har bir so‘z shaklining turkumi va grammatik kategoriyalari avtomatik tegglanadi, shuningdek, sintaktik darajada gap tarkibidagi so‘zlarning tobe va hokim munosabatlari iyerarxik daraxt ko‘rinishida modellashtirilib, semantik darajada esa so‘zlarning matn kontekstidagi haqiqiy ma’nosi va o‘zaro bog‘liqliklari sun’iy intellekt vositalari yordamida chuqur tahlil qilinadi.

Zamonaviy kompyuter lingvistikasida sun’iy intellekt va neyron tarmoqlar texnologiyalarining jadal joriy etilishi matn korpuslarini yaratish va ularga ishlov berish jarayonini sifat jihatidan mutlaqo yangi bosqichga ko‘tardi, chunki bugungi kunda keng qo‘llanilayotgan katta til modellari va chuqur o‘qitish algoritmlari inson aralashuvini minimal darajaga tushirgan holda ulkan hajmdagi matnlarni avtomatik ravishda tahlil qilish, tasniflash va lingvistik jihatdan boyitish imkonini bermoqda. Ushbu jarayonda so‘zlarni kompyuter tushunadigan raqamli shaklga keltirish uchun ularni vektorli taqdim etish usullari, xususan, zamonaviy Word Embedding texnologiyalari va BERT arxitekturasi muhim poydevor bo‘lib xizmat qiladi, biroq bu universal yondashuvlarni o‘zbek tilining o‘ziga xos agglyutinativ tabiati, ya’ni so‘z o‘zagiga zanjirsimon tarzda birin-ketin qo‘shimchalar qo‘shilib kelishi va boy so‘z yasalish imkoniyatlariga moslashtirish chuqur lingvistik adaptatsiyani talab qiladi. Morfologik teglash va sintaktik daraxtlarni qurish jarayonlarida aynan shu sun’iy intellekt modellari matn tarkibidagi har bir leksemaning turkumi, grammatik shakli hamda gapdagi tobe va hokim so‘zlar orasidagi munosabatlarni ehtimolliklar nazariyasi va neyron tarmoq qatlamlari orqali yuqori aniqlikda hisoblab chiqib, avtomatik ravishda aniqlab beradi va natijada korpusning ilmiy salohiyatini keskin oshiradi.

O‘zbek tili korpusini yaratish va unga sun’iy intellekt asosida avtomatik ishlov berish jarayoni tilimizning o‘ziga xos tipologik tabiati, xususan, uning kuchli rivojlangan agglyutinativ xususiyati bilan chambarchas bog‘liq bo‘lib, bu holat so‘z o‘zagiga zanjirsimon tarzda birin-ketin qo‘shiladigan so‘z yasalish va o‘zgarish affikslarining cheksiz kombinatsiyalarini keltirib chiqaradi hamda har bir leksemaning morfologik tahlilini murakkablashtiradi. Agglyutinativ tillarda bitta o‘zakdan o‘nlab yangi so‘z shakllari va ma’no qatlamlari hosil qilinishi sababli, an’anaviy dasturlash algoritmlari ko‘pincha bu so‘z qoliplarini to‘laqonli qamrab ola olmaydi, bu esa korpus uchun mo‘ljallangan avtomatik teglash modellaridan yuqori darajadagi moslashuvchanlikni talab qiladi.

Mavjud o‘zbek tili milliy korpusi, tarjima va parallel korpuslardagi amaliy tahlil jarayonlarida ham bir qator jiddiy lingvistik muammolar yaqqol ko‘zga tashlanmoqda, xususan, bir xil shaklga ega bo‘lgan so‘zlarning turli turkumlarga o‘tishi yoki har xil ma’no anglatishi natijasida kelib chiqadigan omonimiya hodisalari, shuningdek, gap tarkibida so‘zlarning o‘zaro sintaktik birikuvchanligidagi noaniqliklar avtomatik tahlil jarayonida xatoliklarni keltirib chiqarmoqda. Masalan, birgina so‘z shaklining kontekstga qarab ham ot, ham fe’l, yoki ham ravish, ko‘makchi vazifasida kelishi sun’iy intellekt modellarining adashishiga sabab bo‘lishi mumkin, chunki sintaktik bog‘lanishlardagi iyerarxik munosabatlar har doim ham yuzaki statistik qoidalar asosida to‘g‘ri aniqlanavermaydi.

Ana shu qiyinchiliklarni bartaraf etish va o‘zbek tili uchun avtomatik annotatsiya sifatini tubdan oshirishda zamonaviy transfer learning (bilimlarni ko‘chirish) usullaridan foydalanish eng maqbul yo‘l sifatida e’tirof etilmoqda. Resurs bazasi boy bo‘lgan boshqa tillar uchun oldindan o‘qitilgan yirik universal neyron tarmoq modellarining (pretrained language models) bilimlarini o‘zbek tilining o‘ziga xos lingvistik xususiyatlariga moslashtirish, ya’ni fine-tuning qilish orqali qisqa vaqt ichida kamroq hajmdagi matnlar asosida ham yuqori aniqlikdagi morfologik va sintaktik tahlil


Mualliflik huquqi egasi tomonidan ushbu material Creative Commons Attribution 4.0 (CC BY 4.0) litsenziyasi asosida tarqatiladi.

natijalariga erishish mumkin bo'ladi. Bu usul o'zbek tilidagi kam resurslilik muammosini qisman yumshatibgina qolmay, balki sun'iy intellekt vositalari yordamida yaratilayotgan milliy korpusning kelgusidagi ilmiy ishonchliligi, xalqaro standartlarga muvofiqligi va amaliy qo'llanilish samaradorligini kafolatlaydi.

zamonaviy sun'iy intellekt texnologiyalari va chuqur neyron tarmoq modellarini til korpuslarini yaratish jarayoniga tatbiq etish o'zbek tili korpusining nafaqat umumiy hajmini, balki lingvistik annotatsiya sifatini ham sifat jihatidan mutlaqo yangi bosqichga ko'tarishda hal qiluvchi ahamiyat kasb etib, an'anaviy qo'lda bajariladigan mashaqqatli va vaqt talab qiladigan usullardan voz kechib, transfer learning va katta til modellari imkoniyatlaridan foydalanish o'zbek tilining agglyutinativ tabiatidan kelib chiqadigan qiyinchiliklarni muvaffaqiyatli yengish, omonimiya hamda sintaktik noaniqliklarni bartaraf etish hamda qisqa muddatda millionlab so'zdan iborat ishonchli elektron bazalarni shakllantirish imkonini beradi. Morfologik va sintaktik tahlil jarayonlarini to'liq avtomatlashtirish o'zbek tilida tabiiy tilni qayta ishlash yo'nalishini jadal rivojlantirish uchun mustahkam poydevor yaratib, kelgusida mukammal mashina tarjimasi tizimlarini, imlo va uslub xatolarini avtomatik to'g'rilovchi intellektual tahrirchilarni, shuningdek, ovozli ma'lumotlarni matnga o'giruvchi zamonaviy nutqni tanish tizimlarini yaratish va amaliyotga joriy etish uchun zarur bo'lgan ilmiy-texnik bazani ta'minlaydi. Mazkur yo'nalishdagi kelgusidagi ilmiy tadqiqotlarni yanada chuqurlashtirish maqsadida o'zbek tilining fonetik, morfologik va sintaktik xususiyatlarini to'laqonli hisobga oluvchi maxsus ochiq manbali dasturiy modellarni ishlab chiqish, xalqaro miqyosdagi ilg'or korpus arxitekturalari tajribasiga tayangan holda milliy, ta'limiy va parallel korpuslarni o'zaro integratsiya qilish hamda yosh tadqiqotchilarni kompyuter lingvistikasi va neyron tarmoqlarni o'zbek tili materiallariga qo'llash bo'yicha amaliy ko'nikmalarini oshirishga qaratilgan chora-tadbirlarni amalga oshirish tavsiya etiladi.

**Foydalanilgan adabiyotlar:**

1. Yo'ldoshev B. O'zbek leksikografiyasi va korpus lingvistikasi asoslari. – Toshkent: Fan va texnologiya, 2020.
2. To'ychiboyev B. O'zbek tili milliy korpusini yaratishning nazariy va amaliy muammolari. // O'zbek tili va adabiyoti jurnali, 2022. – 3-son. – B. 45–52.
3. Jumaniyozov R. Kompyuter lingvistikasi va o'zbek tili korpusi. – Toshkent: Akademnashr, 2021.
4. Mamatov A. Hozirgi o'zbek adabiy tilining semantik sintaksisi va korpus texnologiyalari. – Toshkent: Mumtoz so'z, 2023.
5. Safarov O. Lingvistik tadqiqotlarda zamonaviy texnologiyalar va sun'iy intellekt elementlari. – Samarqand: SamDU nashri, 2022.
6. Karimov S. O'zbek tilshunosligida matn tahlili va elektron bazalar. – Toshkent: Universitet, 2021.
7. Rahimov A. Tabiiy tilni qayta ishlash (NLP) va o'zbek tilining aglutinativ xususiyatlarini modellashtirish. // O'zbekiston tilshunosligi jurnali, 2024. – 2-son. – B. 60–68.



Mualliflik huquqi egasi tomonidan ushbu material Creative Commons Attribution 4.0 (CC BY 4.0) litsenziyasi asosida tarqatiladi.