O'ZBEK TILI OVOZLI KORPUSINI RIVOJLANTIRISHNING ZAMONAVIY LINGVISTIK VA TEXNIK MASALALARI.

Umarova Zuhra Zafar qizi
Toshkent davlat sharqshunoslik universiteti,
Sharq xalqlari tillari va adabiyoti instituti,
Tarjimashunoslik, tilshunoslik va xalqaro
jurnalistika oliy maktabi Kompyuter
lingvistika guruhi 2-kurs talabasi
Tel: 947799305 zuhraumarova2006@gmail.com

Annotatsiya: *Ushbu ilmiy maqola hisoblash tilshunosligining eng dinamik va dolzarb sohalaridan biri bo'lgan o'zbek tilining nutq korpusining shakllanishi, me'moriy tuzilishi, lingvistik annotatsiya bosqichlari va texnik infratuzilmasining rivojlanishini har tomonlama tahlil qiladi. Tadqiqotda sun'iy intellekt tizimlari, ovozli yordamchilar, avtomatik nutqni aniqlash (ASR) va matndan nutqqa (TTS) modellarini tayyorlashda neyron tarmoqlarining imkoniyatlari o'rganiladi. Maqolada o'zbek tilining nutq bazasini boyitishda duch keladigan fonetik, dialektal va raqamli muammolar tasniflanadi, xalqaro standartlarga mos keladigan amaliy yechimlar va me'moriy dasturiy ta'minot modellari taklif etiladi.*

Kalit so'zlar: *Nutq korpusi, hisoblash lingvistikasi, sun'iy intellekt, avtomatik nutqni aniqlash (ASR), matnni nutqqa aylantirish (TTS), lingvistik annotatsiya, akustik model, neyron tarmoqlari, dialektologiya,raqamlashtirish.*

Abstract: *This scientific article comprehensively analyzes the formation, architectural structure, linguistic annotation stages, and technical infrastructure development of the Uzbek speech corpus—one of the most dynamic and relevant areas of computational linguistics. The study examines the capabilities of neural networks in training artificial intelligence systems, voice assistants, Automatic Speech Recognition (ASR), and Text-to-Speech (TTS) models. The article classifies phonetic, dialectal, and digital challenges encountered in enriching the Uzbek speech database, and proposes practical solutions and architectural software models that comply with international standards.*

Keywords: *Speech corpus, computational linguistics, artificial intelligence, Automatic Speech Recognition (ASR), Text-to-Speech (TTS), linguistic annotation, acoustic model, neural networks, dialectology, digitalization.*

Аннотация: *В данной научной статье всесторонне анализируются формирование, архитектурная структура, этапы лингвистической аннотации и развитие технической инфраструктуры узбекского речевого корпуса — одного из наиболее динамичных и актуальных направлений компьютерной лингвистики. В*

исследовании рассматриваются возможности нейронных сетей в обучении систем искусственного интеллекта (ИИ), голосовых помощников, моделей автоматического распознавания речи (ASR) и синтеза речи (TTS). В статье классифицируются фонетические, диалектологические и цифровые проблемы, возникающие при обогащении речевой базы данных узбекского языка, а также предлагаются практические решения и архитектурные программные модели, соответствующие международным стандартам.

Ключевые слова: *Речевой корпус, компьютерная лингвистика, искусственный интеллект, автоматическое распознавание речи (ASR), синтез речи (TTS), лингвистическая аннотация, акустическая модель, нейронные сети, диалектология, цифровизация.*

KIRISH

XXI asrning o'quvchi choragiga kelib, axborot texnologiyalari va sun'iy intellekt (Artificial Intelligence) tizimlari insoniyat hayotining barcha jabhalariga shiddat bilan kirib bordi. Ush global texnologiya inqilobiga asoslangan tillarning hisoblash makonda mavjud bo'lishi va boshqaruv fayli boshqaruvi yashovchanligini milliy o'ziga xosligini saqlab qolishning eng asosiy shartiga aylandi. Bugun kunda qaytariladigan kompyuter matnli ma'lumotlarni qayta ishlash emas, balki inson nutqini tinglaydi, tushunadigan va inson tilida javob beradigan intellektual hamkor darajasiga ko'tariladi. O'zbek tili raqami ekotizimini kerak, tilining nufuzini xalqaro miqyosda korpus aniqling o'zbek tilivistikasini, tashqi ko'rinishi, uning davlat qismi hisoblangan ovozli (nutq) korpuslarini rivojlantirish bilan chambachas bog'liqdir.O‘zbekiston Respublikasi Prezidentining o‘zbek tilining davlat tili sifatidagi nufuzi va mavqeini tubdan tuzatish, tilimizni zamonaviy axborot texnologiyalari bilan integratsiya qilish hamda milliy korpuslarni aniqlashga oid qabul qilgan tarixiy farmon va qarorlari ushbu sohadagi ilmiy-amaliy yordamga mustahkam huquqiy va iqtisodiy zamin yaratdi. Shunga, xitoy yoki ingliz tili, rus, o'zbek tilidagi mukammal, yuqori sifatli akustik va lingvistik dizayndagi audio ma'lumotlar bazasi hali ham yetarli hajmda shakllanmagan. Mavjud resurslarning tarqoqligi, xalqaro ochiq manbali platformalarga integratsiyalashmaganligi qo'shimcha IT-soha va tilshunos olimlar oldiga strategik harakatlarni'ymoqda.Inson va kompyuter o'rtasida og'zaki muloqotni ta'minlash uchun mashina xotira millionlab soatlik inson nutqi namunalari va fayl aniq yozmalarini ishlab chiqish talab qilish. Ush ilmiy maqolaning maqsadi - o'zbek tili ovozli korpusini kuzatishning fundamental va amaliy muammolarini aniqlash, uning to'g'ridan-to'g'ri ta'minlash hamda sohadagi mavjud texnik to'g'rilash bo'yicha optimal algoritmlarni bartaraf etish bo'yicha.

Korpus lingvistikasi va nutq texnologiyalari (Speech Technologies) sohasining nazariy asoslari jahon va ilmiy keng tadqiq etilgan. Umumjahon miqyosida nutq korpuslarini olish

XX asrning yarmidan yo'l bo'lib, bu ilishda B. Juang, L. Rabin, J. Allen kabi olimlarning fundamental ishlari nutqni avtomatik (ASR) va yashirin Markov modellari (Hid Markov Models - HMM) rivojlanishiga turtki bo'ldi. “Mozilla” tashkilotining “Common Voice” loyihasi hamda Google kompaniyasining “Crowdsourcing” tashabbuslari bilan xalqaro miqyosda ochiq nutq korpuslarini ishlab chiqishda metodologik baza bo'lib xizmat qilmoqda.O'zbek tilshunosligi va kompyuter lingvistikasida korpus masalalari keyin jadallashdi. Elov rahbarligidagi professor B tadqiqot tili o'zbek milliy korpusining (uzbekcorpus.uz) nazariy modellarini ishlab chiqarish chiqdilar. Olimlar A. Po‘latov, M. Primova va Sh. Yusupovlarning ilmiy ishlarida o'zbek tili grammatikasi va fonetikasini algoritmlash, so'zlarning morfologik va sintaktik tahlili masalalari yoritilgan.Buning uchun, o'zbek tilining og'zaki nutq korpusi, uning akustik xususiyatlari, hududiy dialektlarning neyron tarmoqlariga ta'siri va fonetik annotatsiya qilishlari alohida yaxlit tadqiqot ob'ekti sifatida tizimli o'rganilmagan. Mavjud ko'proq yetishtirish nutq korpusiga yo'naltirilgan bo'lib, ovozli ma'lumotlar bazasini boyitishning texnik standartlari tahlili qilmaydi.Uchta tadqiqot mavzusi o'zbek tilidagi ovozli korpusini ishlab chiqish va rivojlantirish kompleks tizimi bir zamonaviy lingvistik qator metodlardan foydalanish mumkin edi:Deskriptiv (tavsifiy): O'zbek tili unli va undosh tovushlarining nutq usullaridagi akustik ko'rsatkichlarini va fayl kompyuter dasturlarida aks etish texnologiyalarida qo'llandi.Statistik tahlil va Matematik modellashtirish: Nutq namunalarini raqamlashtirishda signal-shovqin nisbati (SNR - Signal-to-Noise Ratio) hamda chastota tahlilini amalga oshirishda aniqlandi.Lingvistik annotatsiya va Segmentatsiya: Audiofayllarni so'z, bo'g'in va fonema darajasida sifati hamda xalqaro IPA (International Phonetic Alphabet) standartlariga moslashtirish usullarini tahlil qiladi. Neuron tarmoqlarini modellashtirish (Deep Learning): Nutqnida baholash qo'ygan zamonaviy *Transformer* arxitekturalari va protsessor o'zbek tili akustik bazaga ko'rsatkichlarni eksperimental qiyoslandi. Ovozli korpus oddiy audioyozuvlar to'plami emas, balki lingvistik va akustik qattiq qat'iy moslashtirilgan ma'lumotlar bazasidir. U sun'iy inlekt modellarini o'qitish uchun yo'qotilgan "Dataset" (to'plami)tel ma'lumotlarini ishlash. O'zbek tili ovozli korpusi o'z ichiga ikki asosiy komponentni olishi kerak:Akustik komponent Turli sharoitlarda, turli mikrofonlar yordamida yozib olingan yuqori sifatli audiofayllar (WAV yoki FLAC formatida, 16 kHz dan kam bo'lmaganda).Tekstual komponent: Ush audiolarning har bir soniyasiga mos keladigan aniq, xatolardan xoli yozma transkripsiyasi (matni).Ovozli korpusning tashqi annotatsiya (belgilanish) darajasi bilan o'lchanadi. Annotatsiya - bu audiofaylning ma'lum bir bo'lagiga tegishli lingvistik ma'lumotlarni yig'ish jarayonidir. O'zbek tili uchun annotatsiya jarayoni to'rt bosqichda amalga oshirish shart:Metamatnli annotatsiya: So'zlovchining jinsi, yoshi, qaysi hududdan (shevasi),yozib olingan xonadagi zaryad va mikrofon turi haqida ma'lumotlar kiritiladi.Ortografik annotatsiya: Ovozli nutq aynan eshitish matnga ko'chiriladi. Bunda

imlo tartibga rioya qilish bilan birga, so'zlovchining nutqiy parazit so'zlari (masalan: "ee", "xo'sh", "haligi") ham maxsus teglar (tags) yordam yordam.Fonetik annotatsiya: So'zlarning talaffuz shakli IPA transkripsiyasida aks ettiriladi. Bu ayniqsa o'zbek tilidagi qisqa va cho'ziq unlilar yoki korpusga xos urg'u o'rganishda muhim.Semantik-annotatsiya: Nutq matnidagi so'z turkumi va gapdagi gapdagi jarayonlar, bu esa kompyuterning nutq mazmunini oshirishga yordam beradi (Natural Language Understanding - NLU). O'zbek Tili Raqamlashtirishdagi Texnik va Lingvistik muammolariHisoblar manbalari o'zbek tili ovozli korpusini va ishlashida bir qator jiddiy to'siqlar aniqlandi. Ularni ikki guruhga berishi mumkin:

Lingvistik muammolar: Dialektal xilma-xillik: O'zbek tili uch kir lahja guruhiga (qarluq, qipchoq va o'g'uz) bo'linadi. Toshkent, Samarqand, Xorazm yoki Farg'ona vodiysi boshqaruvining nutqi fonetik jihatdan farq qiladi. Agar korpus faqat adabiy tilda yozilsa, sun'iy intellekt viloyatlardagi jonli nutqni tushunmaydi.Garmonizm (singharmonizm) qoldiqlari va ur'u: Ba'zi shevalarda singharmonizm stresslarning saqlangani, so'z yuklaydigan tovushlarning tushib qolishi yoki o'zidan (masalan: "kelayapman" "kevomman", "kelyapplarni") akustik model chalkashtiradi.Grafika (Alifbo) muammosi: O'zbek jamiyatida ham lotin, ham kirill alifbosi parallel qo'llanilmoqda. Ovozli bazani kuzatishda transkripsiyalarni avtomatik konvertatsiya qilish universal tizim zarur.

Texnik muammolar:Signal sifati va shovqinlar: Ommaviy axborot vositalari yoki ijtimoiy tarmoqlardan olingan audiolarda musiqiy fon xatolar yuqoriligi tufayli o'rganish mumkin model."Ochiq kodli" resurslar tanqisligi: kabi gigantlar o'zbekcha ovozli ma'lumotlar bazasiga ega bo'lsa-da, Google Yandex. Mahalliy olimlar va startaplar uchun bepul yuklab olish.Crowdscing platformasi: Telegram bot yoki mobil ilova orqali respublikaning barcha qurilmalaridagi matnlar sotib olishga o'zbekcha va ovozlarni yozib olish.Dasturiy filtrlash (Spectral Subtraction): Yozib olingan audiolardagi toshlar Pythondagi *Librosa* va *SciPy* kutubxonalari yordamida avtomatik filtrlanadi.Neuron tarmoq verifikatsiyasi: Ovoz va matn mosligi dastlabki bosqichda OpenAI Whisper orqali tekshirilib, moslik darajasi 95% dan yuqori bo'lgan fayllar korpusning asosiga mustahkamlanadi.

O'zbek tili ovozli korpusini bo'yicha olingan natijalar shuni ko'rsatadiki, akademik tilshunoslik tomonidan cheklanib qolish kutilgan samarani bermaydi. Kompyuter lingvistikasi - bu fanlararo (interdisiplinar) soha bo'lib, tilshunos olimlar bilan dasturchi-muhandislar va ma'lumotlar bo'yicha mutaxassislarning (Data Scientists) ta'minlash hamkorligini talab qiladi.Taklif qilingan model asosida yig'iladigan ma'lumotlarda boshqa davlat boshqaruvi va ijtimoiy soha inqilobiy o'zgarishlarga sabab bo'ladi. Masalan, elektron hukumat (E-government) tizimlarida fuqarolarga inter ovozli xizmat ko'rsatish, sud va huquq-tartibot organlarida so'roq jarayonlarini avtomatik yozma bayonnoma (proto) holatiga keltirish, inklyuziv ta'limda - ko'rish talabalarga darsliklarni ovozli o'qib berishni

talab qiladi. Bu davlat tomonidan milliy dasturiy ta'minot va ovozli operatsion tizimlarni qo'llab-quvvatlash uchun yagona ochiq audio-platforma (Open Speech Data Bank) tuzilishi kerak.

Xulosa

O'zbek tili ovozli korpusini tizimli rejalashtirish va uning hisobida infratuzilmasini tubdan isloh qilish milliy tilni hisoblash asrda saqlab qolish hamda zamonaviy sun'iy intellekt texnologiyalari bilan integratsiya qilishning fundamental asosi hisoblanadi. Ovozli bazani kuzatish yo'lidagi eng asosiy lingvistik to'siq bo'lgan hududiy dialektlarning fonetik-akustik farqlari muammolarini ishlab chiqarish ko'p komponentli manba annotatsiya joriy etish orqali samarali hal etish mumkin. Zamonaviy kraudsorsing metodologiyasini sun'iy intellektga zamonaviy intellektual filtrlar bilan birlashtirish minimal moliya va resurs minglab soatlik ochiq ma'lumotlar bazasini talab qiladi va yuqori sifatli ma'lumotlar to'plamini ushlab turadi.

FOYDALANILGAN ADABIYOTLAR

1. Elov B., Primova M., Amirkulov M. O'zbek tili korpusi: Tarixiy rivojlanishi, maqsad va vositalari. Monografiya. – Toshkent: Fan va texnologiya nashriyoti, 2023. – 180 b.
2. Yusupov Sh. Kompyuter lingvistikasida nutqiy korpus kirish va lingvistik annotatsiya asoslari. O'quv qo'llanma. – Farg‘ona: Klassik nashriyoti, 2022. – 215 b.
3. Po'latov A. Kompyuter lingvistikasi (O'zbek tiliga asoslangan materiallar). – Toshkent: Akademnashr, 2011. – 280 b.
4. Rabiner L., Juang B.H. Nutqni aniqlash asoslari. – Prentice Hall, 1993. – 507 b.
5. O'zbek tili elektron korpusi rasmiy platformasi. [Elektron manba]. URL: uzbekcorpus.uz (Murojaat sanasi: 06/08/2026).
6. Mozilla Common Voice Open loyihasi. [Elektron manba]. URL: mozilla.org (Murojaat sanasi: 06/01/2026).
7. Xalqaro ilmiy-amaliy konferensiya materiallari: "Sun'iy intellekt va milliy tillarni raqamlashtirish muammolari". – Toshkent: O‘zMU nashriyoti, 2024. – B. 89-97.
8. Abduaxadov A Neyron vositalari yordamida o'zbekcha nutq signallarini qayta ishlash texnologiyalari // Axborot texnologiyalari va muammolari jurnali. – Samarqand, 2025. – No 3. – B. 42-49.