O‘ZBEK TILINI RAQAMLI TEXNOLOGIYALAR YORDAMIDA SAQLASH VA RIVOJLANTIRISHNING LINGVISTIK MEXANIZMLARI

Qodirova Rayyona Iqboljon qizi
Farg‘ona Davlat Universiteti talabasi
rayyonaqodirova008@gmail.com
Ilmiy rahbar: ***Mamatkodirov Maxammadali Mamatisakovich***
Farg‘ona Davlat Universiteti

Annotatsiya. Mazkur maqolada globallashuv va raqamli transformatsiya sharoitida o‘zbek tilini saqlash, boyitish hamda xalqaro axborot makoniga integratsiya qilishning asosiy lingvistik mexanizmlari tahlil qilingan. Tadqiqotda o‘zbek tilining agglyutinativ tabiatidan kelib chiquvchi morfologik, sintaktik va semantik modellarni tabiiy tilni qayta ishlash (NLP) tizimlariga moslashtirish masalalari yoritilgan. Milliy til korpusini shakllantirish, avtomatik morfologik analizatorlar, mashina tarjimasi, nutqni tanish (ASR) va sintezlash (TTS) texnologiyalarini yaratishning ilmiy-amaliy asoslari ko‘rib chiqilgan.

Kalit so‘zlar: raqamli lingvistika, o‘zbek tili milliy korpusi, tabiiy tilni qayta ishlash (NLP), agglyutinativ tizim, morfologik analizator, mashina tarjimasi, sun’iy intellekt, tilni raqamlashtirish.

Аннотация. В данной статье анализируются основные лингвистические механизмы сохранения, обогащения и интеграции узбекского языка в международное информационное пространство в условиях глобализации и цифровой трансформации. В исследовании освещены вопросы адаптации морфологических, синтаксических и семантических моделей, вытекающих из агглютинативной природы узбекского языка, к системам обработки естественного языка (NLP). Рассмотрены научно-практические основы формирования Национального корпуса языка, создания автоматических морфологических

анализаторов, систем машинного перевода, распознавания (ASR) и синтеза речи (TTS).

Ключевые слова: цифровая лингвистика, Национальный корпус узбекского языка, обработка естественного языка (NLP), агглютинативный строй, морфологический анализатор, машинный перевод, искусственный интеллект, цифровизация языка.

Abstract. This article analyzes the core linguistic mechanisms for preserving, enriching, and integrating the Uzbek language into the global information space amidst globalization and digital transformation. The study highlights the adaptation of morphological, syntactic, and semantic models rooted in the agglutinative nature of the Uzbek language to Natural Language Processing (NLP) systems. Furthermore, it examines the scientific and practical foundations for developing the Uzbek National Corpus, automated morphological analyzers, machine translation tools, Automatic Speech Recognition (ASR), and Text-to-Speech (TTS) technologies.

Keywords: digital linguistics, Uzbek National Corpus, Natural Language Processing (NLP), agglutinative structure, morphological analyzer, machine translation, artificial intelligence, language digitalization.

KIRISH

Zamonaviy raqamli tamaddun sharoitida har bir milliy tilning yashovchanligi va taraqqiyot darajasi uning internet tarmog‘ida, sun’iy intellekt tizimlarida hamda dasturiy ta’minot mahsulotlarida qay darajada faol qo‘llanilishi bilan belgilanadi [1]. Jahon tilshunosligida kechayotgan globallashuv jarayonlari bir tomondan madaniy-axborot almashinuvini tezlashtirsa, ikkinchi tomondan raqamli resurslari yetarli darajada shakllanmagan tillar uchun jiddiy xavf tug‘dirmoqda [1]. Shu nuqtayi nazardan, o‘zbek tilini raqamli texnologiyalar vositasida saqlash, uning boy leksik-grammatik zaxirasini elektron muhitga ko‘chirish va zamonaviy axborot-kommunikatsiya texnologiyalari (AKT) talablariga moslashtirish bugungi kunning eng dolzarb ilmiy-amaliy vazifalaridan biridir [2].

O‘zbekiston Respublikasida davlat tilining nufuzini oshirish va uni raqamlashtirish borasida tizimli islohotlar amalga oshirilmoqda [3]. Xususan, 2020–2030-yillarda o‘zbek tilini rivojlantirish va til siyosatini takomillashtirish konsepsiyasida o‘zbek tilining kompyuter dasturlari, elektron lug‘atlar va milliy korpusini yaratish ustuvor yo‘nalish sifatida belgilab berilgan [3]. Ushbu strategik vazifalarni muvaffaqiyatli amalga oshirish esa eng avvalo tilning ichki tuzilishini matematik va algoritmik modellashtirishga asoslangan lingvistik mexanizmlarni ishlab chiqishni taqozo etadi [4].

Tilni raqamli saqlash deganda nafaqat matnlarni elektron formatga o‘tkazish (skanerlash yoki raqamlashtirish), balki kompyuter tizimlari tomonidan tabiiy tilni anglash, tahlil qilish, sintezlash va qayta ishlash (NLP — Natural Language Processing) mexanizmlarini yaratish tushuniladi [1]. O‘zbek tili turkiy tillar oilasiga mansub agglyutinativ til bo‘lib, uning so‘z yasalishi, affiksal morfologiyasi va gap qurilishi hind-yevropa tillaridan (masalan, ingliz yoki rus tillaridan) tubdan farq qiladi [5]. Shu sababli xorijiy tillar uchun yaratilgan tayyor kompyuter modellarini o‘zbek tiliga to‘g‘ridan-to‘g‘ri tatbiq etib bo‘lmaydi; buning uchun o‘zbek tilining o‘ziga xos lingvistik mexanizmlarini chuqur tadqiq etish zarur [2, 4].

MAVZUNING MAQSADI VA VAZIFALARI

Tadqiqotning asosiy maqsadi — raqamli texnologiyalar va sun’iy intellekt tizimlari orqali o‘zbek tilini saqlash, boyitish hamda global axborot makoniga integratsiya qilishga xizmat qiluvchi asosiy lingvistik mexanizmlarni aniqlash, tizimlashtirish va ularning ilmiy-amaliy modellarini asoslashdan iborat [2, 4].

Mazkur maqsadga erishish uchun tadqiqot doirasida quyidagi ilmiy-amaliy vazifalar belgilab olindi:

1. o‘zbek tilini raqamlashtirish jarayonida yuzaga keladigan fonetik, orfografik va morfologik muammolarni tahlil qilish hamda ularni bartaraf etishning lingvistik yechimlarini ishlab chiqish [5, 6];

2. o‘zbek tilining agglyutinativ tabiatiga mos keluvchi kompyuter morfologiyasi (stemming, lemmatizatsiya va affiksal tahlil) mexanizmlarini ochib berish [2, 4];

3. O‘zbek tili milliy korpusini shakllantirishda matnlarni teglash (annotatsiyalash), meta-ma’lumotlar bazasini yaratish va dialektal-tarixiy qatlamlarni saqlash tamoyillarini asoslash [6, 7];

4. neyron mashina tarjimasi (NMT), nutqni avtomatik tanish (ASR) va nutqni sintezlash (TTS) texnologiyalarida o‘zbek tilining sintaktik va prosodik xususiyatlarini modellashtirish mexanizmlarini ko‘rsatib berish [1, 6];

5. raqamli muhitda terminologik standartlashtirish va neologizmlarni boshqarish orqali tilning leksik-semantik rivojlanish mexanizmlarini taklif etish [3, 7].

MATERIALLAR VA USULLAR

Tadqiqot materiali sifatida O‘zbekiston Respublikasining davlat tilini rivojlantirishga oid me’yoriy-huquqiy hujjatlari [3], O‘zbek tili milliy korpusi (uzbekcorpus.uz) tarkibidagi badiiy, ilmiy, publitsistik va rasmiy matnlar bazasi [7], elektron izohli hamda terminologik lug‘atlar, shuningdek, o‘zbek tili uchun ishlab chiqilgan zamonaviy NLP dasturiy modullari (morfologik analizatorlar va neyron tarjima modellari) o‘rganildi [4, 6]. Tadqiqot davomida tizimli-struktural tahlil, kompyuter-lingvistik modellashtirish, korpus tahlili, qiyosiy-tipologik tasniflash hamda algoritmik formalizatsiya usullaridan foydalanildi [1, 2].

O‘tkazilgan tahlillar shuni ko‘rsatadiki, o‘zbek tilini raqamli texnologiyalar yordamida saqlash va rivojlantirish beshta o‘zaro bog‘liq lingvistik mexanizm asosida amalga oshiriladi [4, 6]. Birinchi mexanizm — grafik va orfografik unifikatsiya mexanizmidir. O‘zbek yozuvida lotin va kirill alifbolarining parallel qo‘llanilishi, shuningdek, o‘zbekcha maxsus harflar (o‘, g‘, sh, ch) va tutuq belgisining (’) kompyuter kodlash tizimlarida (Unicode) turlicha talqin qilinishi ma’lumotlar bazalarida parchalanishni keltirib chiqaradi [6]. Masalan, birgina «o‘» harfi elektron matnlarda apostrof, teskari tirnoq yoki maxsus diakritik belgi ko‘rinishida o‘ndan ortiq har xil kodlar bilan yozilishi matnlarni qidirish va indeksatsiyalash samaradorligini keskin pasaytiradi [6]. Shu bois yagona Unicode standartiga asoslangan transliteratsiya va orfografik normalizatsiya algoritmlari barcha raqamli tizimlarning poydevori hisoblanadi [2, 6].

Ikkinchi mexanizm — agglyutinativ morfologik tahlil va modellashtirish mexanizmi. O‘zbek tilida bitta o‘zakka o‘nlab so‘z yasovchi va shakl yasovchi qo‘shimchalar ketma-ket (zanjirsimon) qo‘shilishi natijasida yuzlab grammatik shakllar hosil bo‘ladi (masalan: kitob — kitoblar — kitoblarimiz — kitoblarimizdagilardan) [5]. Kompyuter lingvistikasida bunday murakkab tuzilmalarni qayta ishlash uchun ikki bosqichli morfologiya (Two-Level Morphology) hamda chekli avtomatlar (Finite-State Transducers — FST) mexanizmi qo‘llaniladi [1, 4]. Bu mexanizm so‘zni o‘zak (lemma) va affikslarga ajratish (lemmatizatsiya), har bir qo‘shimchaning grammatik ma’nosini (kelishik, egalik, zamon, shaxs-son) aniqlash hamda fonetik o‘zgarishlarni (masalan, q → g‘, k → g almashinuvi: qishloq + i = qishlog‘i) avtomatik tiklash imkonini beradi [4, 5].

Uchinchi mexanizm — korpus lingvistikasi va ko‘p qatlamli annotatsiyalash mexanizmi. Tilning leksik boyligi, tarixiy obidalari va hududiy shevalarini kelajak avlodga bezavol yetkazishda Milliy til korpusi hal qiluvchi rol o‘ynaydi [7]. Korpus mexanizmi matnlarni shunchaki saqlash bilan cheklanmay, ularga morfologik (POS-tagging — so‘z turkumlarini belgilash), sintaktik (gap bo‘laklari va sintaktik bog‘lanish daraxtlarini qurish) hamda semantik teglarni biriktiradi [4, 7]. Ayniqsa, o‘zbek tilining qipchoq, qarluq va o‘g‘uz lahjalariga oid audio va matnli ma’lumotlarni geografik meta-ma’lumotlar bilan birga korpusga kiritish yo‘qolib borayotgan dialektal birliklarni raqamli konservatsiya qilishning eng samarali usulidir [7].

To‘rtinchi mexanizm — sintaktik-semantik modellashtirish va neyron mashina tarjimasi (NMT). O‘zbek tilida gap bo‘laklarining odatiy tartibi SOV (Ega — To‘ldiruvchi — Kesim) qolipiga asoslangan bo‘lib, ingliz tilidagi SVO qolipidan farq qiladi [5]. Zamonaviy Transformer arxitekturasiga asoslangan katta til modellari (LLM) va mashina tarjimasi tizimlarini o‘qitishda parallel korpuslar (o‘zbekcha-inglizcha, o‘zbekcha-ruscha va boshqa tillar juftligi) hamda sub-word tokenizatsiya (BPE — Byte Pair Encoding) mexanizmlaridan foydalaniladi [1, 6]. Bu esa o‘zbek tilidagi ko‘p ma’noli so‘zlar (polisemiya) va idiomatik iboralarning kontekstual ma’nosini aniq tarjima qilish imkonini yaratadi [2, 6].

Beshinchi mexanizm — akustik-fonetik va prosodik modellashtirish (ASR va TTS). Tilning jonli nutq shaklida yashovchanligini ta’minlash uchun o‘zbek tilining vokalizm (6 ta unli fonema va ularning pozitsion variantlari) hamda konsonantizm tizimi, urg‘u va intonatsiya qonuniyatlari raqamli akustik modellarga o‘tkaziladi [1, 5]. Nutqni tanish (ASR) va matndan nutq yaratish (TTS) tizimlari imkoniyati cheklangan shaxslar uchun axborotdan foydalanish imkoniyatini kengaytirish bilan birga, o‘zbek tilining adabiy talaffuz me’yorlarini (orfoepiya) ommalashtirishga xizmat qiladi [2, 6]. Ushbu beshta mexanizmning o‘zaro integratsiyasi o‘zbek tilini to‘laqonli raqamli til maqomiga olib chiqishning ilmiy-amaliy asosini tashkil etadi [4, 6].

XULOSA

O‘tkazilgan tadqiqot natijalari shuni tasdiqlaydiki, raqamli texnologiyalar asrida o‘zbek tilini saqlash va rivojlantirish masalasi sof gumanitar doiradan chiqib, fanlararo — kompyuter lingvistikasi va sun’iy intellekt muhandisligi darajasiga ko‘tarilgan [1, 4]. O‘zbek tilining agglyutinativ tabiati unga yuqori darajadagi algoritmik muntazamlik bag‘ishlaydi; bu esa to‘g‘ri ishlab chiqilgan lingvistik mexanizmlar yordamida tilni kompyuter tizimlarida yuqori aniqlik bilan modellashtirish imkonini beradi [2, 5].

Xulosa o‘rnida ta’kidlash joizki, birinchidan, o‘zbek tilidagi barcha raqamli resurslar uchun yagona Unicode kodlash va orfografik normalizatsiya standartlarini joriy etish til ma’lumotlarining parchalanishini bartaraf etadi [6]. Ikkinchidan, O‘zbek tili milliy korpusini morfologik, sintaktik va dialektal qatlamlar bilan uzluksiz boyitib borish tilning tarixiy va hududiy rang-barangligini saqlab qolishning eng ishonchli kafolatidir [3, 7]. Uchinchidan, o‘zbek tiliga moslashtirilgan neyron tarjima, nutqni tanish va sintezlash modellarini keng joriy etish orqali o‘zbek tilining ilm-fan, texnologiya va xalqaro raqamli muloqot tili sifatidagi mavqeyini yanada mustahkamlash mumkin [1, 6].

FOYDALANILGAN ADABIYOTLAR

[1] Jurafsky D., Martin J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. — 3rd ed. — Stanford University, 2024. — 648 p.

[2] Po‘latov A.K. Kompyuter lingvistikasi: o‘quv qo‘llanma. — Toshkent: Akademnashr, 2011. — 520 b.

[3] O‘zbekiston Respublikasi Prezidentining 2020-yil 20-oktabrdagi «Mamlakatimizda o‘zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to‘g‘risida»gi PF-6084-son Farmoni // Qonun hujjatlari ma’lumotlari milliy bazasi.

[4] Abduraxmonova N.Z. Kompyuter lingvistikasi asoslari: darslik. — Toshkent: Nodirabegim, 2021. — 240 b.

[5] Rahmatullayev Sh. Hozirgi adabiy o‘zbek tili: darslik. — Toshkent: Universitet, 2006. — 464 b.

[6] Elov B.B., Hamroyeva Sh.M., Xusainova Z.Y. O‘zbek tilining kompyuter modellari va raqamli resurslarini yaratish muammolari // O‘zbek tili va adabiyoti. — 2022. — № 4. — B. 45–52.

[7] Hamroyeva Sh.M. O‘zbek tili mualliflik korpusini tuzishning lingvistik asoslari: filol. fan. bo‘yicha falsafa doktori (PhD) dis. avtoref. — Buxoro, 2018. — 48 b.