KOMPYUTER LINGVISTIKASI: SUN'IY INTELLEKT VA TABIIY TILNI QAYTA ISHLASH (NL)

Abdullayeva Barno Baxromjon qizi
Andijon davlat chet tillar instituti
Andijon davlat chet tillari institute Xorijiy filologiya va turizm fakulteti
Filologiya va tillarni o`qitish koreys tili 1-bosqich talabasi
Abdullayeva92@gmail.com

Annotatsiya. Mazkur maqola kompyuter lingvistikasining sun'iy intellekt davridagi o'rni va tabiiy tilni qayta ishlash (NLP) texnologiyalarining lingvistik asoslarini aniqlashga bag'ishlangan. Tadqiqotda NLP usullarining qoidaga asoslangan, statistik va neyron bosqichlari qiyosiy yoritilib, o'zbek tilining agglyutinativ tabiati bu texnologiyalar uchun qanday imkoniyat va qiyinchiliklar tug'dirishi ko'rsatiladi. Muallif so'nggi yillarda e'lon qilingan o'zbek tili resurslarini (teglangan ma'lumotlar, daraxtbanka, nutq korpusi, tarjima to'plamlari, til modellari) lingvistik darajalar bo'yicha tizimlashtirib, ularning yetarliligini baholash uchun to'rt mezon taklif etadi. Olingan natijalar o'zbek tili uchun til texnologiyalarini rivojlantirishda hamda tilshunoslik yo'nalishidagi o'quv dasturlarini yangilashda qo'llanilishi mumkin.

Kalit so'zlar: kompyuter lingvistikasi, tabiiy tilni qayta ishlash, sun'iy intellekt, neyron tarmoq, til modeli, morfologik tahlil, o'zbek tili.

Аннотация. Статья посвящена определению места компьютерной лингвистики в эпоху искусственного интеллекта и выявлению лингвистических основ технологий обработки естественного языка (NLP). В работе сопоставлены правилоориентированный, статистический и нейросетевой этапы развития NLP и показано, какие возможности и трудности создаёт для этих технологий агглютинативный строй узбекского языка. Автор систематизирует опубликованные в последние годы ресурсы узбекского языка (размеченные данные, трибанк, речевой корпус, переводческие наборы, языковые модели) по уровням языка и предлагает четыре критерия оценки их достаточности. Результаты могут быть использованы при развитии языковых технологий для узбекского языка и обновлении учебных программ по языкознанию.

Ключевые слова: компьютерная лингвистика, обработка естественного языка, искусственный интеллект, нейронная сеть, языковая модель, морфологический анализ, узбекский язык.

Abstract. This article examines the place of computational linguistics in the age of artificial intelligence and the linguistic foundations of natural language processing (NLP). The rule-based, statistical and neural stages of NLP are compared, and the opportunities and difficulties that the agglutinative nature of Uzbek creates for these technologies are shown. The author systematizes Uzbek-language resources published in recent years (annotated data, a treebank, a speech corpus, translation datasets, language models) by linguistic level and proposes four criteria for assessing their sufficiency. The results can be used to develop language technologies for Uzbek and to update linguistics curricula.

Keywords: computational linguistics, natural language processing, artificial intelligence, neural network, language model, morphological analysis, Uzbek language.

Kirish. Sun'iy intellekt texnologiyalari so'nggi yillarda inson va kompyuter muloqotini tubdan o'zgartirdi: matnni boshqa tilga o'giradigan, savolga javob beradigan, nutqni yozuvga aylantiradigan

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

tizimlar kundalik hayotning oddiy qismiga aylandi. Bu tizimlarning barchasi til bilan ishlaydi. Shu sababli tabiiy tilni qayta ishlash (inglizcha Natural Language Processing, NLP) hisoblash fanlari bilan tilshunoslik tutashgan nuqtada rivojlanadi, uning nazariy asosini esa kompyuter lingvistikasi tashkil etadi. Kompyuter lingvistikasi til hodisalarini hisoblash modellari orqali tavsiflaydi, NLP esa bu modellarni amaliy dasturlarga aylantiradi.

Dunyodagi 7000 dan ortiq tilning juda kichik qismi til texnologiyalarida aks etgan, tillar o'rtasidagi bu tengsizlik ayniqsa resurslar jihatidan yaqqol ko'rinadi (Joshi va b., 2020). O'zbek tili ham resurslari cheklangan tillardan biri. Bobojonova va hammualliflarining ma'lumotiga ko'ra, o'zbek tili taxminan 40 million kishining ona tili bo'lib, turkiy tillar orasida turk tilidan keyin ikkinchi o'rinda turadi, ammo NLP resurslari jihatidan kam ta'minlangan til hisoblanadi (Bobojonova va b., 2025). Uning agglyutinativ tuzilishi, ya'ni so'z shakllarining qo'shimchalar zanjiri orqali hosil bo'lishi, tilshunoslik bilimini bu sohada ayniqsa muhim qiladi.

Tadqiqotning maqsadi kompyuter lingvistikasi va NLP o'rtasidagi aloqani sun'iy intellekt davri nuqtai nazaridan yoritish hamda o'zbek tili uchun mavjud resurslarni lingvistik darajalar bo'yicha tizimlashtirishdan iborat. Quyidagi vazifalar belgilandi: NLP rivojlanish bosqichlarida tilshunoslikning o'rnini aniqlash; o'zbek tilining hisoblash nuqtai nazaridan o'ziga xos qiyinchiliklarini tahlil qilish; 2021–2026-yillarda e'lon qilingan resurslarni tizimlashtirish; amaliy takliflar ishlab chiqish.

Mavzuga oid adabiyotlarning tahlili. Hisoblash yordamida tilni qayta ishlash g'oyasi XX asr o'rtalarida shakllandi. A. Tyuring 1950-yilgi maqolasida mashina fikrlay oladimi degan savolni til orqali muloqot sinovi bilan bog'ladi. 1954-yilda Jorjtaun universiteti va IBM hamkorligidagi tajribada oltmishdan ortiq rus tilidagi gap inglizchaga mashina yordamida o'girildi va bu soha uchun katta umid uyg'otdi (Hutchins, 2004). Biroq 1966-yilda e'lon qilingan ALPAC hisoboti mashina tarjimasining o'sha davrdagi natijalarini qoniqarsiz deb baholab, moliyalashtirishning qisqarishiga ta'sir ko'rsatdi. Shu yili J. Vaytsenbaum yaratgan ELIZA dasturi esa oddiy andozalarga tayangan holda ham suhbat tuyg'usini hosil qilish mumkinligini ko'rsatdi.

1990-yillarda statistik yondashuv ustun bo'ldi; K. Manning va H. Shyutse kitobi bu yo'nalishning asosiy qo'llanmalaridan biriga aylandi (Manning, Schütze, 1999). 2013-yilda T. Mikolov va hammualliflari so'zlarni zich vektorlar ko'rinishida tasvirlash usulini taklif qildi: bunda ma'nosi yaqin so'zlar vektor fazoda ham yaqin joylashadi. 2016-yilda R. Sennrich va hamkorlari noyob so'zlarni so'z osti birliklariga bo'lish orqali tarjima sifatini oshirdi. 2017-yilda A. Vaswani va hammualliflari diqqat mexanizmiga asoslangan Transformer arxitekturasini e'lon qildi; BERT (Devlin va b., 2019) va GPT-3 (Brown va b., 2020) kabi modellar shu arxitekturaga tayanadi. 2022-yilda NLLB jamoasi ikki yuz tilni qamrab oluvchi tarjima modelini taqdim etdi.

Lingvistik nuqtai nazardan bu yutuqlar bahs uyg'otadi. E. Bender va A. Koller faqat shaklga tayanib o'qitilgan tizim ma'noni o'zlashtirishi haqidagi da'volar asosli bo'lishi kerakligini ta'kidlagan; keyinroq Bender va hammualliflari katta til modellarining xarajati va xavflarini muhokama qilgan. Darsliklar orasida D. Jurafski va Dj. Martinning "Speech and Language Processing" kitobi neyron usullarni ham qamrab oladi. Rus tilidagi adabiyotda korpus lingvistikasining asoslari tizimli bayon etilgan (Zaxarov, Bogdanova, 2020). O'zbek tilshunosligida so'z turkumlari va morfologiya an'anaviy grammatikalarda batafsil tavsiflangan (Mahmudov, Nurmonov, 1995; Rahmatullayev, 2006); bu tavsiflar hisoblash modellari uchun zarur lingvistik asosdir.

So'nggi yillarda o'zbek tili uchun resurslar tez ko'paydi: BERTbek til modeli (Kuriyozov va b., 2024), UPOS teglari asosidagi benchmark (Bobojonova va b., 2025), nomlangan obyektlar uchun

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

ma'lumotlar to'plami (Mengliev va b., 2025), ochiq nutq korpusi USC (Musaev va b., 2021), janubiy o'zbek tili tarjima resurslari (Mamasaidov va b., 2025), UzUDT daraxtbankasi (Matlatipov, Aripov, 2026) va til modellarini sinovdan o'tkazuvchi UzLiB benchmarki shular jumlasidandir. Biroq ko'rib chiqilgan manbalar doirasida bu ishlar asosan alohida vazifalar sifatida yoritilgan; ularni lingvistik darajalar kesimida yagona tizimga solib, yetarlilik mezonlari bo'yicha baholash masalasi alohida o'rganilmagan. Mazkur maqola ana shu bo'shliqni to'ldirishga qaratilgan.

Tadqiqot metodologiyasi. Tadqiqotda tarixiy-tavsifiy, qiyosiy va kontent-tahlil usullaridan foydalanildi. Tarixiy-tavsifiy usul NLP bosqichlarini ko'rib chiqishda, qiyosiy usul qoidaga asoslangan, statistik va neyron yondashuvlarni solishtirishda, kontent-tahlil esa o'zbek tiliga oid ilmiy nashrlar va ochiq resurslarni o'rganishda qo'llandi. Resurslarning yetarliligini baholash uchun to'rtta mezon belgilandi: (1) lingvistik qamrov (resurs qaysi til darajasini, ya'ni fonetika, morfologiya, sintaksis, semantika yoki pragmatikani qamrab oladi); (2) ochiqlik (resurs tadqiqotchilar uchun ommaga ochiqmi); (3) baholash imkoniyati (natijalarni solishtirish uchun umumiy ko'rsatkich yoki benchmark bormi); (4) hajm va xilma-xillik (janr, yozuv turi va til variantlari yetarlicha berilganmi). Baholash sifat jihatidan, nashrlarda keltirilgan ma'lumotlarga tayangan holda amalga oshirildi.

Tahlil va natijalar. NLP tarixini uch bosqichga bo'lish mumkin. Birinchi bosqichda (taxminan 1950–1980-yillar) tizimlar tilshunoslar yozgan qoidalar va lug'atlarga tayandi: grammatika qo'lda kodlashtirildi. Ikkinchi bosqichda (1990-yillardan) qoidalar o'rnini katta korpuslardan hisoblangan ehtimollar egalladi, tilshunosning vazifasi esa ma'lumotni tayyorlash va teglashga ko'chdi. Uchinchi bosqichda (2010-yillardan) neyron tarmoqlar va Transformer asosidagi modellar ustun keldi. Bunda ham tilshunoslik ahamiyatini yo'qotmadi: modelni o'qitish uchun korpus, uni baholash uchun test to'plami, natijani izohlash uchun esa til haqidagi bilim kerak bo'ladi.

NLP tizimi odatda ketma-ket bosqichlardan iborat bo'lib, ularning har biri tilshunoslikning muayyan bo'limiga mos keladi. Nutq bilan ishlaydigan tizimlarda fonetika va fonologiya bilimi tovushni yozuvga yoki yozuvni tovushga aylantirish uchun kerak. Matn bilan ishlashda avval matn so'zlarga ajratiladi (tokenizatsiya), so'ng har bir so'zning o'zagi va qo'shimchalari aniqlanadi (morfologik tahlil), undan keyin gapdagi so'zlar o'rtasidagi bog'lanish topiladi (sintaktik tahlil). Yuqori darajada so'z va gapning ma'nosi (semantika) hamda gapning vaziyatga bog'liq ma'nosi (pragmatika) aniqlanadi. O'zbek tilida bu zanjirning dastlabki bo'g'inlari, ayniqsa tokenizatsiya va morfologik tahlil, boshqa ko'pchilik tillarga qaraganda ko'proq og'irlik qiladi, chunki so'z chegarasi va qo'shimchalar ma'noning katta qismini tashiydi.

O'zbek tilining agglyutinativ tabiati NLP uchun bir necha qiyinchilik tug'diradi. Birinchidan, bitta o'zakdan juda ko'p shakl hosil bo'ladi. Masalan, "kitoblarimizdagilarga" so'zi bitta o'zak (kitob) va oltita qo'shimchadan (-lar, -imiz, -da, -gi, -lar, -ga) iborat bo'lib, ingliz tilida "to those in our books" kabi butun bir birikmaga mos keladi. Bunday shakllarning aksariyati matnda kam uchraydi, shuning uchun so'zni butunligicha lug'atga kiritadigan tizim ko'plab noma'lum so'zlarga duch keladi. Muammoni qisman so'zni statistik bo'laklarga ajratish usullari yengillashtiradi (Sennrich va b., 2016), biroq bunday bo'laklar har doim morfemalar chegarasiga to'g'ri kelavermaydi. Bobojonova va hammualliflari ham so'z osti bo'laklash tufayli juda ko'p qo'shimchali so'zlarni yagona birlik sifatida teglash qiyinligini cheklov sifatida qayd etgan. Shu sababli tilshunoslikka asoslangan morfologik tahlilning ahamiyati saqlanib qolmoqda.

Ikkinchidan, o'zbek tilida ikki yozuv, lotin va kirill yozuvi qo'llanadi. Lotin yozuviga asoslangan alifbo 1993-yilda qonun bilan joriy etilgan, biroq kirill yozuvidagi matnlar hamon keng tarqalgan. Bobojonova va hammualliflari ikki yozuv uchun alohida modellarni sinab ko'rdi: so'z

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

turkumini aniqlashda ko'p tilli mBERT modeli lotin yozuvidagi matnda 86,0 foiz, kirill yozuvida esa 80,2 foiz aniqlikka erishgan, ya'ni kirill yozuvidagi o'zbek matni bu modelda yomonroq ifodalangan. Lotin yozuvida esa o' va g' harflaridagi modifikator belgi hamda tutuq belgisi (') tokenizatsiyada so'zni noto'g'ri bo'lib yuborgan. Demak, yozuvni me'yorlashtirish masalasi tashqaridan imlo masalasiga o'xshasa-da, NLP uchun bevosita texnik ahamiyatga ega.

Resurslarni lingvistik darajalar bo'yicha tizimlashtirish natijalari 1-jadvalda keltirilgan.

O'zbek tili uchun NLP resurslarining lingvistik darajalar bo'yicha tasnifi

Vazifa	Lingvistik daraja	Resurs (muallif, yil)	Asosiy ko'rsatkich
So'z turkumini teglash	Morfologiya	BBPOS (Bobojonova va b., 2025)	500 gap, 5831 so'z; o'rtacha aniqlik 91%
Sintaktik tahlil	Sintaksis	UzUDT (Matlatipov, Aripov, 2026)	684 gap, 7582 token; kelishuv > 0,90
Nomlangan obyektlarni aniqlash	Leksika, semantika	NER ma'lumotlar to'plami (Mengliev va b., 2025)	Ma'lumotlar to'plami va neyron model
Til modeli	Barcha darajalar	BERTbek (Kuriyozov va b., 2024)	Uch vazifada mBERT'dan yuqori
Nutqni aniqlash	Fonetika	USC (Musaev va b., 2021)	105 soat, 958 diktor; WER 17,4%
Mashina tarjimasi	Semantika	Janubiy o'zbek resurslari (Mamasaidov va b., 2025)	39 994 parallel gap; 997 gaplik to'plam
Til modellarini baholash	Grammatika, imlo	UzLiB benchmarki	Tanlov javobli savollar to'plami

Jadvaldan ko'rinadiki, morfologiya va nutq darajalari uchun resurslar nisbatan ko'proq. So'z turkumini teglashda neyron modellar qoidaga asoslangan dastur (UzbekTagger, Sharipov va b., 2023) dan yuqori natija bergan: qoidali tagger 75,6 foiz, bir tilli modellar esa 90,9 va 91,6 foiz aniqlikka erishgan. Farq lingvistik jihatdan ham ko'rinadi: "Tortmani tortma" gapida birinchi so'z ot, ikkinchisi fe'l bo'lishi kerak. Qoidali tagger ikkalasini ham ot deb belgilagan, mBERT ham xato qilgan, bir tilli modellar esa to'g'ri aniqlagan. Biroq bu natijalar kichik ma'lumot to'plamiga tayanadi: modellar 17 ta universal turkumdan beshtasini ular kam uchrashi sababli o'rgana olmagan.

Sintaksis darajasida UzUDT daraxtbankasi 684 gap (7582 token)dan iborat bo'lib, annotatorlar kelishuvi lemma, so'z turkumi va morfologik belgilar bo'yicha 0,90 dan yuqori. Mualliflar graf asosidagi (Stanza) va o'tish asosidagi (spaCy) tahlilchilarni solishtirib, birinchisi murakkab bog'lanishlarni aniqlashda, ikkinchisi morfosintaktik teglashda kuchliroq ekanini ko'rsatgan; boshqa daraxtbanka ma'lumotlarini qo'shish esa aniqlikni sezilarli oshirgan. Nutq darajasida USC korpusi CC BY 4.0 litsenziyasi ostida tarqatiladi va dastlabki tajribalarda test to'plamida so'z xatolari ulushi 17,4 foizni tashkil etgan (Musaev va b., 2021).

Ma'lumot sifati masalasi ham tilshunoslik nuqtai nazaridan e'tiborga loyiq. Bobojonova va hammualliflari tajribasida lotin yozuvidagi TahrirchiBERT taxminan 5 milliard tokenli matnda o'qitilgan, biroq bu matn optik belgilarni tanish (OCR) xatolari tufayli shovqinli bo'lgan; kirill

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

yozuvidagi UzBERT esa taxminan 142 million so‘zlik sifatli matnda o‘qitilgan va ancha kichik hajmga qaramay biroz yuqori natija bergan (91,6 foizga qarshi 90,9 foiz). Mualliflar bu farqni ma’lumot sifati va noto‘g‘ri oldindan bo‘laklash bilan izohlaydi. BERTbek ustidagi tadqiqotda ham o‘qitish ma’lumotlarining hajmi va sifati natijaga ta’sir qilishi ko‘rsatilgan (Kuriyozov va b., 2024). Demak, korpus tuzishda matnni tozalash, imloviy me’yorga keltirish va janr xilma-xilligini ta’minlash faqat texnik emas, tilshunoslik vazifasi hamdir.

Mashina tarjimasida Mamasaidov va hammualliflari Afg‘onistonda taxminan 5 million kishi so‘zlashadigan janubiy o‘zbek varianti uchun parallel gaplar to‘plamini yaratib, NLLB-200 modelini shu ma’lumotlar bilan qayta o‘qitdi. Ularning xulosasiga ko‘ra, katta til modellari janubiy o‘zbekchani tushunishda yaxshi natija beradi, biroq shu variantda matn hosil qilish uchun maxsus o‘qitish zarur. Bu o‘zbek tilining ham bir xil massiv emasligini, variantlar alohida resurs talab qilishini ko‘rsatadi. UzLiB benchmarkida savollar o‘zbek tilshunosligiga oid Telegram kanallaridagi viktorinalardan olingan; bu model sifatini baholash uchun me’yoriy tilni biladigan ekspert zarurligini bildiradi. Shu bilan birga, Bender va Koller ta’kidlaganidek, shakl bo‘yicha yaxshi natija ma’noning “tushunilgani”ni isbotlamaydi, shuning uchun benchmark natijalarini ehtiyotkorlik bilan izohlash lozim.

Teglash qo‘llanmasi masalasi ham lingvistik qaror talab qiladi. An’anaviy o‘zbek grammatikasida 12 ta so‘z turkumi ajratiladi, Bobojonova va hammualliflari esa 17 ta universal turkumdan (UPOS) foydalanib, ular o‘rtasidagi moslik jadvalini tuzgan. Yuzdan ortiq tegdan iborat kengaytirilgan morfologik tag to‘plamlari ham taklif etilgan (Abdullayeva va b., 2022; Sharipov va b., 2022), biroq ular asosidagi teglangan ma’lumotlar hozircha mavjud emas.

To‘rt mezon bo‘yicha jamlaganda, qamrov jihatidan morfologiya, sintaksis va nutq darajalari yoritilgan, semantika qisman (tarjima, nomlangan obyektlar) qamrab olingan, ko‘rib chiqilgan manbalar doirasida pragmatika va diskurs uchun ochiq benchmark aniqlanmadi. Holbuki murojaat shakllari (sen, siz) kabi hodisalar o‘zbek tilida muhim pragmatik ma’no tashiydi va dialog tizimlari uchun alohida ma’lumot talab etadi. Ochiqlik jihatidan mualliflar ko‘pincha ma’lumotlarni ommaga ochiq e’lon qilgan; baholash jihatidan aniqlik va WER kabi umumiy ko‘rsatkichlar mavjud; hajm jihatidan esa resurslar hali kichik. Tilshunosning o‘rni resurslarni yaratishda ayniqsa ko‘rinadi: teglash qo‘llanmasini ishlab chiqish, annotatsiya, annotatorlar kelishuvini ta’minlash va xatolarni tahlil qilish aynan lingvistik malaka talab etadi.

Xulosa va takliflar. Tadqiqot shuni ko‘rsatdiki, sun’iy intellekt davrida kompyuter lingvistikasi ahamiyatini yo‘qotmagan: neyron modellarni o‘qitish va baholash uchun sifatli lingvistik ma’lumot kerak. O‘zbek tili uchun morfologiya, sintaksis va nutq darajalarida dastlabki ochiq resurslar yaratilgan, biroq ularning hajmi kichik, semantika va ayniqsa pragmatika darajalari kam yoritilgan.

Tadqiqot natijalariga tayanib quyidagi takliflarni bildirish mumkin. Birinchidan, universal teglar (UPOS) va an’anaviy o‘zbek grammatikasidagi so‘z turkumlari o‘rtasidagi moslikni rasmiy qo‘llanma sifatida mustahkamlash va yirik teglangan korpuslar yaratish zarur. Ikkinchidan, o‘ va g‘ harflari hamda tutuq belgisini yagona kodlash qoidasi va lotin–kirill transliteratsiyasi uchun me’yoriy ko‘rsatma ishlab chiqish kerak. Uchinchidan, morfemalarga asoslangan va statistik so‘z osti bo‘laklash usullarini o‘zbek tili matnlarida solishtiruvchi tajribalar o‘tkazish lozim. To‘rtinchidan, pragmatika, murojaat shakllari va diskursni qamrab oluvchi test to‘plamlari yaratish hamda filologiya yo‘nalishlari o‘quv dasturlariga korpus tuzish va annotatsiya amaliyotini kiritish maqsadga muvofiq.

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

Foydalanilgan adabiyotlar ro'yxati

1. Abdullayeva O., Uzokova M., Elov B. O'zbek tilida POS tegging masalasi: muammo va takliflar // Uzbekistan: Language and Culture. – 2022. – Vol. 2, No. 5. – P. 51–68.
2. Mahmudov N., Nurmonov A. O'zbek tilining nazariy grammatikasi. – Toshkent: O'qituvchi, 1995.
3. Rahmatullayev Sh. Hozirgi adabiy o'zbek tili. – Toshkent: Universitet, 2006.
4. O'zbekiston Respublikasining "Lotin yozuviga asoslangan o'zbek alifbosini joriy etish to'g'risida"gi Qonuni, 1993-yil 2-sentabr, 931-XII-son (1995-yil 6-mayda o'zgartishlar kiritilgan). – lex.uz/docs/-112286.
5. Kuriyozov E., Vilares D., Gómez-Rodríguez C. BERTbek: A Pretrained Language Model for Uzbek // Proceedings of SIGUL 2024 @ LREC-COLING 2024. – Torino: ELRA and ICCL, 2024. – P. 33–44.
6. Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing. – Cambridge (Mass.): MIT Press, 1999.
7. Matlatipov S.G., Aripov M. UzUDT: Uzbek Universal Dependencies Treebank // Proceedings of LREC 2026. – Palma de Mallorca: ELRA, 2026. – P. 11642–11649.
8. Mengliev D., Barakhnin V., Eshkulov M., Ibragimov B., Madirimov S. A comprehensive dataset and neural network approach for named entity recognition in the Uzbek language // Data in Brief. – 2025. – Vol. 58. – Art. 111249.
9. Sharipov M., Mattiev J., Sobirov J., Baltayev R. Creating a morphological and syntactic tagged corpus for the Uzbek language // CEUR Workshop Proceedings. – 2022. – Vol. 3315. – P. 93–98.
10. Sharipov M., Kuriyozov E., Yuldashov O., Sobirov O. UzbekTagger: The rule-based POS tagger for Uzbek language. – arXiv:2301.12711, 2023.

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