INNOVATIVE WORLD
Ilmiy tadqiqotlar markazi

ZAMONAVIY ILM-FAN VA INNOVATSIYALAR NAZARIYASI

ILMIY-AMALIY KONFERENSIYA

2026

Google Scholar doi zenodo OpenAIRE

Andijan, Uzbekistan +998335668868 https://innoworld.net

« ZAMONAVIY ILM-FAN VA INNOVATSIYALAR NAZARIYASI » NOMLI ILMIY, MASOFAVIY, ONLAYN KONFERENSIYASI TO‘PLAMI

3-JILD 5-SON

Konferensiya to’plami va tezislar quyidagi xalqaro ilmiy bazalarda indexlanadi

www.innoworld.net

O‘ZBEKISTON-2026

Tel:+99833 5668868 | Telegram: @Anvarbek_PhD

IJTIMOIY TARMOQLAR KORPUSI ASOSIDA O'ZBEK TILINING ZAMONAVIY LEKSIK TIZIMINI MODELLASHTIRISHNING KOMPUTER-LINGVISTIK ASOSLARI

Toshtemirova Gulnoraxon Baxtiyor qizi
KIUF universiteti o'qituvchisi, Kokand University tayanch doktoranti
toshtemirovagulnora2909@gmail.com

ANNOTATSIYA
Ushbu maqolada ijtimoiy tarmoqlar asosida shakllantirilgan korpus yordamida o'zbek tilining zamonaviy leksik tizimini komputer-lingvistik jihatdan modellashtirish masalalari ko'rib chiqiladi. Tadqiqot davomida Facebook, Telegram va Instagram kabi ijtimoiy tarmoqlardagi o'zbek tilidagi matnlar to'plami tahlil qilinib, ushbu platformalarda qo'llaniladigan yangi so'z birikmalari, so'z yasalishi jarayonlari va leksik innovatsiyalar aniqlangan. Korpus lingvistikasi va tabiiy tilni qayta ishlash (NLP) metodlaridan foydalangan holda leksik birliklarning tarqalish naqshlari, semantik o'zgarishlar hamda yangi so'zlarning tildagi o'rni tadqiq etilgan. Tadqiqot natijasida 3 740 ta yangi leksik birlik qayd etilib, o'zbek tilidagi ijtimoiy tarmoq matnlarining o'ziga xos leksik qatlami mavjudligi aniqlandi va bu qatlam an'anaviy adabiy til leksikasidan sezilarli darajada farq qilishi ko'rsatildi. Olingan natijalar o'zbek tili uchun zamonaviy elektron lug'at va til modellarini yaratishda muhim nazariy va amaliy asos bo'lib xizmat qilishi mumkin.

Kalit so'zlar: korpus lingvistikasi, leksik tizim, ijtimoiy tarmoqlar, komputer lingvistikasi, tabiiy tilni qayta ishlash, leksik innovatsiya, so'z yasalishi, semantik o'zgarish, til modellash, o'zbek tili.

ABSTRACT
This article examines the issues of computer-linguistic modeling of the modern lexical system of the Uzbek language using a corpus formed on the basis of social media platforms. During the research, a collection of texts in the Uzbek language from social networks such as Facebook, Telegram, and Instagram was analyzed, and new word combinations, word formation processes, and lexical innovations used on these platforms were identified. Using corpus linguistics and natural language processing (NLP) methods, the distribution patterns of lexical units, semantic changes, and the role of new words in the language were studied. As a result of the research, 3,740 new lexical units were recorded, the distinctive lexical layer of Uzbek social media texts was identified, and it was shown that this layer differs significantly from traditional literary language lexicon. The obtained results can serve as an important theoretical and practical basis for creating modern electronic dictionaries and language models for the Uzbek language.

Keywords: corpus linguistics, lexical system, social media, computer linguistics, natural language processing, lexical innovation, word formation, semantic change, language modeling, Uzbek language.

1. KIRISH

So'nggi o'n yillikda raqamli kommunikatsiya texnologiyalarining jadal rivojlanishi tilshunoslik fani oldiga yangi va dolzarb vazifalarni qo'ydi. Ijtimoiy tarmoqlar – Facebook, Telegram, Instagram, YouTube va boshqa platformalar – millionlab foydalanuvchilar o'rtasida kunlik muloqotning asosiy maydoniga aylandi. O'zbekistonda ham bu jarayon shiddat bilan kechmoqda: 2023-yilgi ma'lumotlarga ko'ra, mamlakatda 20 milliondan ortiq internet foydalanuvchisi mavjud bo'lib, ularning muhim qismi ijtimoiy tarmoqlarda faol ishtirok etadi [1].[1]

Ijtimoiy tarmoqlarda yaratilgan matnlar tilning jonli, so'zlashuv shaklini aks ettirib, an'anaviy yozma nutqdan ko'p jihatdan farq qiladi. Bu matnlarda yangi so'zlar, o'ziga xos iboralar, qisqartmalar va yangi grammatik konstruksiyalar keng qo'llaniladi. Bunday hodisalar nafaqat lingvistik, balki madaniy, ijtimoiy va texnologik jarayonlarning natijasidir. Shuning uchun ham ijtimoiy tarmoq matnlari o'zbek tilining bugungi holatini o'rganishda noyob va boy manba bo'lib xizmat qiladi.

Komputer lingvistikasi va tabiiy tilni qayta ishlash (NLP) sohasidagi so'nggi yutuqlar tilshunoslikda mutlaqo yangi imkoniyatlarni ochdi [2].[2] Korpus lingvistikasi metodlari yordamida katta hajmdagi matnlarni tahlil qilish, leksik birliklarning tarqalish naqshlarini aniqlash va semantik o'zgarishlarni kuzatish mumkin bo'ldi. Bundan tashqari, vektorli so'z ifodalari (word embeddings) va transformerga asoslangan til modellari tilshunoslik tadqiqotlarini yangi bosqichga olib chiqdi.

O'zbek tili uchun bunday tadqiqotlar hali dastlabki bosqichida turibdi. Mavjud ishlarning aksariyati adabiy tilga yoki jurnalistik matnlarga asoslanadi. Ijtimoiy tarmoqlardagi o'zbek tilidagi matnlarni korpusga kiritib, leksik jihatdan tahlil qilish borasida keng ko'lamli sistemali izlanishlar amalga oshirilmagan. Holbuki, aynan shu matnlar o'zbek tilining zamonaviy rivojlanish tendensiyalarini, yangi so'z yasalish jarayonlarini va semantik siljishlarni to'liq namoyon etadi.

Ushbu maqolaning asosiy maqsadi – ijtimoiy tarmoqlar korpusi asosida o'zbek tilining zamonaviy leksik tizimini komputer-lingvistik metodlar yordamida modellashtirish tamoyillarini asoslash va bu boradagi dastlabki tadqiqot natijalarini ilmiy muomalaga kiritishdir. Tadqiqot vazifalari quyidagilardan iborat: (1) o'zbek tilidagi ijtimoiy tarmoq matnlari korpusini shakllantirishning metodologik asoslarini belgilash; (2) korpusdagi leksik innovatsiyalarni aniqlash va toifalash; (3) vektorli so'z ifodasi modellari yordamida leksik munosabatlarni

[1]O'zbekistonda internet foydalanuvchilar soni haqida batafsil ma'lumot: Digital 2023: Uzbekistan (We Are Social & Hootsuite, 2023). Hisobot mamlakatda 22 milliondan ortiq internet foydalanuvchisi borligini qayd etadi.

[2]NLP – Natural Language Processing (tabiiy tilni qayta ishlash) – kompyuter algoritmlaridan foydalanib, insoniy tilni avtomatik tarzda tahlil qilish, tushunish va generatsiya qilish bilan shug'ullanuvchi ilmiy-amaliy soha.

modellashtirish; (4) an'anaviy va raqamli leksika o'rtasidagi tizimli farqlarni tahlil etish.

Tadqiqotning ilmiy yangiligi shundan iboratki, o'zbek tilshunosligi tarixida birinchi marta ijtimoiy tarmoq matnlari asosida maxsus korpus shakllantirilib, zamonaviy NLP metodologiyasi tatbiq etilgan. Amaliy ahamiyati esa o'zbek tili uchun yangi leksikografik manbalar, til modellari va tabiiy tilni qayta ishlash vositalarini yaratishda bevosita foydalanish imkoniyatida ifodalanadi.

2. MATERIALLAR VA METODLAR

2.1. Korpus shakllantirishning uslubiy asoslari

Tadqiqotning empirik negizini 2021–2024-yillar davomida ijtimoiy tarmoqlardan to'plangan o'zbek tilidagi matnlar korpusi tashkil etadi. Korpusni shakllantirishda J. Sinclair tomonidan asoslangan korpus lingvistikasining tamoyillaridan foydalanildi [4]. Xususan, quyidagi mezonlarga qat'iy amal qilindi: representativlik (turli mavzular, janrlar va foydalanuvchi guruhlarini qamrab olish); autentiklik (tabiiy, tahrirlanmagan nutq namunalarini kiritish); kengayuvchanlik (yangi matnlar bilan to'ldirish imkoniyati); qonuniylik (ma'lumot to'plashda shaxsiy ma'lumotlarni himoya qilish va etik me'yorlarga amal qilish).

Korpus quyidagi platformalardagi matnlardan tashkil topdi: Facebook – O'zbekistondagi eng faol jamoat guruhlari va sahifalar (fan, madaniyat, iqtisod, sport, siyosat sohalari bo'yicha); Telegram – ko'p obunachili o'zbekcha kanallar va guruhlardagi matnlar; Instagram – mashhur bloggerlar, jamoat arboblari va tashkilotlarning izohlari hamda tavsifnomalari. Matnlar Python dasturlash tili va maxsus veb-skrayperlar yordamida to'plandi. Natijada jami 2,31 million so'z va 180 000 ga yaqin gapni o'z ichiga olgan 45 000 ta matn to'plami hosil qilindi. Bu ko'rsatkich kichik-o'rtacha korpus miqyosiga mos keladi va dastlabki statistik tahlil uchun yetarli hisoblanadi [5].

Shaxsiy ma'lumotlarni himoya qilish maqsadida barcha foydalanuvchi nomlari anonimlashtirilib, shaxsni aniqlashga imkon beruvchi ma'lumotlar o'chirib tashlandi. Tadqiqot O'zbekiston Respublikasining «Shaxsga doir ma'lumotlar to'g'risida»gi qonuniga muvofiq olib borildi.

2.2. Matnlarni oldindan qayta ishlash

To'plangan xom matnlar bir qator oldindan qayta ishlash (preprocessing) bosqichlaridan o'tkazildi. Birinchi bosqichda matnlardan HTML kodlari, URL-manzillar, emoji va boshqa nolingvistik elementlar tozalandi. Ikkinchi bosqichda tokenizatsiya – matnni alohida so'z va belgilarga ajratish – amalga oshirildi. O'zbek tili uchun mavjud tokenizatorlar ko'proq adabiy til uchun mo'ljallanganligi sababli, ijtimoiy tarmoq matnlarining o'ziga xos xususiyatlarini hisobga oluvchi maxsus tokenizatsiya qoidalari ishlab chiqildi. Bu qoidalar qisqartmalar, emotikon so'zlar,

Tel:+99833 5668868 | Telegram: @Anvarbek_PhD

harflarning qo'shib yozilishi va defis bilan bog'langan birikmalarga alohida e'tibor beradi.

Uchinchi bosqichda lemmatizatsiya – so'zlarni o'zak shakliga keltirish – amalga oshirildi. O'zbek tilining agglyutinativ xususiyati tufayli so'z shakllarining xilma-xilligi juda katta bo'lib, bu lemmatizatsiyani sezilarli darajada murakkablashtiradi. Masalan, «o'qimoq» fe'lining korpusda 80 dan ortiq turli shakli uchradi. To'rtinchi bosqichda matnlar morfologik belgilar bilan boyitildi – har bir so'zga nutq qismi (Part-of-Speech) yorlig'i biriktirildi. Bu jarayon uchun mavjud o'zbek tili morfologik analizatorlariga asoslangan aralash uslub – qoidaga asoslangan (rule-based) va statistik metodlarning kombinatsiyasi – qo'llanildi.

2.3. Leksik tahlil metodlari

Tadqiqotda quyidagi asosiy metodlardan foydalanildi:

Chastota tahlili: korpusdagi har bir leksik birlikning uchrashuv chastotasi hisoblandi. Korpus chastota lug'ati tuzildi va an'anaviy me'yoriy lug'atlardagi ro'yxat bilan solishtirildi. Bu metod orqali eng faol va eng kam qo'llaniladigan so'zlar, shuningdek, mavjud lug'atlarda aks etmagan yangi birliklar aniqlandi.

Kollokatsia tahlili: C.D. Manning va H. Schütze [6] tavsiya etgan statistik ko'rsatkichlardan – o'zaro ma'lumot (mutual information, MI), t-sinovdan va log-likelihood – foydalanib, so'zlarning bir-biri bilan turg'un bog'lanish naqshlari o'rganildi. Bu so'zlarning semantik sohasini va qo'llanilish kontekstini belgilashda muhim vosita bo'lib xizmat qildi.

Vektorli so'z ifodalari: Korpus asosida Word2Vec [7][3] va GloVe [8][4] modellaridan foydalanib, har bir leksik birlikni ko'p o'lchovli vektor fazosida ifodalovchi so'z vektorlari hosil qilindi. Vektorlar o'rtasidagi kosinusli masofani o'lchash[5] orqali so'zlar o'rtasidagi semantik yaqinlik aniqlandi. Word2Vec modeli CBOW arxitekturasida, 300 o'lchovli vektor fazosida va 5 so'zli oyna (window) parametrida o'qitildi.

Klasterlash tahlili: Hosil qilingan so'z vektorlari k-means va iyerarxik klasterlash algoritmlari yordamida semantik guruhlarga ajratildi. Optimal klasterlar soni siluet koeffitsienti (silhouette coefficient) va «tirsakli» (elbow) metod yordamida belgilandi. Bu usul leksik maydonlarni avtomatik tarzda aniqlashga va ularning chegaralarini belgilashga imkon berdi.

---

[3]Word2Vec modeli 2013-yilda Google tadqiqotchilari T. Mikolov va hamkasblari tomonidan taklif qilingan. Modelning ikki arxitekturasi mavjud: CBOW (Continuous Bag of Words) va Skip-gram. Qarang: [7].

[4]GloVe – Global Vectors for Word Representation – J. Pennington, R. Socher va C.D. Manning (Stanford University) tomonidan 2014-yilda ishlab chiqilgan so'z vektorlari modeli. Word2Vec dan farqli o'laroq, GloVe global statistik ma'lumotlarga asoslanadi. Qarang: [8].

[5]Kosinusli o'xshashlik (cosine similarity) = $\cos(\theta) = (A \cdot B)/(|A| \cdot |B|)$. Qiymat 1 ga yaqinlashganda so'zlar semantik jihatdan yaqin, 0 ga yaqinlashganda mustaqil, −1 ga yaqinlashganda qarama-qarshi ma'noni anglatadi.

Vaqt dinamikasi tahlili: Korpusdagi matnlar yaratilish sanasi bo'yicha choraklik davrlarga ajratilib, so'zlar chastotasining 2021-dan 2024-yilgacha bo'lgan o'zgarish dinamikasi o'rganildi. Neologizmlarning til tizimiga kirish jarayonini kuzatish va viral leksemalarning tarqalish tezligini hisoblash uchun ham ushbu vaqt qatlamlanishi metodidan foydalanildi.

3. NATIJALAR

3.1. Korpusning umumiy statistik tavsifi

Shakllantirilgan korpusning asosiy parametrlari quyidagicha: jami 2,31 million so'z shakli (token), 142 500 noyob leksema (type), type-token nisbati (TTR) – 0,062. Matnlarning o'rtacha uzunligi 51,3 so'z (median: 38 so'z). Mavzular bo'yicha taqsimot: ijtimoiy-siyosiy matnlar – 28%, ko'ngilochar mazmun – 24%, tijorat va reklama – 18%, sport – 14%, ta'lim – 10%, boshqalar – 6%.

Eng yuqori chastotali 50 ta leksemadan iborat ro'yxatni tahlil qilganda, funksional so'zlar (olmosh, bog'lovchi, ko'makchi) bilan bir qatorda «video», «post», «kontent», «link», «story» kabi inglizcha so'zlar o'rin olgan. Bu ijtimoiy tarmoq leksikasining o'ziga xos xususiyatini va ingliz tili ta'sirining kuchliligini yaqqol ko'rsatadi. Qiyoslar uchun: adabiy tilga asoslangan korpuslarda bunday so'zlar odatda birinchi 300 ta chastota ro'yxatiga ham kirmaydi.

3.2. Leksik innovatsiyalar tasnifi

Tadqiqot davomida 3 740 ta yangi leksik birlik aniqlandi. Ular quyidagi toifalarga bo'lindi:

To'g'ridan-to'g'ri o'zlashma so'zlar: ingliz tilidan to'g'ridan-to'g'ri o'zlashtirilgan, o'zbek alifbosiga mos yozuv bilan ishlatiladigan so'zlar. Masalan: «bloger», «instagramchi», «vlogger», «xeshteq», «lajk» (like), «koment» (comment), «ripost» (repost), «rils» (reels), «stori» (story). Bu guruhda jami 1 280 ta leksema (34,2%) aniqlandi va leksik innovatsiyalarning eng katta qismini tashkil etadi.

Gibrid so'z yasalishi: o'zbek va ingliz til unsurlarini biriktirib hosil qilingan yangi birliklar. Masalan: «post» + «-la-» → «postlamoq»; «lajk» + «-la-» → «lajklamoq»; «chat» + «-lash-» → «chatlamoq»; «follow» + «-chi» → «followerchi». Bu guruhda 870 ta leksema (23,3%) aniqlandi. Gibrid so'z yasalishi o'zbek tilining chet el materialini o'z morfologik tizimi orqali o'zlashtira olish kuchini ko'rsatib, o'ziga xos interfeysi sifatida namoyon bo'lmoqda.

Semantik kengayish: mavjud o'zbek so'zlarining ijtimoiy tarmoq kontekstida yangi ma'nolar kasb etishi. «Sahifa» so'zi «profil sahifasi» ma'nosida; «do'st» so'zi «ijtimoiy tarmoqdagi do'st» ma'nosida; «guruh» – «online guruh» ma'nosida; «obuna» – «kanal obunachisi» ma'nosida faol qo'llanilmoqda. Bu guruhda 640 ta leksema (17,1%) qayd etildi. Semantik kengayish hodisasi tilning yangi reallikka moslashishining eng iqtisodiy yo'li hisoblanadi.

Tel:+99833 5668868 | Telegram: @Anvarbek_PhD

Qisqartmalar va akronimlar: DP (direct message/shaxsiy xabar), SM (social media), PR, IT, UGC (user-generated content) kabi inglizcha akronimlar bilan bir qatorda, o'zbek tilidagi qisqartmalar ham faol ishlatilmoqda: «tb» (tegishli bo'lmagan), «ybb» (yaxshi bo'ling barchaga), «hm» (ha, mayli), «sbb» (sababi), «iltk» (iltimos). Jami 420 ta birlik (11,2%) aniqlandi.

Neologizmlar: ijtimoiy tarmoq muhitida paydo bo'lgan yangi o'zbek so'zlari: «tarmoqchi», «kontent yaratuvchi», «raqamli xaritachi», «onlaynchi», «virtual maydon», «izohchi» (kommenter). Bu toifada 530 ta leksema (14,2%) qayd etildi.

3.3. Vektorli model natijalari

Word2Vec modeli orqali hosil qilingan so'z vektorlari semantik tahlil uchun keng imkoniyatlar ochdi. Model sifatini baholash uchun an'anaviy usul – so'z analogiyalari sinovi – qo'llanildi. Masalan: «bloger» – «video» + «rasm» = «instagramchi» kabi analogiyalar modelning to'g'ri natija berishini ko'rsatdi.

Semantik klasterlar tahlili quyidagi asosiy leksik maydonlarni aniqladi: (1) raqamli aloqa vositalari (chat, messenjer, video zvonok, link); (2) kontent yaratish va tarqatish (post, video, reel, story, xeshteq); (3) foydalanuvchi o'zaro ta'siri (lajk, izoh, ulashish, obuna); (4) tijorat va reklama (reklama, promo, brend, PR); (5) siyosiy diskurs (muholifat, referendum, saylov, partiya); (6) ko'ngilochar mazmun (meme, trend, viral, challenge).

«Bloger» so'zining eng yaqin semantik qo'shnilari sifatida «influencer», «vlogger», «tarmoqchi», «kontent yaratuvchi», «lider» so'zlari aniqlandi (kosinusli o'xshashlik: 0,87 dan 0,92 gacha). Bu esa mazkur tushunchaning o'zbek tilida qanday idrok etilishini va uning atrofidagi semantik maydonni aniq ko'rsatadi.

3.4. Vaqt dinamikasi: leksik o'zgarishlarning xronologiyasi

Vaqt bo'yicha tahlil natijasida leksikaning o'zgarishida aniq bosqichlar aniqlandi. 2021-yilda asosan «post», «kontent», «lajk», «like», «share» kabi iboralar dominant bo'lgan. 2022-yildan e'tiboran «rils» (Reels), «stori» (Story), «live» (jonli efir) kabi yangi formatlar bilan bog'liq leksika faol kirdi. Bu Meta korporatsiyasining Instagram platformasini yangilash siyosatiga bevosita bog'liq.

2023-2024-yillarda sun'iy intellekt va chatbot bilan bog'liq – «chatGPT», «neuroset», «AI», «prompt», «generatsiya», «hallutsinatsiya» (model xatosi ma'nosida) – so'zlar keskin o'sish ko'rsatdi. Mazkur so'zlar guruhining 12 oylik o'sish sur'ati 340% ni tashkil etdi. Bu raqam tilning texnologik o'zgarishlarga qanchalik tez va sezgir munosabat bildirishi mumkinligini isbotlaydi.

4. MUHOKAMA

4.1. Ijtimoiy tarmoq leksikasining o'ziga xos xususiyatlari

Tadqiqot natijalari shuni ko'rsatadiki, o'zbek tilidagi ijtimoiy tarmoq matnlari adabiy til va so'zlashuv tilidan farqli ravishda o'zining mustaqil leksik qatlamiga ega. Bu qatlam bir necha muhim jihatdan diqqatga sazovordir.

Tel:+99833 5668868 | Telegram: @Anvarbek_PhD

Birinchidan, ingliz tilidan so'z o'zlashtirishning jadalligi an'anaviy holat bilan solishtirilganda tubdan farq qiladi. Adabiy tilda so'z o'zlashtirilgach, u fonetik va morfologik jihatdan o'zbek tiliga moslashtiriladi, akademik muhokamadan o'tadi va me'yorlashtiriladi. Ijtimoiy tarmoqlarda esa bu jarayon o'ta tez kechadi va bir xil so'zning bir necha imloviy variantlari bir vaqtning o'zida parallel qo'llanilishi kuzatiladi: «lajk/layk/like», «koment/komment/comment», «bloger/blogger». Bu holat til me'yorlashtirish uchun yangi muammolarni keltirib chiqarmoqda.

Ikkinchidan, so'z yasalishining o'zbek tilida an'anaviy ravishda kamdan-kam qo'llaniladigan yangi usullari faollashmoqda. Masalan, «xeshteq» (#) belgisi yordamida hosil qilingan teg-so'zlar an'anaviy grammatika kategoriyalariga to'liq sig'maydi, ammo nutqda mustaqil leksik birlik sifatida ishlatiladi. D. Crystal ta'kidlaganidek, internet tili o'zining maxsus leksik va grammatik qoidalarini ishlab chiqadi [15].

Uchinchidan, kontekst sezgirligi an'anaviy leksikaga qaraganda ancha yuqori. «Like» so'zi o'zbek tilidagi qo'llanilishida nafaqat «yoqtirish» fe'li, balki «baholash» tizimi, «ijobiy munosabat» va hatto «e'tibor belgisi» ma'nolarini ham o'z ichiga oladi. Bu polisemiyaning yangi, raqamli shakli hisoblanadi va leksikografiyada alohida tasnif talab qiladi. Hozirgi o'zbek izohli lug'atlarida bu so'zlarning aksariyati umuman uchramaydi yoki juda umumiy tarzda beriladi [12].

4.2. Korpus lingvistikasi va NLP: metodologik muammolar va yechimlar

Tadqiqot davomida bir qator metodologik muammolarga duch kelindi. Birinchi va eng jiddiy muammo – matnlarning «shovqinliligi» (noisiness). Ijtimoiy tarmoq matnlari imlo xatolariga, nutq xatolariga va nostandart yozuvlarga boy. Masalan, «bilmyman» (bilmayman), «ketdy» (ketdi), «broo» (birodar) kabi yozuvlar tez-tez uchraydi. Bu oldindan qayta ishlash bosqichini sezilarli darajada murakkablashtirdi va maxsus xato-tuzatish moduli yaratishni taqozo etdi.

Ikkinchi muammo – kodlarning almashishi (code-switching) [9].[6] Ko'plab matnlarda o'zbek, rus va ingliz tillari aralashib keladi. «Kecha postle qilgan materialim prosto bomb bo'ldi, barchaga nravitsya!» kabi gaplar bu hodisaning yorqin namunasi. Bunday ko'p tilli muhit uchun til aniqlash va tokenizatsiya algoritmlarini maxsus moslashtirish talab etildi.

Uchinchi muammo – «viral» leksikaning o'tkinchligi. Ijtimoiy tarmoqlarda ko'p so'zlar o'ta qisqa muddatda keng tarqalib, so'ngra deyarli qo'llanishdan chiqib ketadi. Bunday «bir mavsumlilk» leksikani ham qo'lga kiritib, tadqiqot obyekti sifatida belgilash tilshunoslik uchun muhim metodologik masaladir.

[6]Code-switching (kod almashinuvi) – bir nutqiy aktda ikki yoki undan ortiq tildan foydalanish hodisasi. Ikki tilli va ko'p tilli jamiyatlarda keng tarqalgan bo'lib, so'nggi yillarda raqamli aloqa muhitida yanada intensivlashgan. Qarang: [9].

BERT asosidagi kontekstual til modellari [3][7] ushbu muammolarning bir qismini hal qilishda samarali bo'ldi. Kontekstual vektorlar yordamida «post» so'zining «postblog» (blog yozuvi) va «postsovet» (sovetdan keyingi) kabi turli ma'nolarini kontekstga qarab aniq ajratish imkoni yaratildi. Biroq BERT modelini o'zbek tili uchun muvaffaqiyatli o'qitish hali yetarli hajmdagi annotatsiyalangan korpusni talab qilmoqda.

4.3. O'zbek tilshunosligi va leksikografiyasi uchun amaliy ahamiyati

Tadqiqot natijalari bir necha muhim amaliy yo'nalish uchun poydevor vazifasini o'taydi. Birinchidan, zamonaviy o'zbek tili lug'atlarini yangilash va boyitish uchun hujjatlashtirilgan manba bo'lib xizmat qiladi. Yo'ldoshev M. ta'kidlaganidek [13], o'zbek tilining leksik qatlami doimo harakatda bo'lib, uni dinamik kuzatib borish tilshunos oldidagi asosiy vazifalardan biridir. Shakllantirilgan korpus va aniqlangan 3 740 ta yangi birlik ushbu vazifani bajarishda bevosita qo'llanilishi mumkin.

Ikkinchidan, o'zbek tili uchun yangi NLP vositalarini yaratishda – mashinali tarjima tizimlarini, his-tuyg'u tahlili (sentiment analysis) va ma'lumotlarni avtomatik ajratib olish (information extraction) dasturlarini – bu korpus va uning asosida qurilgan modellar muhim ahamiyat kasb etadi. Tursunov, Muxtorov va Rahmatullayev ijodiy merosida belgilangan o'zbek tili leksikasining semantik xaritasi [12] bugungi raqamli sharoitda yangi mazmun bilan to'ldirilishi zarur.

Uchinchidan, ijtimoiy tarmoq matnlarini qayta ishlash tizimlari (masalan, kontent moderatsiya algoritmlari, spam-filtrlari) uchun o'zbek tilidagi yangi leksikani bilish majburiy zaruriyat hisoblanadi. Shu sababdan ham, Matlatipov va hamkasblari [14] belgilagan o'zbek tili uchun NLP vositalarini yaratish yo'lidagi qadamlar bugungi kunda dolzarbligini yanada oshirdi.

Bundan tashqari, tadqiqot ijtimoiy lingvistika va sotsiologiyaning kesishmasidagi muhim muammoni ham ko'taradi. P. Bourdieuning lingvistik kapital nazariyasi nuqtai nazaridan [11][8], raqamli muhitda inglizcha texnik atamalarni to'g'ri bilish va qo'llash o'zbek foydalanuvchilar uchun ijtimoiy-madaniy kapitalning haqiqiy ko'rinishiga aylandi. «Influencer», «kontent», «brend» kabi so'zlarni o'rinli ishlatish kishining raqamli savodxonligi va ijtimoiy maqomining belgisiga aylangan.

4.4. Cheklovlar va kelajakdagi tadqiqot yo'nalishlari

Hozirgi tadqiqotning quyidagi cheklovlari mavjud: birinchidan, korpusning hajmi (2,31 million so'z) tilshunoslik standartlari uchun o'rtacha miqyosga to'g'ri

[7]BERT – Bidirectional Encoder Representations from Transformers – 2018-yilda Google AI tadqiqot guruhi tomonidan ishlab chiqilgan kontekstual til modeli. 2019-yilda NAACL konferentsiyasida e'lon qilingan. Qarang: [3].

[8]Lingvistik kapital (Bourdieu, 1991) – muayyan til shakllarini bilish va ulardan foydalanish orqali olinadigan ijtimoiy afzalliklar majmui. Raqamli muhitda inglizcha texnik atamalarni bilish shunday kapital rolini o'ynaydi.

keladi va umumlashtirish uchun ba'zi hollarda yetarli bo'lmasligi mumkin; ikkinchidan, matnlar asosan yozma nutqni ifodalab, og'zaki nutq shakllarini qamrab olmaydi; uchinchidan, Telegram kanallaridan olingan matnlar platforma API cheklovlari sababli to'liq hajmda olinmadi.

Kelajakda tadqiqotni quyidagi yo'nalishlarda kengaytirish zarur: (1) korpus hajmini 10 million so'z va undan yuqoriga yetkazish va yangi platformalar (TikTok, YouTube izohlari) matnlarini qo'shish; (2) turli mintaqaviy ijtimoiy tarmoq matnlarini kiritib, dialektal va mintaqaviy xususiyatlarni qiyosiy o'rganish; (3) transformer asosidagi modellarni o'zbek tiliga moslashtirib, yanada chuqur semantik tahlil olib borish; (4) diaXronik (tarixiy dinamik) leksik o'zgarishlarni uzoq muddatli kuzatish tizimini yaratish; (5) o'zbek tili uchun birinchi ijtimoiy tarmoq lingvistik yo'l xaritasini ishlab chiqish.

5. XULOSA

Ushbu tadqiqot o'zbek tilshunosligi tarixida ijtimoiy tarmoqlar asosida maxsus korpus shakllantirilib, komputer-lingvistik metodlar yordamida leksik tizim o'rganilgan dastlabki keng ko'lamli urinishlardan biri hisoblanadi. Tadqiqot jarayonida quyidagi asosiy xulosalar chiqarildi.

Birinchi xulosa: ijtimoiy tarmoqlardagi o'zbek tilidagi matnlar adabiy va so'zlashuv tillaridan tizimli ravishda farqlanadigan alohida leksik qatlamni tashkil etadi. Aniqlangan 3 740 ta yangi leksik birlik – bu nafaqat raqamli ko'rsatkich, balki o'zbek tilining zamonaviy hayotga moslashuvchanligini ham tasdiqlaydi.

Ikkinchi xulosa: yangi leksemalarning asosiy manbai ingliz tili bo'lib (34,2%), so'z o'zlashtirishning tezligi va intensivligi bundan avvalgi davrlarga nisbatan keskin oshgan. Biroq gibrid so'z yasalishi (23,3%) ham kuchli tendensiya sifatida namoyon bo'lmoqda – bu o'zbek tilining o'z ichki morfologik mexanizmi orqali yangi materiyani ijodiy o'zlashtira olish qobiliyatini ko'rsatadi.

Uchinchi xulosa: Word2Vec va GloVe asosida qurilgan leksik vektorlar modeli ijtimoiy tarmoq leksikasidagi semantik munosabatlarni aniqlashda ishonchli natija berdi. Mazkur modellar o'zbek tili uchun keyingi NLP vositalarini yaratishda bevosita foydalaniladigan til resursi bo'lib xizmat qilishi mumkin.

To'rtinchi xulosa: leksikaning vaqt dinamikasi texnologik tendensiyalar bilan bevosita va kuchli bog'liq ekanligi isbotlandi. Sun'iy intellekt bilan bog'liq leksikaning 12 oy ichida 340% o'sishi til va texnologiya o'rtasidagi simbiozning yaqqol namunasi hisoblanadi.

Beshinchi xulosa: metodologik jihatdan, ijtimoiy tarmoq matnlarini qayta ishlashda matnlarning «shovqinliligi» va kod almashinuvi hodisasi alohida vositalarni talab qiladi. Bu muammolarni hal qilishga yo'naltirilgan yondashuvlar o'zbek tili NLP sohasining ustuvor masalalariga aylangan.

Umuman olganda, ushbu tadqiqot o'zbek tili leksikografiyasini rivojlantirish, tabiiy tilni qayta ishlash vositalarini yaratish va raqamli lingvistika sohasidagi keyingi ilmiy izlanishlar uchun muhim nazariy-metodologik asos yaratadi. Ijtimoiy

tarmoqlar o'zbek tilining eng jonli va faol rivojlanayotgan qatlami bo'lib, ularni o'rganmaslik tilshunoslikda katta bo'shliqni anglatadi. Bu yo'nalishdagi izlanishlarni tizimli tarzda davom ettirish fanning ham, jamiyatning ham talabi bo'lib qolmoqda.

FOYDALANILGAN ADABIYOTLAR

1. We Are Social & Hootsuite. Digital 2023: Uzbekistan. – 2023. – URL: https://datareportal.com/reports/digital-2023-uzbekistan (murojaat sanasi: 10.03.2024).
2. McEnery, T., Hardie, A. Corpus Linguistics: Method, Theory and Practice. – Cambridge University Press, 2012. – 294 p.
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. – 2019. – P. 4171–4186.
4. Sinclair, J. Corpus, Concordance, Collocation. – Oxford University Press, 1991. – 179 p.
5. Bird, S., Klein, E., Loper, E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. – O'Reilly Media, 2009. – 504 p.
6. Manning, C.D., Schütze, H. Foundations of Statistical Natural Language Processing. – MIT Press, 1999. – 680 p.
7. Mikolov, T., Chen, K., Corrado, G., Dean, J. Efficient Estimation of Word Representations in Vector Space // arXiv preprint arXiv:1301.3781. – 2013.
8. Pennington, J., Socher, R., Manning, C.D. GloVe: Global Vectors for Word Representation // Proceedings of EMNLP. – 2014. – P. 1532–1543.
9. Winford, D. An Introduction to Contact Linguistics. – Blackwell, 2003. – 384 p.
10. Vaswani, A., Shazeer, N., Parmar, N. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998–6008.
11. Bourdieu, P. Language and Symbolic Power. – Harvard University Press, 1991. – 302 p.
12. Tursunov, U., Muxtorov, J., Rahmatullayev, Sh. Hozirgi o'zbek adabiy tili. – Toshkent: O'zbekiston, 1992. – 400 b.
13. Yo'ldoshev, M. O'zbek badiiy matnida ko'chma ma'no muammolari. – Toshkent: Fan, 2008. – 286 b.
14. Matlatipov, G., Tanaka-Ishii, K., Umarov, B., Matlatipova, M. Context-Dependent Machine Translation of the Uzbek Language // Proceedings of SLTU-2012: Workshop on Spoken Language Technologies for Under-Resourced Languages. – 2012. – P. 183–186.
15. Crystal, D. Language and the Internet. – Cambridge University Press, 2001. – 272 p.