TOVUSH O‘ZGARISH HOLATLARIDA O‘ZBEK TILI UCHUN LEMMALAR VA STEMLAR MA’LUMOTLARI BAZASINI YARATISH

Maqsud Siddiqovich SHARIPOV
Texnika fanlari nomzodi
Urganch davlat universiteti
Urganch, O‘zbekiston
maqsbek72@gmail.com

Annotatsiya

Ushbu maqolada o‘zbek tilidagi fleksiyon konteksda lemmatizatsiya va stemming jarayonlarini kuchaytirishga qaratilgan maxsus ma’lumotlar bazasini ishlab chiqishning afzalligi tahlil etiladi.

Tayanch so‘zlar: Fleksiya, tovush o‘zgarishi, lemma, lemmalash, o‘zak, o‘zakni topish.

СОЗДАНИЕ БАЗЫ ДАННЫХ ЛЕММ И СТЕМОВ ДЛЯ УЗБЕКСКОГО ЯЗЫКА В СЛУЧАЯХ ЗВУКОИЗМЕНЕНИЯ

Максуд Сиддикович ШАРИПОВ
Кандидат технических наук
Ургенчский государственный университет
Ургенч, Узбекистан

Аннотация

В статье анализируются преимущества разработки специализированной базы данных, направленной на совершенствование процессов лемматизации и стемминга во флексионном контексте в узбекском языке.

Ключевые слова: Флексия, изменение звука, лемма, лемматизация, основа, стемминг.

O‘zbek tili, turkiy tillar oilasiga mansub bo‘lib, asosga qo‘shimchalar qo‘shish orqali so‘z shakllarini keng o‘zgartirish imkonini beruvchi boy agglutinativ tuzilishga ega. Ushbu morfologik murakkablik tilning o‘ziga xos xususiyati bo‘lishi bilan birga, tabiiy tilni qayta ishlash (inglizcha: Natural Language Processing - NLP) va kompyuter lingvistikasi sohalarida, ayniqsa, lemmatizatsiya va stemming jarayonlarida sezilarli qiyinchiliklar tug‘diradi. Bu jarayonlar so‘z shakllarini asosiy (lemma) yoki ildiz (stem) shakliga keltirishni o‘z ichiga oladi va axborot izlash, mashinaviy tarjima hamda matnni normallashtirish kabi ko‘plab tabiiy tilni qayta ishlash dasturlarining asosi hisoblanadi. Biroq, an’anaviy lemmatizatsiya va stemming yondashuvlari o‘zbek tilining murakkab fonetik o‘zgarishlari va morfologik xususiyatlari tufayli samarasiz bo‘lishi

mumkin, chunki bu omillar so‘z shakllarining dinamik o‘zgarishiga olib kelib, ularning asl tuzilishini noaniq qiladi. Ushbu muammolarni hal qilish uchun o‘zbek tilining o‘ziga xos xususiyatlarini inobatga olgan, lingvistik jihatdan asoslangan kuchli resurs zarur.

Ushbu tadqiqot mazkur muammolarni hal etish maqsadida o‘zbek tilidagi fonetik o‘zgarishlar va morfologik xilma-xillikni aniq aks ettiruvchi maxsus ma’lumotlar bazasini ishlab chiqishga qaratilgan. Bunday resurs tabiiy tilni qayta ishlash tizimlarining aniqligi va samaradorligini oshirish uchun muhim ahamiyatga ega. O‘zbek tilida keng tarqalgan fonetik o‘zgarishlar asosiy shakllarni aniqlashni murakkablashtiradi, chunki bunday o‘zgarishlar so‘z shakllari bilan ularning ildizlari yoki lemmalari o‘rtasidagi bog‘liqlikni noaniq qiladi. Tadqiqotimiz ushbu fonetik o‘zgarishlarning so‘z tuzilishiga ta’sirini inobatga olgan holda, lingvistik vositalarning mustahkamligini oshirishga xizmat qiladi va ularning tilning real hayotdagi murakkabliklarini aniqroq aks ettirishini ta’minlaydi.

O‘zbek tilida so‘z yasalishida fleksiyalash muhim rol o‘ynaydi, shu sababli ilg‘or va fonetik o‘zgarishlarni hisobga oluvchi lemmatizatsiya hamda stemming usullariga bo‘lgan ehtiyoj juda dolzarbdir. Ushbu fleksiyalash jarayonlarini inobatga oladigan ishonchli usullarsiz hisoblash tizimlari matnni aniq qayta ishlashda qiyinchiliklarga duch kelishi mumkin. Lemmatizatsiya va stemming – matnni normalizatsiya qilish uchun zarur bo‘lgan asosiy oldindan ishlov berish texnikalaridir. Lemmatizatsiya fleksiyalangan so‘z shakllarini ularning lug‘aviy asosiy ko‘rinishiga (lemmalariga) moslashtirib, turli variantlar orasida semantik yaxlitlikni ta’minlaydi. Stemming esa so‘zdan qo‘shimchalarni olib tashlash orqali uning ildiz shaklini ajratib, matn indeksatsiyasi kabi jarayonlarda hisoblash samaradorligini oshirishga yordam beradi. Fleksiyalash jarayoni bilan ushbu usullarning o‘zaro bog‘liqligi mashinalarning inson tilini aniq tushunishi va yaratishini ta’minlashda muhim ahamiyat kasb etadi, ayniqsa, o‘zbek tili kabi kam resursli tillarda.

Mavjud ilmiy ishlarga chuqur tahliliy yondashuv lemmatizatsiya va stemming jarayonlari turli lingvistik modellarda qanday amalga oshirilganini

hamda ushbu usullarning murakkab morfologik va fonologik tizimlarga ega tillarga tatbiq etilgandagi cheklovlarini aniqlash muhimligini ko‘rsatadi. Turkiy tillar va boshqa morfologik jihatdan boy tillarga oid tadqiqotlar tizimli ravishda tahlil qilinib, ularning metodologiyalari, vositalari va resurslari o‘rganilgan. Natijada, tabiiy tilni qayta ishlash (NLP) sohasida erishilgan yutuqlar va hanuzgacha mavjud bo‘lgan qiyinchiliklar aniqlanib, ushbu yo‘nalishda samarali yondashuvlar ishlab chiqish zarurligi ta’kidlangan.

Ushbu maqolada [1] o‘zbek tili uchun atoqli otlarni aniqlash (Named Entity Relationship - NER) ma’lumotlar to‘plami taqdim etilgan bo‘lib, u 1 160 ta jumla va 19 000 ta belgilangan so‘z shakllarini o‘z ichiga oladi. Mualliflar, shuningdek, o‘zbek tili uchun ikkita atoqli otlarni aniqlash algoritmini ishlab chiqib, kam resursli tillarda tabiiy tilni qayta ishlash va mashinaviy o‘rganish tadqiqotlariga muhim hissa qo‘shgan. Ushbu ish o‘zbek tilida atoqli otlarni aniqlash tizimlarini shakllantirish uchun mustahkam asos yaratib, chat-botlar va muhim ma’lumotlarni ajratib olish vositalari kabi amaliy dasturlarni qo‘llab-quvvatlash imkonini beradi. Nomuhim so‘zlarni [2] filtratsiya qilish samarali matn qayta ishlash uchun muhim bo‘lib, semantik mazmunni yo‘qotmasdan so‘rovlar murakkabligini kamaytirishga yordam beradi. Biroq, mavjud modellar agglutinativ tillar, jumladan, o‘zbek tili uchun yetarli darajada samarali emas, chunki bu tillarda stop so‘zlar qo‘shimchalar bilan yashiringan bo‘lishi mumkin. Ushbu tadqiqot agglutinativ tillarda nomuhim so‘zlarni aniqlash uchun kollokatsiya usulini taklif qiladi. Unigram, bigram va kollokatsiya tahlili qo‘llanib, o‘zbek tilidagi "Maktab korpusi" (School corpus) asosida nomuhim so‘zlarni filtratsiya qilish samaradorligi oshiriladi. Ushbu maqola [3] o‘zbek va qozoq tillari uchun mashinaviy tarjima maqsadida tuzilgan parallel korpusni taqdim etadi. Korpus uch bosqichda shakllantirilgan: avval mavjud ma’lumotlardan foydalanilgan, keyin ochiq manbalardan olingan moslashtirilgan matnlar qo‘shilgan va yakuniy bosqichda ikki tilli mutaxassislar tomonidan qo‘lda tarjima qilingan. Turli mavzularni qamrab olgan ushbu ma’lumotlar to‘plami neyron tarmoqlarga asoslangan mashinaviy tarjima vazifalarini qo‘llab-quvvatlaydi hamda Hugging Face platformasida joylashtirilgan. Bu resurs o‘zbek-qozoq

tarjimasi va boshqa kam resursli turkiy tillar juftligi uchun muhim manba bo‘lib xizmat qiladi. Ushbu tadqiqot [4] Qirg‘iziston, Mo‘g‘uliston va O‘zbekistonda fermerlarning qurg‘oqchilik haqidagi subyektiv tushunchalari bilan ob'yektiv meteorologik ma’lumotlar o‘rtasidagi tafovutni tahlil qiladi. 2,830 ta fermer xo‘jaligi darajasidagi kuzatuvlar va Standartlashtirilgan yog‘ingarchilik-evapotranspiratsiya indeksi (SPEI) ma’lumotlaridan foydalangan holda o‘tkazilgan tadqiqot natijalari fermerlar tomonidan qurg‘oqchilikni noto‘g‘ri baholash holatlarini aniqladi: Qirg‘izistonda 67%, Mo‘g‘ulistonda 32% va O‘zbekistonda 46% fermerlar qurg‘oqchilikni noaniq baholagan. Tadqiqot shuningdek, smartfon asosidagi ob-havo ma’lumotlari qurg‘oqchilik haqidagi tushunchalarning aniqligini oshirishda muhim rol o‘ynashini ta’kidlaydi va past raqamli savodxonlikka ega fermerlar uchun internetga kirish va ob-havo xizmatlarini yaxshilash zarurligini ilgari suradi. Ushbu maqola [5] o‘zbek tili uchun lemmatizatsiya algoritmini yaratish jarayonini muhokama qiladi. Ushbu maqolada [6] o‘zbek tili matnlarida fe’llarni aniqlash uchun qo‘shimcha va affikslarga asoslangan qoida asosidagi yondashuv taklif etiladi. Ushbu maqolada [7] mualliflar o‘zbek tili uchun qoidaviy stemming algoritmini taqdim etadilar. Ushbu tadqiqot maqolasi [8] kam resursli o‘zbek tili uchun so‘z turkumlari (Part-of-speech - POS) bilan belgilangan dataset va teglash vositasini taqdim etadi.

Ushbu tadqiqot [9] Turk ishora tili uchun yirik ko‘lamli, multi-modal ma’lumotlar to‘plami – AUTSL ni taqdim etadi. Mazkur dataset 43 nafar ishtirokchi tomonidan bajarilgan 226 ishora harakatlarini o‘z ichiga oladi va 38,336 ta video namunasi Microsoft Kinect v2 yordamida (RGB, chuqurlik va skelet ma’lumotlari) yozib olingan. AUTSL foydalanuvchidan mustaqil baholash imkonini beruvchi shaklda tuzilgan bo‘lib, real sharoitlarni aniqroq aks ettirishga mo‘ljallangan. Tadqiqotda Konvolyutsion Neyron Tarmoqlar xususiyatlar chiqarish uchun va Long Short-Term Memory (LSTM) tarmoqlari vaqt bo‘yicha modellashtirish uchun ishlatilgan. Sinov natijalariga ko‘ra, Montalbano datasetida 96.11% aniqlikka, AUTSL datasetida tasodifiy bo‘linmalar bilan 95.95% natijaga erishilgan, foydalanuvchidan mustaqil test esa 62.02% natija ko‘rsatgan, bu esa

haqiqiy dunyodagi ishora tili tanish tizimlari uchun dolzarb muammolar mavjudligini tasdiqlaydi. Shuningdek, morfologik jihatdan boy tillar, masalan, turk tili uchun an’anaviy NLP usullari tokenizatsiya, lemmatizatsiya va xususiyatlar muhandisligi kabi keng qamrovli oldindan qayta ishlash bosqichlarini talab qiladi. Bunday tillarda ma’lumotlarning kamligi va yuqori o‘lchamli bo‘lishi kabi muammolarni bartaraf etish uchun maxsus yondashuvlar zarur. Ushbu tadqiqot [10] turk tilida BERT modelining samaradorligini o‘rganib, oldindan qayta ishlash yukini minimallashtirishga qaratilgan. Baholash beshta turkcha NLP vazifalarini – kayfiyat tahlili, kiberbulling aniqlash, matn tasnifi, hissiyotlarni aniqlash va spamni aniqlashni o‘z ichiga oladi hamda BERT modelining ishlash natijalari an’anaviy mashinaviy o‘rganish modellarining natijalari bilan taqqoslanadi. Natijalar shuni ko‘rsatadiki, BERT an’anaviy modellarni ortda qoldirib, murakkab oldindan qayta ishlash bosqichlarisiz turkcha matnlarni samarali qayta ishlay oladi. Ushbu tadqiqot [11] Complement Naive Bayes (CNB), Random Forest (RF), Decision Tree (C4.5/J48) va Support Vector Machines (SVMs) kabi mashinaviy o‘rganish modellarini baholaydi. Natijalar shuni ko‘rsatadiki, SVM modellar eng yuqori samaradorlikka erishgan bo‘lib, tasdiqlangan dataset asosida o‘qitilgan modellar xom ma’lumotlar asosida o‘qitilgan modellar bilan solishtirganda yuqori natijalarga erishgan. Ushbu natijalar hissiyotlarni ajratib olish vazifalarida dataset sifatining muhimligini ta’kidlaydi. Matndan hissiyotlarni ajratib olish tabiiy tilni qayta ishlashning asosiy vazifalaridan biri bo‘lib, nazorat ostidagi o‘rganish uchun yirik, annotatsiyalangan datasetlarni talab qiladi. Biroq, turk tili uchun hissiyot yorliqlangan to‘liq datasetlarning yetishmovchiligi mavjud. Ushbu tadqiqot turk tilida hissiyotlarni ajratib olish uchun yangi datasetni taqdim etadi va matnlarni olti xil hissiyot – baxt, qo‘rquv, g‘azab, qayg‘u, jirkanish va hayrat – bo‘yicha tasniflaydi. Ma’lumotlar 4,709 nafar ishtirokchi tomonidan o‘tkazilgan so‘rov orqali to‘plangan bo‘lib, natijada 27,350 ta yozuv shakllantirilgan va annotatorlar tomonidan tasdiqlangan. Ushbu tadqiqot [12] kayfiyat tasnifi uchun to‘rtta mashinaviy o‘rganish modelini ishlab chiqib, ularning balanslangan va balanslanmagan ssenariylardagi samaradorligini baholaydi. Eng yaxshi model

polarizatsiya tasnifi uchun 0,81 F1-koʻrsatkichiga, ball tasnifi uchun esa 0,39 F1-koʻrsatkichiga erishgan. Dataset va nozik sozlangan modellar Creative Commons Attribution 4.0 International License asosida GitHub platformasida ochiq holda taqdim etilgan. Qozoq tili uchun kayfiyat tahlili katta, annotatsiyalangan datasetlarning yetishmovchiligi sababli qiyinchiliklarga duch kelgan. Ushbu tadqiqot KazSAnDRA nomli datasetni taqdim etadi, bu Qozoq tili uchun birinchi va eng yirik ochiq kayfiyat tahlili datasetidir. Unda 180,064 ta sharh mavjud boʻlib, ular 1 dan 5 gacha boʻlgan reytinglar bilan baholangan. Ushbu dataset polarizatsiya va ball tasnifi vazifalarini qoʻllab-quvvatlaydi. Ushbu tadqiqot [13] IOB2 sxemasiga muvofiq annotatsiyalangan qozoqcha atoqli otlarni tanish (NER) datasetini taqdim etadi. Ushbu dataset 112,702 ta jumla va 25 ta entity sinfi boʻyicha 136,333 ta annotatsiyani oʻz ichiga oladi. Annotatsiya jarayoni ikki nafar ona tilida soʻzlashuvchi qozoq tili mutaxassisi tomonidan, izchillikni ta’minlash maqsadida nazorat ostida amalga oshirilgan. NER uchun ilgʻor mashinaviy oʻrganish modellarini ishlab chiqish natijasida eng yaxshi model test toʻplamida 97.22% F1-koʻrsatkichiga erishgan. Dataset, annotatsiya yoʻriqnomalari va trening kodlari Creative Commons Attribution 4.0 License asosida ochiq holda taqdim etilgan boʻlib, kelgusidagi tadqiqotlar uchun foydalanish mumkin. Ushbu tadqiqot [14] kalit soʻzlarni ajratib olish uchun Random Forest va XgBoost (Extreme Gradient Boosting) algoritmlaridan foydalanishni oʻrganadi hamda ularning samaradorligini ikkita datasetda baholaydi: keng qoʻllaniladigan 500N-KPCrowd dataseti (ingliz tilidagi yangiliklar) va qozoq tilidagi dataset. Tadqiqot natijalariga koʻra, qozoq tilidagi dataset uchun eng yuqori natijaga erishgan model 0.97 F1-koʻrsatkichini qayd etgan, bu ilmiy adabiyotlarda e’lon qilingan eng yuqori natijadir. Taqqoslash uchun, 500N-KPCrowd datasetida eng yaxshi F1-koʻrsatkich 0.70 ga teng boʻlgan.

Ushbu tadqiqot [15] Indoneziyaning mintaqaviy tillari uchun soʻnggi yillarda stemming va lemmatizatsiya boʻyicha erishilgan yutuqlarni tizimli ravishda koʻrib chiqadi. 2014 yildan 2023 yilgacha chop etilgan 35 ta tanlangan tadqiqotga asoslangan holda, izlanish natijalari raqamli lugʻatlar va morfologik

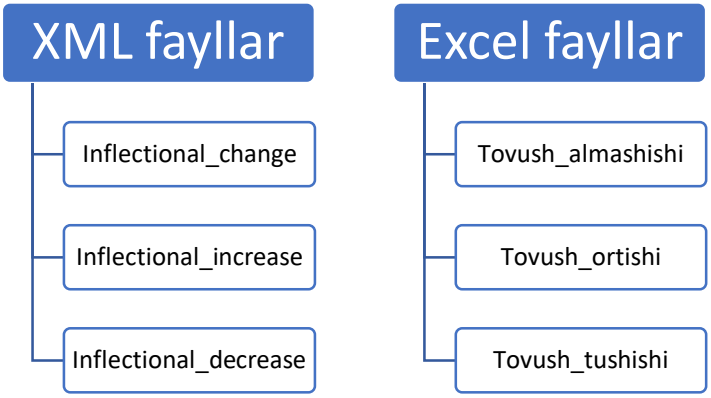
ma’lumotlarning stemming samaradorligini oshirishdagi muhim rolini ta’kidlaydi. Shuningdek, tadqiqotda katta korpuslarga ega tillargagina mos keluvchi korpusga asoslangan usullarning cheklovlari qayd etilgan. Shunga qaramay, o‘tgan o‘n yil ichida stemming va lemmatizatsiya texnikalarida sezilarli taraqqiyot kuzatilgan. Ushbu maqola [16] jumlalarni qayta qidirish masalasini o‘rganadi, bu hujjatni qayta qidirishga o‘xshash bo‘lib, faqat kichikroq tahlil birligiga asoslanadi. Tadqiqotda TF-IDF (so‘z chastotasi – teskari hujjat chastotasi) usuli qo‘llanilib, oldindan ishlov berish texnikalari, jumladan, stop-so‘zlarni olib tashlash, stemming va lemmatizatsiya tatbiq etilgan. Natijalar shuni ko‘rsatadiki, oldindan ishlov berish jarayonlari jumlalarni qayta qidirish samaradorligini oshiradi, bu esa hujjatlarni qayta qidirishga o‘xshash natijalarga olib keladi. Uzoqroq so‘rovlar uchun lemmatizatsiya yaxshi natija bergan bo‘lsa, stemmingning samaradorligi nisbatan pastroq bo‘lgan. Eksperimentlar Text Retrieval Conference (TREC) ning yangilik izlash yo‘nalishi ma’lumotlari asosida o‘tkazilgan.

O‘zbek tilidagi so‘zlarning fleksiyalanish bazasini yaratish maqsadida turli manbalardan ma’lumotlar to‘plash ishlari amalga oshirildi. Ushbu ma’lumotlar bazasi muallif rahbarligida lingvistlar tomonidan tuzilgan bo‘lib, unda quyidagi manbalardan foydalanilgan:

- *O‘zbek tilining izohli lug‘ati*
- *Maktab ona-tili darsliklari*
- *O‘zbek tili grammatika va qoidalar qo‘llanmasi*
- *Tilshunoslik sohasidagi maqolalar*

Ma'lumotlar GitHub platformasida saqlangan bo‘lib, unda batafsil ma’lumot va metadata mavjud. Bu ma’lumotlar bazasiga qulay murojaat va tahlil qilish imkoniyatini ta’minlaydi hamda quyidagi havola orqali ochiq tarzda foydalanish mumkin: https://github.com/MaksudSharipov/UzbekInflectionalWords.

Ma'lumotlar ikki turdagi fayllarda saqlangan: Excel va XML (eXtensible Markup Language). Bu formatlar ma’lumotlarni tizimli ravishda saqlash, qayta ishlash va tahlil qilish imkonini beradi.



1-rasm. Ma’lumotlar bazasi fayl turlari va nomlanishi

XML formatidan ma’lumotlarni saqlashda foydalanish juda muhim bo‘lib, u ma’lumotlarning tuzilmaviy tashkil etilishini, kengaytiriluvchanligini va o‘zaro moslashuvchanligini ta’minlaydi. XML hujjatlarni inson va mashina uchun tushunarli standartlashtirilgan formatda kodlash imkonini beradi, bu esa murakkab va ierarxik ma’lumotlarni saqlash uchun qulay tanlov hisoblanadi. Bundan tashqari, XML turli tizimlar o‘rtasida ma’lumot almashinuvini osonlashtiradi va turli platformalar o‘rtasida moslik va integratsiyani talab qiladigan dasturlar uchun muhim format hisoblanadi. Ma’lumotlar bazasi **1-jadvalda** keltirilgan teglarni o‘z ichiga oladi.

Teg nomi	**Ta’rifi**
<data></data>	*barcha ma'lumotlarni saqlovchi asosiy teg*
<row></row>	*har bir ma'lumot yozuvini (so‘z, stem, lemma, fleksiyalash) saqlovchi teg*
<inflection_type> </inflection_type>	*fleksiyalash turini (tovush almashishi, ortish, tushishi) saqlovchi teg*
<index></index>	*har bir yozuvning ketma-ket raqamini saqlovchi teg*
<word></word>	*fleksiyalangan so‘zni saqlovchi teg*
<stem_inflectional> </stem_inflectional>	*fleksiyalangan asosiy shaklni saqlovchi teg*
<stem_correct> </stem_correct>	*so‘zning asosi (fleksiyalanmagan) uchun teg*
<lemma> </lemma>	*so‘zning lemmasini saqlovchi teg*
<formed> </formed>	*fleksiyalash shakllanishini saqlovchi teg*
<rule> </rule>	*fleksiyalash qoidalarini saqlovchi teg*

1-jadval. XML faylidagi teglar ro‘yxati

Ushbu ma’lumotlar bazasi o‘zbek tilidagi so‘zlarning asosiy shakllarini (stem) va fleksiyalanish jarayonlari, xususan, qo‘shimchalar qo‘shilishi natijasida

hosil bo‘ladigan yangi shakllarni kataloglashtirish bo‘yicha ilk keng qamrovli tashabbuslardan biri hisoblanadi. Mazkur resurs so‘z yasalishi va o‘zbek tilining morfologik tuzilishini o‘rganish uchun noyob va qimmatli manba sifatida xizmat qiladi.

Ushbu ma’lumotlar bazasining yaratilishi o‘zbek tilshunosligi, xususan, morfologiya, leksikografiya va tabiiy tilni qayta ishlash (NLP) sohalaridagi tadqiqotlarni rivojlantirishda muhim qadam hisoblanadi. Mazkur resurs kelajakdagi hisoblash modellarini, mashinaviy o‘rganish ilovalarini hamda avtomatlashtirilgan matn qayta ishlash vositalarini yaratish uchun mustahkam asos bo‘lib xizmat qilishi mumkin.

Natijalar. Ushbu tadqiqot davomida fleksiyalangan so‘z shakllarini o‘z ichiga olgan uchta alohida fayl yaratildi, har biri ma’lum bir ma’lumotlar to‘plamini saqlaydi.

- **Tovush almashishi: 59 ta**
- **Tovush ortishi: 56 ta**
- **Tovush tushishi: 39 ta**

Ushbu ma’lumotlar to‘plamlari XML formatida tuzilgan va saqlangan bo‘lib, bu ma’lumotlarning standartlashtirilgan va oson foydalaniladigan ko‘rinishda ifodalanishini ta’minlaydi. Ayniqsa, XML formati ma’lumotlarning aniq va ierarxik tashkil etilishini saqlashga yordam bergan, bu esa kelajakdagi qayta ishlash va tahlil jarayonlarini yanada qulay qiladi.

Ushbu tadqiqot o‘zbek tilining lemmatizatsiya va stemming jarayonlarida yuzaga keladigan asosiy muammolarni, jumladan, fonetik o‘zgarishlar va fleksiyalangan morfologiyaning murakkabliklarini hal etishga qaratilgan maxsus lemma va asosiy so‘z shakllari (stem) ma’lumotlar bazasini taqdim etadi. Tadqiqot doirasida 154 fleksiyalash birligi - 59 tovush almashishi, 56 tovush ortishi va 39 tovush tushishi - Excel hamda XML formatlarida kataloglashtirildi. Ushbu resurs tabiiy tilni qayta ishlash (NLP) tizimlarining aniqligini oshirishga xizmat qilib, o‘zbek tilining dinamik tuzilishini aniq aks ettirishga yordam beradi. Tadqiqot natijalari an’anaviy usullarning yetarli emasligini ko‘rsatib, agglutinativ tillar,

jumladan, o‘zbek tili uchun fonetik o‘zgarishlarni hisobga oluvchi resurslarning zarurligini tasdiqlaydi.

Turli lingvistik manbalarga asoslangan va <word>, <lemma> kabi batafsil XML teglar bilan tuzilgan ushbu ma’lumotlar bazasi tadqiqotchilar va tizimlar uchun mustahkam va o‘zaro moslashuvchan vosita sifatida xizmat qiladi. U ilgari olib borilgan turkiy tillarga oid tadqiqotlardagi mavjud bo‘shliqni to‘ldirib, matnni normalizatsiya qilish jarayonlarini qo‘llab-quvvatlaydi hamda o‘zbek tilining lingvistik merosini saqlashga hissa qo‘shadi. Ushbu tadqiqot kam resursli tillar uchun kengaytirilishi mumkin bo‘lgan yechim taklif etib, hisoblash tilshunosligi sohasini yangi bosqichga olib chiqadi.

Ushbu ma’lumotlar bazasini yanada kengaytirish orqali qo‘shimcha fleksiyalash naqshlarini qamrab olish va lemmatizatsiya hamda stemming algoritmlarini takomillashtirish uchun mashinaviy o‘rganish modellari bilan integratsiya qilish mumkin. Bundan tashqari, ushbu resurs avtomatlashtirilgan leksikografiya, mashinaviy tarjima va o‘zbek tilining morfologik tahlil vositalarini rivojlantirish kabi ilg‘or qo‘llanmalar uchun asos bo‘lib xizmat qilishi mumkin. Bunday rivojlanishlar uning tabiiy tilni qayta ishlash (NLP) tadqiqotlari va texnologiyalaridagi ahamiyatini oshirib, kengroq qo‘llanilishini ta’minlaydi.

FOYDALANILGAN ADABIYOTLAR:

1.Mengliev D, Barakhnin V, Abdurakhmonova N, Eshkulov M. Developing named entity recognition algorithms for Uzbek: Dataset insights and implementation. Data Brief 2024; 54: 110413.

2.Madatov K, Bekchanov S, Vičič J. Dataset of stopwords extracted from Uzbek texts. Data Brief 2022; 43.

3.Allaberdiev B, Matlatipov G, Kuriyozov E, Rakhmonov Z. Parallel texts dataset for Uzbek-Kazakh machine translation. Data Brief 2024; 53: 110194.

4.Emileva B, Kuhn L, Bobojonov I, Glauben T. The role of smartphone-based weather information acquisition on climate change perception

accuracy: Cross-country evidence from Kyrgyzstan, Mongolia and Uzbekistan. Clim Risk Manag 2023; 41: 100537.

5.Sharipov M, Sobirov O. Development of a Rule-Based Lemmatization Algorithm Through Finite State Machine for Uzbek Language. CEUR Workshop Proceedings, CEUR-WS 2022, 154 – 159.

6.Maksud S, Elmurod K, Ollabergan Y, Ogabek S. UzbekVerbDetection: Rule-based Detection of Verbs in Uzbek Texts. 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings, 2024, 17343 – 17347.

7.Sharipov M, Yuldashov O. UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language. CEUR Workshop Proceedings, CEUR-WS 2022, 137 – 144.

8.Sharipov M, Kuriyozov E, Yuldashev O, Sobirov O. UzbekTagger: The rule-based POS tagger for Uzbek language. arXiv preprint arXiv:230112711 2023;

9.Mercanoglu O, Keles H. AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset and Baseline Methods. IEEE Access 2020; 8: 181340–181355.

10.Akın Özçift Kamil Akarsu FY, Söylemez C. Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT): an empirical case study for Turkish. Automatika 2021; 62: 226–238.

11.Tocoglu M, Alpkocak A. TREMO: A dataset for emotion analysis in Turkish. J Inf Sci 2018; 44: 016555151876101.

12.Yeshpanov R, Varol HA. KazSAnDRA: Kazakh Sentiment Analysis Dataset of Reviews and Attitudes. Proceedings of the 2024 Joint

International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL 2024, 9657–9667.

13.Yeshpanov R, Khassanov Y, Varol HA. KazNERD: Kazakh Named Entity Recognition Dataset. Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association 2022, 417–426.

14.Abibullaeva AA, Kazbekova GN, Junisov NM. Keyword Extraction from Kazakh Text with Machine Learning Algorithms. Vestnik KazNPu imeni Abaya, Seriya ``Fiziko-matematicheskie nauki'' 2024; 85: 106–113.

15.Abidin Z, Junaidi A, Wamiliana. Text Stemming and Lemmatization of Regional Languages in Indonesia: A Systematic Literature Review. Journal of Information Systems Engineering and Business Intelligence 2024; 10: 217 – 231.

16.Boban I, Doko A, Gotovac S. Sentence retrieval using Stemming and Lemmatization with different length of the queries. Advances in Science, Technology and Engineering Systems 2020; 5.