

O'ZBEK TILIDAGI MATNLARNI AVTOMATIK DAVRLASH MASALALARI

Kenjayeva Marjona

1-kurs magistranti

mkenjayeva274@gmail.com

ToshDO'TAU

Annotatsiya. Ushbu maqolada o'zbek tilidagi matnlarni avtomatik davrlashning (periodizatsiyaning) lingvistik va algoritmik muammolari tadqiq etilgan. Tadqiqot davomida qadimgi turkiy, eski o'zbek va hozirgi o'zbek tili bosqichlarini farqlovchi leksik, morfologik va orfografik indikatorlar tahlil qilingan. Shuningdek, matnlarni xronologik tasniflashda zamonaviy kompyuter lingvistikasi usullari, jumladan, mashinali o'rganish va neyron tarmoqlar (NLP) modellarining samaradorligi ochib berilgan. Olingan natijalar o'zbek milliy korpusini takomillashtirish hamda tarixiy qo'lyozmalarni raqamli identifikatsiya qilish tizimlarini yaratish uchun poydevor bo'lib xizmat qiladi.

Kalit so'zlar: *o'zbek tili tarixi, avtomatik davrlash, kompyuter lingvistikasi, NLP lingvistik markerlar.*

Abstract: This article investigates the linguistic and algorithmic challenges of automatic periodization of texts in the Uzbek language. Throughout the study, the lexical, morphological, and orthographic indicators that distinguish the stages of Old Turkic, Old Uzbek (Chagatai), and Modern Uzbek languages are analyzed. Furthermore, the efficiency of modern computational linguistics methods, including machine learning and neural network (NLP) models, in the chronological classification of texts is demonstrated. The obtained results serve as a foundation for improving the Uzbek National Corpus and creating digital identification systems for historical manuscripts.

Keywords: *history of the Uzbek language, automatic periodization, computational linguistics, NLP (Natural Language Processing), machine learning, national corpus, text classification, linguistic markers.*

Kirish. Zamonaviy tilshunoslik va axborot texnologiyalari integratsiyasi natijasida yuzaga kelgan kompyuter lingvistikasi bugungi kunda milliy tillarning raqamli mavjudligini ta'minlashda asosiy vosita bo'lib xizmat qilmoqda. O'zbek tili o'zining ko'p asrlik taraqqiyoti davomida turli tarixiy bosqichlarni, leksik o'zgarishlarni va bir necha grafik tizimlarni bosib o'tgan murakkab agglyutinatив tillardan biridir. Shu sababli, o'zbek tilidagi ulkan matnlar massivini (qadimiy qo'lyozmalardan tortib zamonaviy elektron nashrlargacha) tizimlashtirish va ularning xronologik mansubligini aniqlash zarurati tug'ilmoqda. Matnlarni avtomatik davrlash (periodizatsiya) masalasi bu nafaqat texnik jarayon, balki tilning tarixiy evolyutsiyasini raqamli algoritmlar yordamida modellashtirishdir. Bugungi kunda global tarmoqda o'zbek tilidagi matnlar hajmi geometrik progressiya bilan ortib borayotgan bir sharoitda, ushbu matnlarning qaysi davrga tegishli ekanligini qo'lda (manual) aniqlash imkonsizdir. Bu esa sun'iy intellekt va mashinali o'qitish usullari yordamida matnlarni avtomatik klassifikatsiya qilishning dolzarbligini belgilaydi.

Ushbu tadqiqotning asosiy maqsadi o'zbek tilining tarixiy taraqqiyot bosqichlarini (Qadimgi turkiy, Eski o'zbek va Hozirgi o'zbek tili) tavsiflovchi lingvistik parametrlarni aniqlash va ular asosida matnlarni avtomatik davrlarga ajratish algoritmlarini ishlab chiqishdir. Tadqiqot obyekti sifatida tanlangan matnlar turli davrlar ijtimoiy-siyosiy hayoti, madaniyati va til qurilishini aks ettirishi bilan ahamiyatlidir. O'zbek tili matnlari bilan ishlashda avtomatik davrlash jarayoni bir qator o'ziga xos to'siqlarga duch keladi:

1. Alifbo va imlo dinamikasi. O'zbek yozuvining arab, lotin va kirill grafikalarida mavjudligi yagona tahlil bazasini yaratishda “grafik konvertatsiya” muammosini keltirib chiqaradi.

2. Leksik-semantik transformatsiya. Bir xil soʻz shakllarining turli asrlarda turlicha maʼno anglatishi yoki arxaik qatlamning yangi qatlam bilan aralashib ketishi (diffuziya).

3. Morfologik oʻzgaruvchanlik. Turkiy tillarga xos boʻlgan boy affiksatsiya tizimining tarixiy oʻzgarishlarini kompyuter modellariga “oʻrgatish” zarurati.

Tadqiqotning ilmiy yangiligi. Mazkur ishda ilk bor oʻzbek tili matnlari uchun “xronologik markerlar” tizimi ishlab chiqiladi. Bu tizim matndagi statistik maʼlumotlar (soʻzlar chastotasi) va chuqur lingvistik belgilar (grammatik konstruksiyalar) kombinatsiyasiga asoslanadi. Shuningdek, matnning yozilish uslubi va davriy qatlamini aniqlashda NLP (Natural Language Processing) texnologiyalarining oʻzbek tili morfologiyasiga moslashtirilgan modellari taklif etiladi. Tadqiqot natijalari quyidagi sohalarida tatbiq etilishi mumkin:

- Milliy korpus yaratish. Matnlarni avtomatik tarzda davriy teglash (tagging) va annotatsiya qilish.
- Raqamli kutubxonachilik. Nomaʼlum matnlar va qoʻlyozmalarning yaratilgan vaqtini soniyalar ichida aniqlash.
- Adabiyotshunoslik va tarix. Matnshunoslik tahlillarini avtomatlashtirish va qiyosiy-tarixiy metodlarni raqamlashtirish.

Asosiy qism

Oʻzbek tili evolyutsiyasining kompyuter tahlili uchun davriy tasnifi. Matnlarni avtomatik davrlashda oʻzbek tilining boy tarixi uchta asosiy lingvistik modelga ajratiladi. Bu modellar algoritmlar uchun tayanch nuqtasi boʻlib xizmat qiladi[5:34]. Qadimgi va eski turkiy til modeli (VII–XIV asrlar): Bu davr matnlari oʻzining oʻziga xos morfologik koʻrsatkichlari (masalan, joʻnalish kelishigining -qa/-ka shakllari, sifat-doshning -gan formasi) va leksik qatlamining sof turkiyligi bilan ajralib turadi. Eski oʻzbek tili modeli (XIV–XX asr boshlari): Bu davrda arab va fors tillaridan kirib kelgan oʻzlashmalarning haddan tashqari yuqoriligi, izofa konstruksiyalari va

klassik uslub ustunlik qiladi. Avtomatik tahlilda ushbu davr matnlari leksik ko'p qatlamliligi bilan murakkablik tug'diradi. Hozirgi o'zbek tili modeli (XX asrdan bugungi kungacha): Bu bosqichda terminologiyani xalqarolashuvi, sintaksisning soddalashishi va grammatik me'yorlarning standartlashuvi kuzatiladi.

Matnlarni avtomatik davrlashning lingvistik indikatorlari. Algoritm matnning davrini aniqlashi uchun undan ma'lum bir “belgilar”ni (features) ajratib olishi kerak. Tadqiqotda quyidagi indikatorlar asosiy o'rin tutadi:

- Morfologik markerlar[3:52-56]. Masalan, hozirgi o'zbek tilida faol bo'lgan -yapti affiksi eski matnlarda umuman uchramaydi. Buning o'rniga -durur yoki -ur kabi arxaik formalar tahlil qilinadi.
- Leksik-statistik ko'rsatkichlar. Matndagi so'zlarning uchrash chastotasi Zipf qonuni asosida tahlil qilinadi[4:167]. Eski matnlarda “ul”, “erdi”, “birla” kabi so'zlar yuqori chastotaga ega bo'lsa, zamonaviy matnlarda ularning o'rnini “u”, “edi”, “bilan” so'zlari egallaydi.
- Grafik va orfografik belgilar. O'zbek tili tarixidagi alifbo almashishlari (arab-lotin-kirill) avtomatik davrlashning eng birlamchi va ishonchli filtri hisoblanadi.

1-jadval. Mant davrini aniqlash uchun tahliliy mezonlar

Tahlil mezonlari	Qadimgi va eski turkiy til	Eski o'zbek tili (Chig'atoy)	Hozirgi o'zbek tili
Dominant alifbo	Runik, Uyg'ur	Arab (Eski o'zbek yozuvi)	Kirill, Lotin
Xarakterli so'zlar	<i>ul, erdi, adig', biqin</i>	<i>kavkab, xurshid, dahr, gulyuz</i>	<i>kosmos, iqtisodiyot, tizim, smartfon</i>
Morfologik markerlar	-qa/-ka, -gan, -sar	-durur, -lig', -urlar	-yapti, -gan, -moqda
O'zlashma qatlam	Sof turkiy, mo'g'ulcha	Arabcha, forsha (juda ko'p)	Ruscha, inglizcha, xalqaro
Sintaktik qurilish	Sodda, o'zakdosh so'zlar	Murakkab izofalar, uzun gaplar	Standartlashgan, qisqa gaplar

Klassifikatsiya algoritmlari va texnik yondashuvlar. Matnlarni davrlarga ajratish masalasi mashinali o'rganishda Klassifikatsiya vazifasi sifatida quyidagi metodlar yordamida yechiladi[1:84]. Statistik modellar (TF-IDF va Naive Bayes): Matndagi har bir so'zning ma'lum bir davr uchun “vazni” hisoblab chiqiladi. Agar matnda ma'lum bir davrga xos so'zlar konsentratsiyasi yuqori bo'lsa, model uni o'sha davrga mansub deb topadi.

Neyron tarmoqlar va Transformerlar (BERT): O'zbek tili uchun moslashtirilgan mBERT (Multilingual BERT) modellari matnning kontekstual ma'nosini tahlil qiladi[2:75-78]. Bu modellar so'zning nafaqat shaklini, balki uning matn ichidagi ma'noviy yuklamasini ham hisobga olgan holda davrlash aniqligini 90% dan oshirish imkonini beradi.

N-gram tahlili: Faqat so'zlarni emas, balki harf va so'z birikmalarining ketma-ketligini (bi-gram, tri-gram) tahlil qilish orqali tilning stilistik o'zgarishlari aniqlanadi.

Avtomatik davrlashdagi texnik va nazariy muammolar. Tadqiqot jarayonida o'zbek tili uchun xos bo'lgan quyidagi muammolar tahlil qilindi:

1. Dataset (ma'lumotlar to'plami) tanqisligi. Ayniqsa, o'rta asrlarga oid matnlarning kompyuter o'qiydigan (OCR qilingan) sifatli bazasining kamligi.
2. Stilistik diffuziya. Zamonaviy yozuvchilarning tarixiy asarlarida qadimiy tildan foydalanishi algoritmlarni chalg'itishi mumkin. Bu muammoni yechish uchun matnning “chuqur sintaktik tuzilishi” tahlili zarur.
3. Transliteratsiya xatoliklari. Arab yozuvidan lotin yoki kirill yozuviga o'girishda fonetik xususiyatlarning yo'qolishi davrlash aniqligiga salbiy ta'sir ko'rsatadi.

Tajriba natijalari va samaradorlik tahlili. O'tkazilgan tajribalar shuni ko'rsatadiki, o'zbek tili matnlarini avtomatik davrlashda eng yuqori natija morfologik tahlil + neyron tarmoqlar kombinatsiyasida kuzatiladi. Statistik usullar

esa qisqa matnlarni (masalan, ijtimoiy tarmoq postlari yoki qisqa xabarlar) davrlashda kutilgan samarani bermaydi.

Xulosa. O'zbek tilidagi matnlarni avtomatik davrlash (periodizatsiya) masalasi an'anaviy tilshunoslik, manbashunoslik va zamonaviy kompyuter texnologiyalarining o'zaro integratsiyasini talab etuvchi o'ta muhim ilmiy yo'nalishdir. Mazkur mavzu doirasida amalga oshirilgan tahlil va o'rganishlar asosida quyidagi ilmiy va amaliy xulosalarga kelindi.

Matnning tarixiy yoshini aniqlashda faqatgina leksik qatlamga (eskirgan yoki yangi so'zlarga) tayanish yetarli aniqlikni bermaydi. Tadqiqotlar shuni ko'rsatdiki, davrlash tizimining eng ishonchli tayanchi bu — morfologik indikatorlar (tarixiy affikslar) va sintaktik konstruksiyalar. Ayniqsa, o'zbek tili kabi agglyutinativ tillarda qo'shimchalar o'zgarishini modellashtirish mashinali o'qitish algoritmlari uchun eng barqaror xronologik filtr bo'lib xizmat qiladi. Matnlarni davrlarga tasniflashda an'anaviy statistik metodlar (masalan, Naive Bayes yoki n-gram) qisqa va aniq ajratilgan matnlarda yaxshi ishlasa-da, murakkab va aralash uslubdagi asarlarda o'zini oqlamaydi. Hozirgi kunda matnning chuqur semantikasini tushunuvchi neyron tarmoqlar (xususan, BERT arxitekturasiga asoslangan transformer modellar) eng yuqori samaradorlikni namoyon etmoqda. Ular so'zning shunchaki shaklini emas, balki uning matn ichidagi tarixiy kontekstini ham baholay oladi. O'zbek tili tahlilida tizimning muvaffaqiyatiga to'sqinlik qiluvchi asosiy omil — turli alifbolar (arab, lotin, kirill) almashinuvi natijasida yuzaga kelgan “grafik va transliteratsion yo'qotishlar” hamda qadimiy matnlarning sifatli raqamlashtirilgan o'quv korpuslarining (dataset) yetishmasligi.

Amaliy ahamiyati va kelgusi rejalar: Ishlab chiqilayotgan avtomatik davrlash modellari nafaqat sun'iy intellekt va kiberxavfsizlik (mualliflikni va yozilish vaqtini aniqlash) sohalarida, balki O'zbek tili milliy korpusini takomillashtirishda ham muhim dasturiy poydevor bo'ladi. Kelajakda ushbu tizimni yanada takomillashtirish

uchun o'rta asrlar va chig'atoy tilidagi qo'lyozmalarni kompyuter o'qiydigan formatga o'tkazish (OCR texnologiyalarini rivojlantirish), matnni nafaqat asrlar bo'yicha, balki o'n yilliklar qamrovida (masalan, 1920-yillar tili, 1980-yillar tili) aniqroq davrlash imkonini beruvchi mikroxtnologik algoritmlarni ishlab chiqish, tarixiy matnlarning grammatik bazasini to'liq qamrab oluvchi elektron tezaurus va lug'atlarni mashinali o'rganish modellariga integratsiya qilish zarur. Xulosa qilib aytganda, o'zbek tili matnlarini avtomatik davrlash – milliy madaniy va ilmiy merosimizni kelajak raqamli texnologiyalariga moslashtirish, tilimiz tarixini aniq, ishonchli va tezkor tahlil qilish imkonini beruvchi istiqbolli tadqiqot maydonidir.

Foydalanilgan adabiyotlar ro'yxati:

1. Alayev R. Tabiiy tillarga ishlov berishda matnlarni avtomatik tasniflash usullari // “Kompyuter lingvistikasi va sun'iy intellekt” respublika ilmiy-amaliy konferensiyasi materiallari. – 2022.
2. Elov B. O'zbek tili raqamli infratuzilmasini yaratishda zamonaviy NLP modellari (BERT, RoBERTa) tahlili // Raqamli iqtisodiyot va IT xabarnomasi. – 2023.
3. Hamroyeva Sh. O'zbek tili morfologik analizatori lingvistik ta'minotini yaratish tamoyillari: filol. fan. bo'yicha falsafa doktori (PhD) dissertatsiyasi avtoreferati. – Buxoro, 2020. – 48 b.
4. Jurafsky D., Martin J. H. Speech and Language Processing. – 3rd ed. (Draft). – Stanford: Stanford University, 2023.
5. Mengliyev B. O'zbek tilida korpus lingvistikasi va uning nazariy asoslari. – Toshkent: Zamin nashr, 2021. – 220 b.