NLP USULLARI YORDAMIDA TIL O‘ZGARISHINI DIAXRONIK KORPUS ASOSIDA O‘RGANISH

Xusainova Zilola Yuldashevna,
f.f.f.d. (PhD), dotsent v.b.
xusainovazilola@navoiy-uni.uz
ToshDO‘TAU

Yangibayeva Surayyo,
II bosqich magistrant
surayyoyangibayeva4@gmail.com
ToshDO‘TAU

Annotatsiya. Mazkur maqolada o‘zbek tilidagi diaxron korpus asosida til o‘zgarishlarini tabiiy tilni qayta ishlash (NLP) usullari yordamida tahlil qilish masalasi yoritiladi. Tadqiqotda 1920–2025-yillar oralig‘idagi matnlar asosida shakllantirilgan diaxron korpusdan foydalanildi. Korpus tarkibidagi matnlar avtomatik ravishda normalizatsiya qilinib, morfologik tahlil, lemmatizatsiya, POS-teglash va statistik modellashtirish bosqichlaridan o‘tkazildi. Tadqiqot natijalari leksik, morfologik va sintaktik qatlamlarda sezilarli diaxron o‘zgarishlar mavjudligini ko‘rsatdi. Xususan, ayrim tarixiy birliklarning iste’moldan chiqishi, yangi terminlarning paydo bo‘lishi, Type-Token Ratio (TTR) ko‘rsatkichining pasayishi hamda gap uzunligining qisqarishi kuzatildi. Tadqiqot natijalari o‘zbek tilining tarixiy taraqqiyotini kompyuter lingvistikasi metodlari asosida tahlil qilish imkoniyatlarini kengaytiradi.

Kalit so‘zlar: *diaxron korpus, NLP, korpus lingvistikasi, lemmatizatsiya, morfologik tahlil, POS-teglash, TTR, o‘zbek tili.*

Abstract. This article examines the issue of analyzing language changes based on the diachronic corpus of the Uzbek language using natural language processing (NLP) methods. The study utilized a diachronic corpus formed on the basis of texts from the 1920s to 2025. The texts within the corpus were automatically normalized and underwent stages of morphological analysis, lemmatization, POS tagging, and statistical modeling. The research results demonstrated the presence of

significant diachronic changes in the lexical, morphological, and syntactic layers. In particular, the decline of certain historical units, the emergence of new terms, a decrease in the Type-Token Ratio (TTR) indicator, and a reduction in sentence length were observed. The research results expand the possibilities of analyzing the historical development of the Uzbek language based on the methods of computer linguistics.

Keywords: *diachronic corpus, NLP, corpus linguistics, lemmatization, morphological analysis, POS tagging, TTR, Uzbek language.*

Kirish. So‘nggi yillarda tabiiy tilni qayta ishlash (Natural Language Processing, NLP) sohasida yirik hajmdagi lingvistik korpuslardan foydalanish tilshunoslik tadqiqotlarining muhim yo‘nalishlaridan biriga aylandi. Ayniqsa, diaxron korpuslar tilning vaqt davomida qanday o‘zgarishini statistik va lingvistik jihatdan kuzatish imkonini beradi [6:304]. Ingliz, turk va rus tillarida yaratilgan yirik diaxron korpuslar asosida leksik evolyutsiya, semantik o‘zgarishlar va grammatik tendensiyalar bo‘yicha ko‘plab tadqiqotlar amalga oshirilgan. O‘zbek tili esa kompyuter lingvistikasi nuqtai nazaridan hali kam resursli tillar qatoriga kiradi. Ayniqsa, tarixiy matnlarni avtomatik qayta ishlash, diaxron teglash va statistik modellashtirish bo‘yicha tadqiqotlar yetarli emas. O‘zbek tilining bir asrlik taraqqiyoti davomida yozuv tizimining bir necha marotaba o‘zgarishi, leksik qatlamlarning yangilanishi hamda ijtimoiy-siyosiy omillar ta’siri til tizimida sezilarli transformatsiyalarni yuzaga keltirgan.

O‘zbek tilining diaxron korpusi NLP usullari yordamida til o‘zgarishlarini avtomatik tahlil qiladi. Tadqiqot doirasida korpus matnlari normalizatsiya qilinib, lemmatizatsiya, morfologik tahlil, POS-teglash va statistik modellashtirish usullari qo‘llanildi.

Diaxron korpuslar asosida til taraqqiyotini o‘rganish zamonaviy korpus lingvistikasining muhim yo‘nalishlaridan biridir. Corpus of Historical American

English (COHA) ingliz tilidagi tarixiy o‘zgarishlarni o‘rganishda keng qo‘llanadi [5:1489–1501]. Turk tilida yaratilgan Turkronicles korpusi esa XX asrdan boshlab tilning leksik va statistik rivojlanishini kuzatish imkonini beradi [2:6958–6966]. Kam resursli tillarda diaxron NLP tadqiqotlari ko‘proq qoidaga asoslangan yondashuvlarga tayangan holda amalga oshiriladi [6]. Agglyutinativ tillarda morfologik analizator va lemmatizatorlar til birliklarini aniqlashda muhim vosita hisoblanadi [4:189–198]. O‘zbek tilida esa POS-teglash va morfologik tahlil bo‘yicha ayrim tadqiqotlar mavjud bo‘lsa-da, diaxron korpus asosidagi kompleks NLP infratuzilmasi hali to‘liq shakllanmagan [1:64-68].

Korpusni shakllantirish. Tadqiqot uchun 1920–2025-yillar oralig‘idagi badiiy adabiyotlar, gazeta-jurnallar, ilmiy maqolalar, rasmiy hujjatlar va internet matnlari asosida diaxron korpus shakllantirildi. Matnlar OCR texnologiyasi yordamida raqamlashtirildi va yagona lotin yozuviga normalizatsiya qilindi.

Matnlarni qayta ishlash jarayonida quyidagi bosqichlar bajarildi:

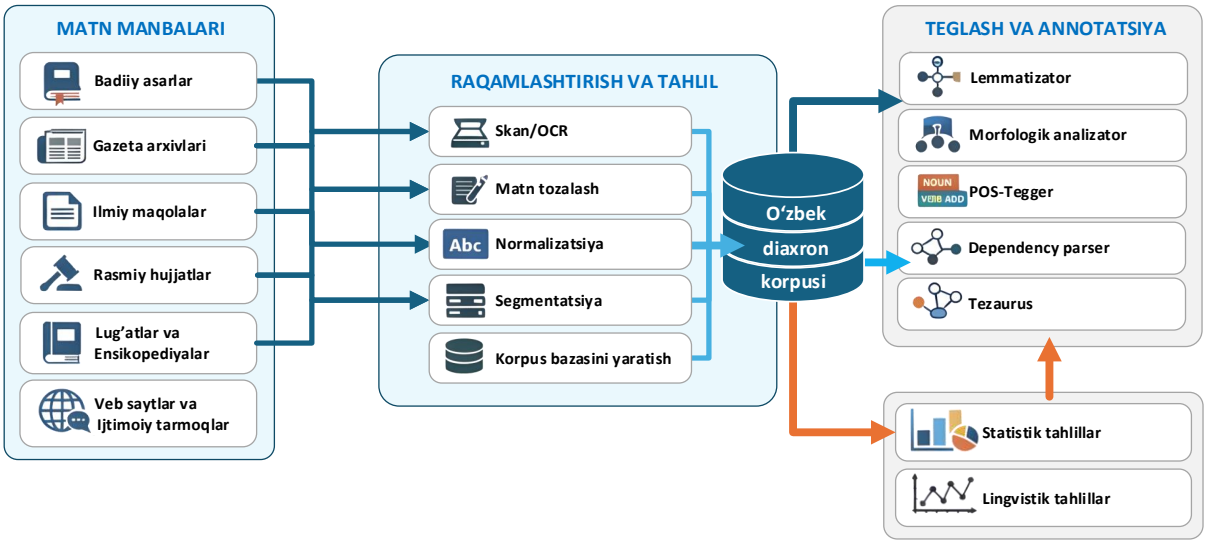
- grafik normalizatsiya;
- ortiqcha belgilarni tozalash;
- tokenizatsiya;
- lemmatizatsiya;
- morfologik tahlil;
- POS-teglash.

NLP modullari. Diaxron korpusni avtomatik teglash uchun quyidagi NLP vositalaridan foydalaniladi:

1. Lemmatizator;
2. Morfologik analizator;
3. POS-teglagich;
4. Sintaktik analizator;

5. Diaxron tezaurus.

Korpusni qayta ishlashning umumiy arxitekturasi 1-rasmda keltirilgan.



1-rasm. Diaxron korpusni NLP asosida qayta ishlash sxemasi

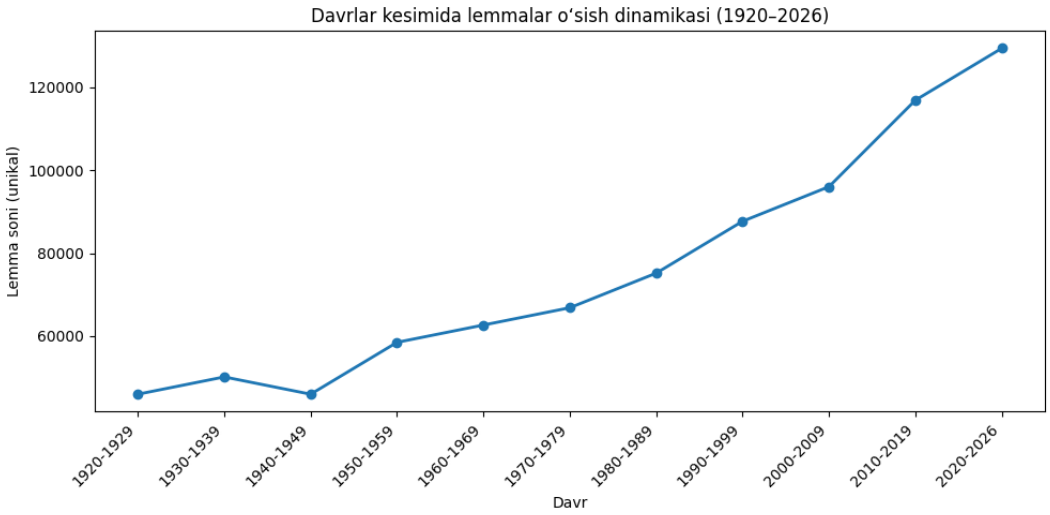
Natijalar va tahlil

Token va lemma dinamikasi

Tahlil natijalariga ko‘ra, korpusdagi tokenlar soni vaqt o‘tishi bilan sezilarli ravishda oshgan. Ayniqsa, 1990-yildan keyingi davrda internet va raqamli manbalar hisobiga korpus hajmi keskin kengaygan.

2-rasm. Tokenlar o‘sish dinamikasi (1920–2025)

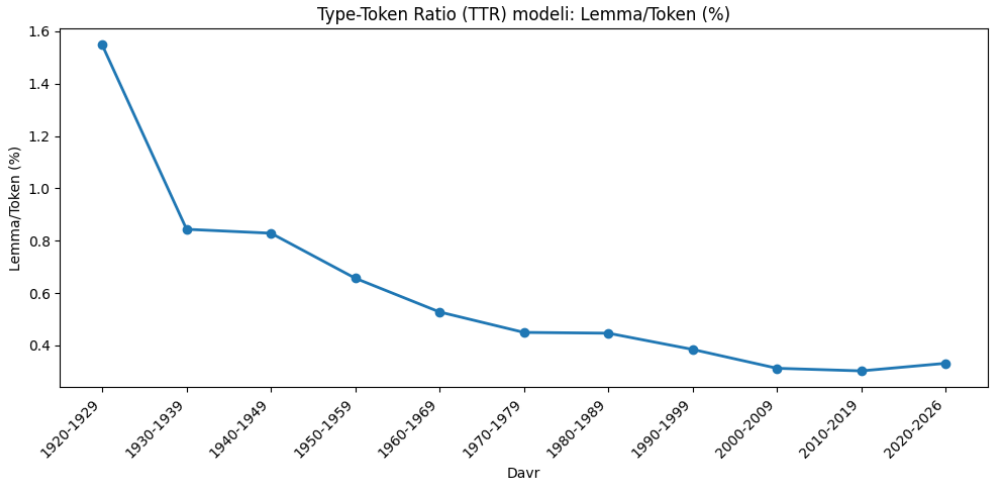
Lemmalar sonining o‘sishi tokenlar soniga nisbatan barqarorroq bo‘lib, bu til tizimining ichki strukturaviy muvozanatini ko‘rsatadi.



3-rasm. Lemma o‘sish dinamikasi

Type-Token Ratio (TTR) tahlili

Korpusda leksik xilma-xillikni aniqlash uchun Type-Token Ratio (TTR) ko‘rsatkichi hisoblandi. Natijalar dastlabki davrlarda TTR yuqori bo‘lganini, keyingi davrlarda esa korpus hajmining ortishi natijasida pasayganini ko‘rsatdi.

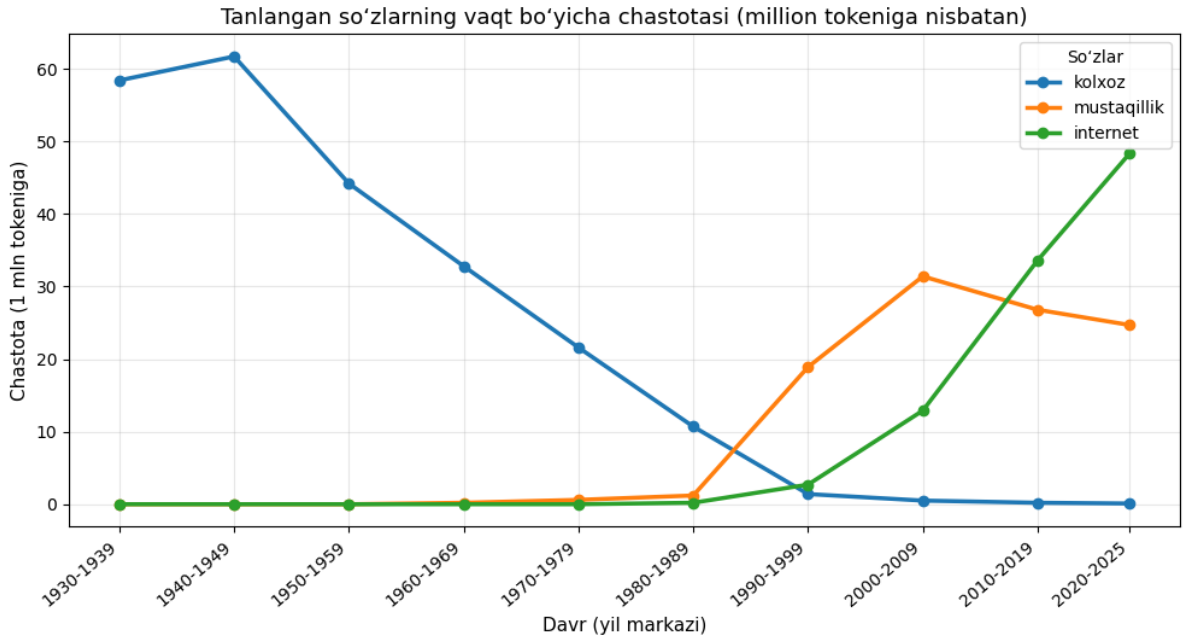


4-rasm. Type-Token Ratio (TTR) modeli

TTR ko‘rsatkichining pasayishi korpus lingvistikasidagi klassik statistik qonuniyatlar bilan mos keladi. Biroq 2020-yillarda yangi internet terminologiyasi va texnologik leksika hisobiga TTR qiymatida qisman o‘sish kuzatildi.

Leksik o‘zgarishlar

Diaxron korpus asosida ayrim tarixiy va zamonaviy birliklarning chastota dinamikasi tahlil qilindi. “Kolxoz”, “mustaqillik” va “internet” kabi so‘zlarning turli davrlardagi qo‘llanish chastotasi tilning ijtimoiy-siyosiy taraqqiyot bilan uzviy bog‘liqligini ko‘rsatadi.



5-rasm. Ayrim so‘zlarning vaqt bo‘yicha qo‘llanish chastotasi

Tahlil natijalari shuni ko‘rsatdiki, sovet davriga oid birliklar zamonaviy matnlarda deyarli uchramaydi, aksincha internet va texnologik terminologiya XXI asrda keskin faollashgan.

Xulosa

Olingan natijalar o‘zbek tilining diaxron taraqqiyoti jamiyatdagi siyosiy, texnologik va madaniy o‘zgarishlar bilan chambarchas bog‘liq ekanligini ko‘rsatdi. Ayniqsa, sovet davriga oid terminologiyaning iste’moldan chiqishi va mustaqillik davri hamda internet diskursiga oid yangi birliklarning paydo bo‘lishi leksik qatlamdagi transformatsiyalarni yaqqol namoyon etdi. Gap uzunligining qisqarishi, affikslar sonining kamayishi va nominal birliklarning ortishi zamonaviy kommunikatsiyaning tezkor va ixcham shaklga o‘tayotganini bildiradi. Diaxron

korpus asosidagi NLP tahlillari tarixiy matnlarni avtomatik qayta ishlash, semantik evolyutsiyani modellashtirish hamda kelajakdagi sun’iy intellekt tizimlari uchun muhim empirik baza yaratadi.

Mazkur tadqiqotda o‘zbek tilining diaxron korpusi asosida NLP usullari yordamida til o‘zgarishlari tahlil qilindi. Tadqiqot natijalari quyidagilarni ko‘rsatdi:

- o‘zbek tilining leksik tarkibi tarixiy davrlar davomida sezilarli o‘zgarishga uchragan;
- TTR ko‘rsatkichining pasayishi korpus hajmi ortishi bilan bog‘liq;
- zamonaviy matnlarda nominal uslub kuchaygan;
- gap uzunligi va morfologik murakkablik qisqargan;
- diaxron korpuslar til evolyutsiyasini statistik modellashtirish uchun samarali vosita hisoblanadi.

Kelgusida ushbu korpus asosida semantik o‘zgarishlarni neyron tarmoqlar yordamida modellashtirish hamda tarixiy matnlar uchun maxsus NLP modellarini yaratish rejalashtirilmoqda.

Foydalanilgan adabiyotlar ro‘yxati

1. Boltayevich E.B., Samariddinovich S.S., Mirdjonovna K.S., Adalı E., Yuldashevna X.Z. POS tagging of Uzbek text using Hidden Markov Model // 2023 8th International Conference on Computer Science and Engineering (UBMK). – 2023. – P. 63–68.

2. Davies M. The Corpus of Historical American English (COHA): 200 years of historical data for linguistic analysis. Brigham Young University Press. 2020. – P. 6958–6966.

3. Hilpert M., Gries S. Quantitative approaches to diachronic corpus linguistics // The Cambridge Handbook of English Historical Linguistics. – Cambridge University Press, 2017.

4. Hämäläinen M., Partanen N. and Alnajjar K. Lemmatization of historical old literary Finnish texts in modern orthography. In Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1: conférence principale. 2021. – P. 189–198.

5. Hamilton W. L., Leskovec, J., and Jurafsky D. Diachronic word embeddings reveal statistical laws of semantic change. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, August. – P. 1489–1501.

6. McEnery T. and Baker P. Corpus Linguistics and 19th-Century Historical Texts. Edinburgh University Press. 2016. – 304 p.

7. Öksüz H., Çöltekin Ç. Turkronicles: A diachronic corpus for Turkish language research // Journal of Turkic Linguistics. – 2025.