DIAXRON KORPUS YARATISH BOSQICHLARI (MUSTAQILLIK DAVRI NASHRLARI MISOLIDA)

Xusainova Zilola Yuldashevna,
filologiya fanlari falsafa doktori (PhD)
xusainovazilola@navoiy-uni.uz
ToshDO‘TAU

Yangibayeva Surayyo Gulimboyevna,
Kompyuter lingvistikasi mutaxasisligi magistranti
surayyoyangibayeva4@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada o‘zbek tilidagi mustaqillik davri nashrlari asosida diaxron korpus yaratish metodologiyasi va uning ilmiy ahamiyati tahlil qilingan. Diaxron korpuslar tilning vaqt o‘tishi bilan yuz berayotgan o‘zgarishlarini aniqlash, tilning rivojlanish dinamikasini modellashtirish va kelajakdagi o‘zgarishlarni prognoz qilish imkonini beradi. Tadqiqot jarayonida materiallarni tanlash, ma’lumotlarni to‘plash, raqamlashtirish, matn tozalash, normalizatsiya, avtomatik belgilash, sintaktik-semantik tahlil, qo‘lda tekshirish, indekslash va natijalarni vizualizatsiya qilish bosqichlari batafsil bayon etilgan. Maqolada ta’kidlangan usullar va yondashuvlar kelajakda til evolyutsiyasini yanada kengroq va aniqlik bilan o‘rganishga zamin yaratadi.

Abstract. This article analyzes the methodology for creating a diachronic corpus based on the publications from the Independence period in the Uzbek language and its scientific significance. Diachron corpora allow identifying language changes over time, modeling the dynamics of language development, and predicting future changes. The research process elaborates on the stages of material selection, data collection, digitization, text cleaning, normalization, automatic tagging, syntactic-semantic analysis, manual verification, indexing, and visualization of results. The methods and approaches highlighted in the article provide a foundation for more comprehensive and precise studies of language evolution in the future.

Аннотация. В данной статье анализируются методология создания диахронического корпуса на основе публикаций периода независимости на узбекском языке и его научная значимость. Диахронические корпусы позволяют выявлять изменения языка с течением времени, моделировать динамику развития языка и прогнозировать будущие изменения. В процессе исследования подробно рассматриваются этапы отбора материалов, сбора данных, оцифровки, очистки текста, нормализации, автоматической разметки, синтаксико-семантического анализа, ручной проверки, индексирования и

визуализации результатов. Представленные методы и подходы создают основу для более широкого и точного изучения эволюции языка в будущем.

Kalit so‘zlar. *diaxron korpus, mustaqillik davri nashrlari, NLP, korpus lingvistikasi, ma’lumotlar bazasi.*

Kirish

Diaxron korpuslar – ma’lum bir davr oralig‘idagi til hodisalarini vaqt o‘tishi bilan izchil kuzatish imkonini beruvchi, lingvistik jihatdan annotatsiyalangan matnlar to‘plamidir. Ushbu korpuslar tilning vaqt o‘tishi bilan yuz berayotgan o‘zgarishlarini aniqlashda asosiy vosita hisoblanadi. NLP tadqiqotlarida ularning ahamiyati tilning dinamik evolyutsiyasini modellashtirishda namoyon bo‘lib, zamonaviy algoritmlar orqali tilning tarixiy, madaniy va ijtimoiy kontekstlari hamda semantik, sintaktik va stilistik jihatlari chuqur o‘rganilishi mumkin. Mustaqillik davri nashrlari misoli orqali diaxron korpuslarning dolzarbligi yanada yaqqol ko‘rinadi, chunki ushbu davr matnlari til evolyutsiyasining muhim bosqichlarini ifodalaydi. Masalan, **1990–2000**-yillardagi matnlarda *“mustaqillik”, “milliy g‘urur”, “taraqqiyot”* kabi leksemalarning chastotasi va kollokatsiyalaridagi o‘zgarishlar til siyosatining yo‘nalishini aks etadi.

NLP tadqiqotlarida diaxron korpuslar **til modellarini ishlab chiqish**, **mashina tarjimasi** va **matn tasnifi** kabi vazifalar uchun qimmatli ma’lumot manbai sifatida xizmat qiladi. Diaxron korpuslar tilning rivojlanish dinamikasini aniqlash orqali o‘zbek tilidagi kelajakdagi o‘zgarishlarni taxmin qilish uchun asos yaratadi. Diaxron korpuslar yordamida yaratilgan modellar til o‘zgarishlarini aniqlik bilan aks ettirish imkonini beradi. Shuningdek, bunday yondashuv til resurslarini kengaytirish va tarixiy matnlarni kompleks tahlil qilish imkoniyatini beradi. Bunday korpuslarning yetishmasligi – nafaqat lingvistika, balki ma’lumotlar fanlari uchun ham cheklov hisoblanadi, chunki ular tarixiy kontekstni hisobga olgan holda aniqroq prognozlashtirishni talab qiladi. Demak, mustaqillik davri materiallari asosida yaratilgan diaxron korpuslar – til tarixini raqamlashtirish, milliy ma’naviy merosni saqlash va AI tizimlarini lingvistik jihatdan moslashtirish yo‘lidagi muhim qadamdir. Diaxron korpuslarni yaratish nafaqat tilshunoslik, balki kompyuter lingvistikasi va sun’iy intellekt sohalarida ham yangi yondashuvlarni ishlab chiqishga turtki bo‘ladi.

Diaxron korpuslar tilning dinamik evolyutsiyasini **obyektiv va kvantitativ** usulda o‘rganish imkoniyatini yaratadi, bu esa zamonaviy NLP tizimlarini tarixiy kontekstga moslashtirishda hal qiluvchi rol o‘ynaydi. Mustaqillik davri matnlari misolida korpuslar **sotsiolingvistik o‘zgarishlar** (masalan, sovet terminologiyasidan milliy leksikaga o‘tish)ni tahlil qilish uchun noyob manba hisoblanadi. Diaxron korpuslar vaqt o‘tishi bilan semantik siljishlarni (masalan, “demokratiya” so‘zining 1991 va 2023-yillardagi konnotatsiyalaridagi farq)

avtomatik aniqlash imkonini beradi. Korpuslarning dolzarbligi shundaki, ular nafaqat lingvistik, balki madaniy merosni raqamli formatda saqlash, ya’ni non-fizaviy arxiv vazifasini ham o‘taydi. Mustaqillik davri gazetalari, qonun hujjatlari va badiiy asarlarini qamrab olgan korpus milliylik yordamida xaritalashtirish imkoniyatini ochadi. Hozirgi kunda o‘zbek tilida bunday korpusning mavjud emasligi til siyosati va ta’lim dasturlarini empirik ma’lumotlar asosida ishlab chiqishni qiyinlashtirmoqda.

Diaxron korpuslarni yaratish bo‘yicha amalga oshirilgan tadqiqotlarda tilning tarixiy rivojlanishi va o‘zgarishlarini tahlil qilishga katta e’tibor qaratilgan. Ushbu ishlar asosan zamonaviy til namunalariga yo‘naltirilgan bo‘lib, tarixiy materiallarning chuqur o‘rganilishi kamroq e’tiborga olingan. Mavjud metodologiyalar orasida leksik-semantik tahlil, sintaktik struktura va stilistik xususiyatlarni o‘rganish usullari keng qo‘llaniladi. Shu bilan birga, ko‘plab ilmiy ishlar tarixiy kontekstni hisobga olmagan holda, zamonaviy til modellari asosida ishlab chiqilgan. Mustaqillik davri nashrlari esa mavjud korpuslarda yetarlicha aks ettirilmagan, bu esa ilmiy bo‘shliqni yuzaga keltiradi.

Diaxron korpuslar yaratish metodologiyasi so‘nggi 20 yil ichida korpus lingvistikasi va NLPning keskin rivojlanishi tufayli keng qamrovli tadqiqotlarga aylandi[1]. Xalqaro miqyosda COHA (Corpus of Historical American English) kabi loyihalar til evolyutsiyasini 200 yillik ma’lumotlar asosida o‘rganishga imkon beradi, ammo Markaziy Osiyo tillari uchun shunga o‘xshash tizimlar deyarli mavjud emas[2]. O‘zbek tilidagi korpuslar (masalan, O‘zbek Milliy Korpusi) asosan zamonaviy matnlarni (2000–2023) qamrab oladi va mustaqillikning dastlabki yillaridagi materiallarni sistematik tarzda o‘z ichiga olmaydi[3]. Turk tillari diaxron korpuslari (masalan, Turk Tarixiy Korpusi (Türkçenin tarihsel derlemi) faqat grammatik o‘zgarishlarni o‘rganishga qaratilgan, leksik-semantik jihatlar esa chetda qolmoqda[4].

Xususan, 1991–2000-yillardagi gazetalar, qonun hujjatlari va she’riy merosning katta qismi hali raqamlashtirilmagan yoki fragmentar tarzda tarqalgan holda saqlanmoqda, bu esa ularni korpusga kiritishni qiyinlashtiradi.

Hozirgi metodologiyalar ko‘pincha statik analizga asoslangan (masalan, faqat bitta yil materiallarini o‘rganish), bu esa uzluksiz diaxron trendlarni aniqlash imkonini cheklaydi. Korpus yaratish jarayonida ***eski nashrlarni raqamlashtirish***, ***tozalash*** va ***tahlil qilishda murakkabliklar*** kuzatilmoqda. Shu sababli, yangi yondashuvlar va innovatsion metodlar qo‘llanilishi zarur. Sohada mavjud bo‘lgan korpuslarning formatlari va tarkibi bir xil emas, ba’zilarida esa mustaqillik davri materiallari yetarlicha qamrab olinmagan. Mavjud ilmiy adabiyotlarda tilning rivojlanishini statistik va model asosida tahlil qilish usullari yetarlicha rivojlanmagan. Shu bois, mustaqillik davri nashrlarining o‘ziga xos xususiyatlarini

hisobga olgan tadqiqotlar olib borish dolzarb hisoblanadi. Ushbu maqolada taklif etilayotgan diaxron korpus 32 yillik mustaqillik davrini (1991-2023) qamrab oladi va janrlararo diversifikatsiya (gazeta, qonun, she’riyat) orqali tilning ko‘p qirrali evolyutsiyasini ko‘rsatadi.

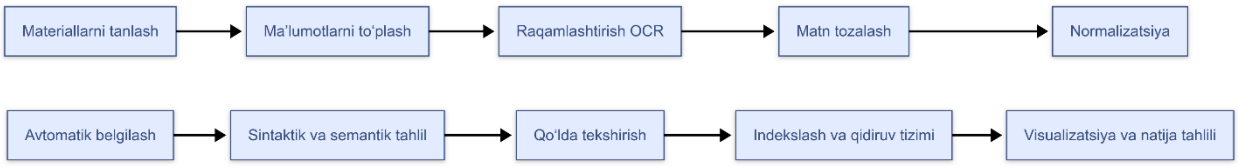
Diaxron korpus yaratish (1-jadval) jarayoni tilning tarixiy rivojlanishini chuqur o‘rganish uchun muhim vositadir. Ushbu jarayonda **birinchi bosqich** sifatida ***mustaqillik davri nashrlari tanlab olinib***, ularning tematik va madaniy ahamiyati tahlil qilinadi. **Ma’lumotlarni to‘plash** jarayonida nashrlar arxivlardan, kutubxonalardan va onlayn resurslardan yig‘iladi. To‘plangan materiallar raqamlashtirish bosqichida OCR texnologiyalari yordamida elektron formatga o‘tkaziladi. Raqamlashtirishdan so‘ng, **matnlar tozalash** jarayonidan o‘tkazilib, noaniq belgilar va xatoliklar tuzatiladi. Keyingi bosqichda, **matnlar normalizatsiya qilinib**, tilshunoslik qoidalari asosida standartlashtiriladi. Diaxron korpus yaratishda leksikografik va morfologik teglash muhim o‘rin egallaydi. Avtomatik teglash vositalari yordamida so‘zlar va ularning qismlarga ajratilishi amalga oshiriladi. Korpus tarkibini yanada boyitish maqsadida **sintaktik tahlil metodlari** ham qo‘llanilishi mumkin. Sintaktik tahlil orqali gap tuzilishi aniqlanib, **tilning muayyan davrdagi grammatik xususiyatlari** o‘rganiladi. Shu bilan birga, **semantik tahlil metodlari** yordamida matnlarning ma’nosi va kontekstual o‘zgarishlari aniqlanadi. Korpusning sifatini ta’minlash uchun qo‘lda tekshirish va qayta ishlash jarayonlari ham muhim hisoblanadi. Yaratilgan ma’lumotlar bazasi doirasida **qidiruv va indekslash tizimlari** ishlab chiqilib, ma’lumotlarga tezkor kirish imkoniyati yaratiladi. Indekslash jarayoni orqali foydalanuvchilar kerakli ma’lumotlarni oson topish imkoniga ega bo‘ladi.

Diaxron tahlil bosqichida vaqt o‘tishi bilan tilning o‘zgarish tendensiyalari batafsil o‘rganiladi. Ushbu tahlillar til evolyutsiyasini, leksik boylik va sintaktik strukturalarning o‘zgarishini ochib beradi. Korpus yaratishning so‘nggi bosqichida **natijalarni vizualizatsiya qilish orqali** tadqiqotchilarga qulay interfeys taqdim etiladi. Natijada, mustaqillik davri nashrlari asosida yaratilgan diaxron korpus tilshunoslik va NLP tadqiqotlari uchun yangi imkoniyatlar ochib beradi.

1-jadval: Diaxron korpus yaratish bosqichlari

Bosqich	Tavsif	Asosiy vazifalar
1. Materiallarni tanlash	Mustaqillik davri nashrlari namunalarini aniqlash va tanlash	Tematik va madaniy ahamiyatni aniqlash
2. Ma’lumotlarni to‘plash	Tanlangan ma’lumotlarni arxivlar, kutubxonalar va onlayn resurslardan yig‘ish	Ma’lumot bazasini shakllantirish
3. Raqamlashtirish	Matnlarni raqamli formatga o‘tkazish	Matnni elektron ko‘rinishga keltirish

4. Matnni tozalash	Raqamlashtirilgan matnlardagi xatoliklar va noaniqliklarni bartaraf etish	Xatoliklarni tuzatish, formatlashni yaxshilash
5. Normalizatsiya	Matnlarni standartlashtirish, tilshunoslik qoidalariga moslashtirish	Terminologiya va yozuv qoidalarini bir xil holga keltirish
6. Avtomatik teglash	Matn bo‘ylab leksik, morfologik va grammatik teglashni amalga oshirish	Teglash vositalari yordamida avtomatik tahlil
7. Sintaktik va semantik tahlil	Matnning grammatik tuzilishi va ma’nosini aniqlash	Sintaksis va semantika asosida tahlil qilish
8. Qo‘lda tekshirish	Avtomatik tahlil natijalarini qo‘lda nazorat qilish	Aniqlik va ishonchlilikni oshirish
9. Indekslash va qidiruv tizimi	Matnlarni indekslash va tizimlashtirish orqali tezkor qidiruv imkoniyatini yaratish	Foydalanuvchi uchun qulay kirish interfeysini ta’minlash
10. Visualizatsiya va natija tahlili	Natijalarni grafiklar va diagrammalar shaklida vizualizatsiya qilish hamda tahlil qilish	Til evolyutsiyasini osonroq anglash va taqdim etish



Xulosa

Ushbu maqolada mustaqillik davri nashrlari asosida diaxron korpus yaratishning asosiy bosqichlari, ya’ni materiallarni tanlash, ma’lumotlarni to‘plash, raqamlashtirish, matn tozalash, normalizatsiya, avtomatik belgilash, sintaktik-semantik tahlil, qo‘lda tekshirish, indekslash va natijalarni vizualizatsiya qilish bosqichlari batafsil ko‘rib chiqildi. Tadqiqot davomida har bir bosqichning o‘ziga xos vazifalari va qo‘llaniladigan metodologiyalar aniq aks etdi. Korpus yaratish jarayonining har bir bosqichi tilning tarixiy rivojlanishini yanada chuqurroq o‘rganishga xizmat qilgani ta’kidlandi. Shu bilan birga, tadqiqot natijalari asosida korpusning sifatini oshirish va yanada boyitish istiqbollari ko‘rib chiqildi. Kelgusida korpusga audio-video materiallarni qo‘shish imkoniyatlari ham o‘rganilishi, shuningdek, ko‘p modalli ma’lumotlar asosida yangi tahlil usullarini joriy etish tavsiya etiladi. Ushbu yondashuv tilshunoslik va NLP sohalarida yangi ilmiy qarashlar va modellashtirish metodlarini ishlab chiqishga asos bo‘lib xizmat qiladi.

Diaxron modellashtirish bo‘yicha kelajakdagi tadqiqotlar esa til evolyutsiyasining murakkab jarayonlarini yanada aniqroq ko‘rsatish imkonini beradi. Tadqiqot davomida aniqlangan cheklovlar va xatoliklar asosida metodologik takomillashtirishga ehtiyoj sezildi. Natijada, yaratilgan korpus nafaqat tarixiy til tahlilida, balki hozirgi va kelajak til modellarini ishlab chiqishda muhim resurs sifatida namoyon bo‘ldi. Ushbu ilmiy izlanish kelgusidagi tadqiqotlar uchun yangi yo‘nalishlar ochib, til evolyutsiyasini yanada keng qamrovda o‘rganishga zamin yaratadi.

Foydalanilgan adabiyotlar:

1. McEnery T. and Baker P. (2016). Corpus Linguistics and 19th-Century Historical Texts. Edinburgh University Press.
2. Davies M. (2020). The Corpus of Historical American English (COHA): 200 years of historical data for linguistic analysis. Brigham Young University Press.
3. Karimov X. va boshq. (2021). O‘zbek Milliy Korpusi va uning zamonaviy til tadqiqotlaridagi o‘rni. O‘zbek tilshunosligi ilmiy tadqiqotlari jurnali, 8(3), 45–60.
4. Aissen J. (2003). Differential Object Marking: Iconicity vs. Economy. Natural Language and Linguistic Theory, 21(3), 435–483.
5. Ataboyev N.B. Mediamatnlar diaxron korpusida til rivojining empirik tahlil tamoyillari. Monografiya. Buxoro. (2024).
6. Elov B., Hamroyeva Sh., Alayev R., Xusainova Z., Yodgorov U. O‘zbek tili korpusi matnlarini qayta ishlash usullari. (2023).
7. Meyer C.F. English Corpus Linguistics: An Introduction. Cambridge University Press. (2002).
8. Xusainova Z., Yangibayeva S., Diaxron korpus va uning arxitekturasi. (2025).