

SCIENCE
PROBLEMS.UZ

ISSN 2181-1342

Actual problems of social and humanitarian sciences
Актуальные проблемы социальных и гуманитарных наук

Ijtimoiy-gumanitar fanlarning dolzarb muammolari

11-maxsus
son (5-jild)

2025

SCIENCEPROBLEMS.UZ

**IJTIMOIIY-GUMANITAR FANLARNING
DOLZARB MUAMMOLARI**

№ S/11 (5) – 2025

**АКТУАЛЬНЫЕ ПРОБЛЕМЫ СОЦИАЛЬНО-
ГУМАНИТАРНЫХ НАУК**

ACTUAL PROBLEMS OF HUMANITIES AND SOCIAL SCIENCES

TOSHKENT-2025

<i>Eshmurodov Nabillo</i> GLOBALLASHUV JARAYONIDA SHAXS MA'NAVİYATINING TRANSFORMATSIYASI VA BEGONALASHUV MUAMMOSI	197-201
<i>Yokubov Suxrob Firdavsovi</i> TASAVVUF TA'LIMOTINING G'OYAVIY MANBALARI	202-206
<i>Najimidinova Dilnoza Gulomjonovna</i> GLOBALLASHUV SHAROITIDA O'ZBEK MILLIY OILASINING IJTIMOIY-FALSAFIY MASALALARI	207-210
<i>Otokulov Abror Sobirkulovich</i> JAMIYATDA MA'NAVIY-AXLOQIYLIKNI YUKSALTIRISHDA SHAXS MAS'ULIYATINI OSHIRISH OMILI	211-217

10.00.00 – FILOLOGIYA FANLARI

<i>Maxamadaliyeva Zuxra Baxrom qizi, Adilova Nasiba Raximboy qizi</i> INGLIZ VA O'ZBEK TILLARIDA MADANIY METAFORALARNING MOHIYATI	218-222
<i>Xusainova Zilola Yuldashevna, Yangibayeva Surayyo Gulumboyevna</i> DIAXRON KORPUS ALGORITMLARI	223-228
<i>Ruzibayeva Aziza Kahramanovna</i> POLITICAL EXILE AND PERSONAL IDENTITY IN LORD BYRON'S LIFE AND WORK	229-233
<i>Karimova Ozodaxon Avazxon qizi</i> A LINGUO-PRAGMATIC STUDY OF INFORMAL SPEECH IN ENGLISH AND UZBEK LANGUAGES: COMPARATIVE ANALYSIS OF COMMUNICATIVE STRATEGIES, SOCIOCULTURAL CONTEXTS, AND PRAGMATIC FUNCTIONS	234-240
<i>Айтмуратова Турсынай Маратовна</i> КОНЦЕПТ «ТРУД» В ЛИНГВОКУЛЬТУРОЛОГИЧЕСКОМ АСПЕКТЕ	241-244
<i>Fayzullayeva Zebuniso Abdukarimovna</i> KONYUNKTIVNING SEMANTIK-PRAGMATIK XUSUSIYATLARI	245-248
<i>Sodikova Vasilakhon, Gavharoy Isroiljon qizi</i> THE ISSUES OF INFORMATION SECURITY AND RELIABILITY IN MODERN JOURNALISM	249-258
<i>Холикова Замира</i> СРАВНИТЕЛЬНО-СОПОСТАВИТЕЛЬНЫЙ АНАЛИЗ ФРАЗЕОЛОГИЗМОВ С КОМПОНЕНТОМ-ЗООНИМОМ «СОБАКА» В РУССКОМ И УЗБЕКСКОМ ЯЗЫКАХ	259-272

12.00.00 – YURIDIK FANLAR

<i>Турсунова Маликахон Улугбековна</i> СООТНОШЕНИЕ ДИПЛОМАТИЧЕСКИХ РАНГОВ С КВАЛИФИКАЦИОННЫМИ КЛАССАМИ ДОЛЖНОСТЕЙ ГОСУДАРСТВЕННОЙ СЛУЖБЫ РЕСПУБЛИКИ УЗБЕКИСТАН: НЕКОТОРЫЕ ВОПРОСЫ ЗНАЧЕНИЯ КЛАССОВ И РАНГОВ В ДИПЛОМАТИЧЕСКОЙ ДЕЯТЕЛЬНОСТИ	273-286
<i>Karaketova Dilnoza Yuldashevna</i> JINOYAT SUBYEKTINI TAVSIFLOVCHI BELGI SIFATIDA RETSIDIV	287-294
<i>Мируктамова Феруза Лутфуллаевна</i> ПРАКТИКА ЗАРУБЕЖНЫХ ГОСУДАРСТВ ПО ЗАЩИТЕ ПРАВ И БЕЗОПАСНОСТИ ДЕТЕЙ ОТ ИНФОРМАЦИОННЫХ УГРОЗ	295-300

Received: 1 December 2025
Accepted: 15 December 2025
Published: 30 December 2025

Article / Original Paper

DIACHRONIC CORPUS ALGORITHMS

Khusainova Zilola

Doctor of Philosophy (PhD) in Philology

Alisher Navoiy Tashkent State University of Uzbek Language and Literature

Yangibayeva Surayyo

Alisher Navoiy Tashkent State University of Uzbek Language and Literature

2st-year Master's Student in Computer Linguistics

Abstract. This article describes the creation of a diachronic corpus of Uzbek fiction published between 1991 and 2021, along with its processing algorithms. Within the framework of corpus linguistics, the processes of text collection, preprocessing, sentence segmentation, metadata formation, and verification were scientifically implemented. As a result, a clean and standardized corpus comprising 116 works was obtained. Using the corpus algorithms, it is possible to analyze the temporal changes of linguistic units, perform statistical analysis by genre and demographic characteristics, and build n-gram models. This study serves as a reliable resource for diachronic research on the Uzbek language and practical investigations in the field of NLP.

Keywords: diachronic corpus, pipeline, sentence segmentation, preprocessing, metadata, NLP.

DIAXRON KORPUS ALGORITMLARI

Xusainova Zilola Yuldashevna

Filologiya fanlari falsafa doktori (PhD)

Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti

E-mail: xusainovazilola@navoiy-uni.uz

<http://www.orcid.org/0000-0003-4357-7515>

Yangibayeva Surayyo Gulumboyevna

Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti

Kompyuter lingvistikasi mutaxassisligi II bosqich magistranti

E-mail: surayyoyangibayeva4@gmail.com

Annotatsiya. Ushbu maqolada 1991–2021-yillar oralig'ida chop etilgan o'zbek badiiy asarlari asosida diaxronik korpus yaratish va uning algoritmlari tavsiflanadi. Korpus lingvistikasi doirasida matnlarni yig'ish, dastlabki ishlov berish, gap segmentatsiyasi, metama'lumotlarni shakllantirish va verifikatsiya jarayonlari ilmiy asosda amalga oshirildi. Natijada 116 ta asardan tashkil topgan, toza va standartlashtirilgan korpus hosil qilindi. Mazkur korpus algoritmlari yordamida lingvistik birliklarning davriy o'zgarishlari, janr va demografik xususiyatlar bo'yicha statistik tahlil va n-gram modellarini yaratish imkoniyati ta'minlandi. Ushbu ilmiy ish o'zbek tili diaxronik tadqiqotlari va NLP sohasidagi amaliy izlanishlar uchun ishonchli manba sifatida xizmat qiladi.

Kalit so'zlar: diaxron korpus, pipeline, gap segmentatsiyasi, dastlabki ishlov berish, metama'lumot, NLP.

DOI: <https://doi.org/10.47390/SPR1342V5SI11Y2025N34>

Kirish. Korpus tadqiqotlari hozirgi kunda kompyuter lingvistikasi ichidagi eng dolzarb yo'nalishlardan biri sifatida e'tirof etiladi. Korpus lingvistikasi til birliklarini elektron formatga keltirish va ularni maxsus tuzilgan korpuslarda jamlash orqali tilning o'rganilishiga oid ilmiy,

nazariy hamda amaliy izlanishlarni tizimli va samarali amalga oshirish imkonini yaratadi. Mazkur maqsad asosida ushbu sohada bir nechta metodologik yondashuvlar shakllangan.

Birinchidan, korpus lingvistikasi tilning hozirgi holatini aks ettiruvchi zamonaviy manbalar asosida korpuslar yaratish bilan shug'ullanadi. Mazkur turdagi korpuslar tilning bugungi holati haqida aniq xulosalar chiqarishga hamda uning kelajakdagi rivoji borasida ilmiy taxminlar bildirishga imkon beradi.

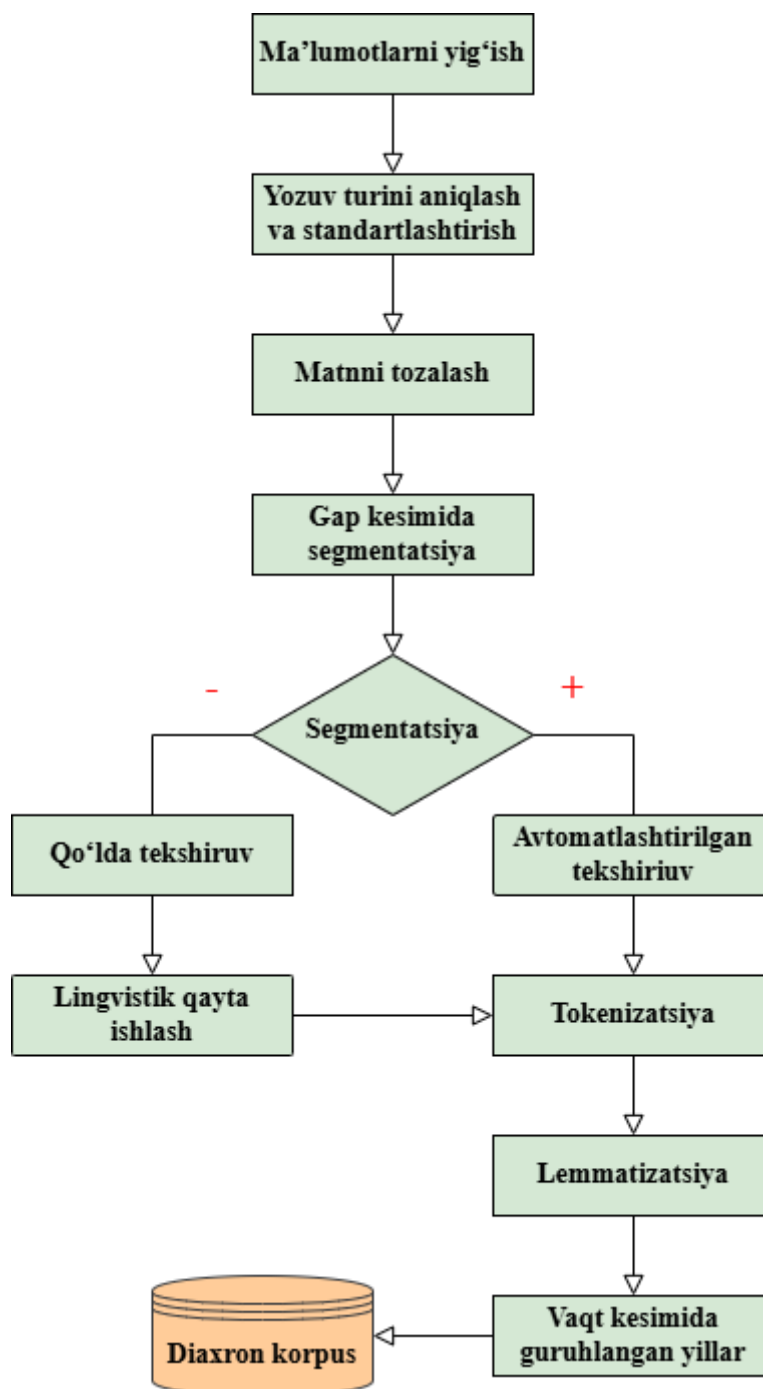
Ikkinchidan, tilning ma'lum bir tarixiy davriga oid manbalardan tashkil topgan korpuslar shakllantiriladi. Bunday korpuslar tahlili orqali tilning rivojlanish bosqichlari aniqlanadi va uning hozirgi maqomini ilmiy asosda belgilash mumkin bo'ladi [1; 67–68-b.].

Diaxron korpuslar aynan shu ikkinchi yondashuvning mahsuli bo'lib, tilning vaqt davomida o'zgarishini kuzatish, hodisalarni tarixiy izchillikda tahlil qilish imkonini beradi. Shu bois diaxronik tadqiqotlar korpus lingvistikasi uchun alohida ahamiyatga ega.

Diaxron korpuslar, odatda, **ma'lum vaqt oralig'idagi** matnlar asosida tuzilib, **til birliklarining tarixiy taraqqiyotini** o'rganishga xizmat qiladi. 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 jihatlar chuqur o'rganilishi mumkin. Mana shu dolzarbligini inobatga olgan holda mustaqillik davri nashrlariga asoslangan (1991–2021) o'zbek tili diaxronik korpusining lingvistik ta'minoti ishlab chiqilmoqda. Korpus lingvistikasi bo'yicha mustaqillik davri nashrlarini qamrab olgan diaxronik korpuslar hali to'liq shakllanmagan, bu esa lingvistik tahlil va texnologik ishlanmalarni cheklaydi. Shu sababli, ushbu maqolada 1991–2021 yillar oralig'ida chop etilgan badiiy asarlar asosida diaxronik korpus yaratish va uning algoritmlarini ishlab chiqish jarayoni tavsiflanadi.

Asosiy qism. Diaxron korpuslarni shakllantirish jarayonida matnlarni yig'ish, tozalash va metama'lumotlarni yaratish bilan bir qatorda, korpusga avtomatik ishlov beruvchi hisoblash jarayonlari ham muhim ahamiyatga ega. Ilmiy izlanishlarda bu jarayonlar korpus algoritmlari sifatida izohlanadi. Korpus algoritmlari matnni tokenlash, normalizatsiya qilish, morfologik va sintaktik belgilash, lemmatizatsiya, n-gram va kollokotsiyalarni statistik aniqlash kabi kompyuterlingvistik protseduralardan iborat. Ushbu algoritmlar korpusni faqat matnlar to'plami sifatida emas, balki lingvistik tadqiqot uchun strukturalangan, indekslangan va avtomatik tahlilga tayyor bo'lgan elektron resurs sifatida shakllantirishga xizmat qiladi. Diaxron korpuslar misolida esa ushbu algoritmlar vaqt omiliga tayanib leksik birliklarning davriy o'zgarishlarini kuzatish, ularni turli davrlar kesimida taqqoslash hamda ularning statistik dinamikasini aniqlash imkonini yaratadi.

Mustaqillik davri nashrlariga asoslangan diaxron korpusni shakllantirish jarayonida qo'llanilgan algoritmik yechimlar bir necha bosqichlar asosida amalga oshirilgan. Agar ushbu jarayonlarni algoritm tarzida tasniflaydigan bo'lsak, mustaqillik davri nashrlariga asoslangan diaxron korpusni yaratishda quyidagi ketma-ketlikka duch kelamiz (1-rasm).



1-rasm. Mustaqillik davri nashrlariga asoslangan diaxron korpus uchun taklif etilgan algoritim

NLPda matnni qayta ishlash jarayoni pipeline (konveyer) deb ataladi. Har qanday tabiiy tilni qayta ishlash modelini yaratishda majburiy ravishda amalga oshiriladigan bosqichlar ketma-ketligi aynan shu pipeline hisoblanadi. Matnni qayta ishlash konveyerini loyihalashtirish va amalga oshirish NLP loyihasining asosiy boshlang'ich bosqichi sifatida qabul qilinadi [2; 183–182-b.].

Ushbu bosqichlar ketma-ketligi korpus qurilishida ham rasmiy algoritmik pipeline sifatida qaraladi va quyidagi asosiy bloklardan iborat:

- ma'lumotlarni yig'ish;
- oldindan qayta ishlash (pre-processing);
- gaplarga bo'lish (segmentatsiya);

- metama'lumotlar hosil qilish;
- tekshirish va verifikatsiya.

Har bir bosqich kompyuter lingvistikasining xalqaro standartlari asosida amalga oshiriladi hamda korpusning izchilligi, aniqligi va lingvistik ishonchliligini ta'minlashga yo'naltirilgan. Natijada, korpus yaratishda qo'llanilgan ushbu algoritmik yondashuv nafaqat ma'lumotlarni texnik jihatdan tartibga solib beradi, balki keyingi diaxronik tahlillar uchun statistik ishonchlilik va to'g'rilikni ham kafolatlaydi.

1. Ma'lumotlarni tayyorlash va pre-processing algoritmi

Diaxron korpusni shakllantirish jarayoni asosan o'nta bosqichlardan iborat bo'lib, ularning dastlabkisi ma'lumotlarni yig'ish, standartlashtirish va tahlilga tayyor holatga keltirish jarayonlarini o'z ichiga oladi [4; 165–166-b.]. Ushbu bosqich korpusning sifat ko'rsatkichlarini belgilovchi eng muhim jarayonlardan biri hisoblanadi.

Korpusning asosiy ma'lumot manbasi sifatida 1991–2021-yillar oralig'ida chop etilgan **286 ta o'zbek badiiy asarlari** (faqat nasriy asarlar) tanlab olindi. Ushbu asarlarning katta qismi kirill yozuvida bo'lgani sababli, ularni yagona grafik shaklga keltirish zarurati tug'ildi. Shu maqsadda **matnlarni tayyorlash va dastlabki qayta ishlash (pre-processing)** algoritmi ishlab chiqildi va amaliyotga joriy etildi.

1.1. Kiril → Lotin konvertatsiyasi

Ma'lumotlarni tahlil qilishda yozuv birligi muhim omil bo'lgani sababli, barcha elektron shakldagi matnlar **UzCorpora** korpusi yordamida lotin yozuviga avtomatik tarzda konvertatsiya qilindi. Bu jarayon natijasida yozuvdagi grafik tafovutlar bartaraf etilib, korpusdagi barcha matnlar yagona ortografik standart asosida qayta ishlashga tayyorlandi.

1.2. Keraksiz va yaroqsiz materialni chiqarish

Tozalash bosqichida ma'lumotlarning texnik yaroqliligi va tahlilga mosligi tekshirildi.

- **140 ta asar** faqat rasm ko'rinishida (PDF formatida) mavjud bo'lib, ularni optik belgilarni aniqlash texnologiyalari (OCR) orqali elektron matnga aylantirish **sifat jihatidan samarasiz** natijalar bergani uchun korpusdan chiqarib tashlandi.
- **146 ta elektron matnli** asardan **30 tasi PDF → TXT konvertatsiya** jarayonida **strukturaviy buzilishlar, belgilar yo'qolishi va kodlash xatolari** tufayli yaroqsiz deb topildi.

Ushbu bosqichda maqsad – korpusda faqat lingvistik jihatdan toza, texnik jihatdan to'liq va semantik yaxlit matnlarni saqlab qolishdan iborat bo'ldi.

1.3. Toza korpusni shakllantirish

Natijada **116 ta toza, yuqori sifatli elektron matn** tanlab olindi va korpusning keyingi bosqichlari – **segmentatsiya, tokenizatsiya, morfologik tahlil va algoritmik annotatsiya** uchun tayyor holatga keltirildi. Har bir matn alohida verifikatsiyadan o'tkazilib, sintaktik yaxlitlik, leksik moslik va format birligi jihatidan nazorat qilindi.

Mazkur pre-processing bosqichi natijasida korpus tarkibi grafik, texnik va lingvistik darajada standartlashtirildi. Shu orqali diaxronik korpusda amalga oshiriladigan keyingi bosqichlar – ya'ni **vaqt oralig'idagi til o'zgarishlarini aniqlash, n-gram modellari qurish, morfologik teglash va statistik tahlil** jarayonlari uchun ishonchli asos yaratildi.

2. Segmentatsiya algoritmi

Diaxronik korpusni shakllantirish jarayonida matnlar gaplar ketma-ketligida strukturaviy jihatdan tashkil etildi. Gap chegaralarini aniqlash mezonlari sifatida quyidagi punktuatsiya belgilaridan foydalanildi:

- nuqta (.),
- so‘roq belgisi (?),
- undov belgisi (!)
- ko‘p nuqta (ellipsis) (...).

Matnlarni dastlabki qayta ishlash (pre-processing) bosqichida ushbu belgilar **Notepad++** muharririda muntazam ifodalar (**regular expressions, regex**) yordamida avtomatik tarzda aniqlanib, har bir gap alohida satrga ajratildi. Shu jarayonda ayrim badiiy matnlarda gaplarning ikki yoki undan ortiq satrda davom etib qolgan holatlari qayd etildi. Mazkur muammolarni bartaraf etish maqsadida **Python dasturlash tilida** qo‘shimcha **post-processing** bosqichi amalga oshirildi. Ushbu bosqichda satrlar mantiqiy tahlil asosida qayta birlashtirilib, gapning integral tuzilmasi saqlab qolindi. Natijada har bir gap mustaqil sintaktik birlik sifatida ajratilib, keyingi lingvistik tahlil uchun tayyor holatga keltirilgan to‘rt yuz mingdan ortiq lingvistik baza shakllantirildi.

3. Metama‘lumot algoritmi

Diaxronik korpus tarkibidagi matnlar bo‘yicha metama‘lumotlar bazasi yaratilishi korpusning ichki tuzilishini tavsiflovchi muhim bosqichlardan biridir. Ushbu bosqichda har bir asar uchun muallif, nashr yili, janr, hajm (so‘z soni), va matn manbasi kabi atributlar aniqlanib, yagona formatda metama‘lumotlar jadvaliga kiritildi.

Tadqiqot doirasida korpus 70 nafar adib qalamiga mansub jami 116 ta asardan tashkil topgan bo‘lib, ulardan **13 nafari ayol, 57 nafari erkak yozuvchilardir**. Mazkur asarlar bir nechta janrlar bo‘yicha taqsimlandi (1-jadval).

1-jadval. Korpusdagi asarlarning janrlar bo‘yicha taqsimoti

<i>roman</i>	<i>qissa</i>	<i>hikoya</i>	<i>esse</i>	<i>novella</i>	<i>hajviya</i>	<i>publisistik miniatura</i>
29 ta	49 ta	30 ta	4 ta	2 ta	1 ta	1 ta

Metama‘lumotlar bazasining yaratilishi keyingi tahliliy jarayonlar uchun asos bo‘lib xizmat qiladi. Xususan, mazkur ma‘lumotlar statistik tahlil, janrlararo taqqoslash, hamda n-gram modellarini hisoblashda qo‘llanildi. Metama‘lumotlar yordamida korpusning **demografik (muallif jinsi), janr jihatidan (badiiy tur va shakl), yozilgan yili jihatidan (nechanchi yilda yozilgan) va xronologik (nashr yillari)** tarkibi aniqlanib, bu ma‘lumotlar asosida diaxron o‘zgarishlarni modellashtirish imkoniyati yaratildi.

Shuningdek, metama‘lumotlar strukturasi **Python** muhiti orqali avtomatik tarzda shakllantirilib, **CSV formatida saqlash** imkonini beruvchi algoritm yordamida optimallashtirildi. Natijada, har bir matn haqidagi tavsifiy ma‘lumotlar korpus tizimiga integratsiyalashgan holda, keyingi semantik va leksik tahlil bosqichlari uchun moslashuvchan va qayta ishlanadigan model ko‘rinishiga keltirildi.

4. Verifikatsiya

Korpusni yakuniy shaklga keltirish jarayonidan so‘ng, uning sifatini ta‘minlash maqsadida **qo‘lda tekshirish (manual verification)** bosqichi amalga oshirildi. Ushbu bosqichda quyidagi ishlar bajarildi:

1. **Segmentatsiya natijalari tekshiruvi** – har bir gapning to‘liq va mantiqiy ajratilganligi tasdiqlandi, satrlararo uzilishlar aniqlanib, Python orqali amalga oshirilgan post-processing natijalari bilan solishtirildi.
2. **Pre-processing jarayonining tekshiruvi** – matnlarni tozalash va punktuatsiya belgilarini normalizatsiya qilish bosqichi natijalari ko‘zdan kechirildi, xatoliklar va noto‘g‘ri ajratilgan belgilar tuzatildi.
3. **Metama’lumotlarning mosligi** – har bir asar uchun muallif, janr, nashr yili va boshqa atributlar to‘g‘ri kiritilganligi tekshirildi, nomuvofiqliklar aniqlanib, tuzatildi.

Natijada, barcha segmentatsiya va pre-processing jarayonlarining natijalari tasdiqlanib, **to‘liq, integral va xatolardan holi diaxron korpus** shakllantirildi. Ushbu verifikatsiya bosqichi korpusning keyingi lingvistik tahlil, n-gram hisoblash va boshqa algoritmik ishlovlar uchun ishonchli asos yaratadi.

Xulosa. Diaxron korpus yaratish jarayoni matnlarni yig‘ish, pre-processing, gap segmentatsiyasi, metama’lumotlarni shakllantirish va verifikatsiya bosqichlarini o‘z ichiga olgan kompleks algoritmik pipeline orqali amalga oshirildi. Natijada, 70 nafar muallifning 116 ta badiiy asarini o‘z ichiga olgan, toza va standartlashtirilgan korpus hosil qilindi. Ushbu korpus lingvistik birliklarning davriy o‘zgarishlarini kuzatish, janrlararo va demografik taqqoslashlar o‘tkazish, shuningdek, n-gram modellarini yaratish imkonini beradi. Qo‘llanilgan algoritmlar matnlarni sistematik tarzda ishlov berish va lingvistik tadqiqotlar uchun tayyor holatga keltirish imkonini berib, diaxronik tahlil jarayonlarining ishonchliligini ta’minlaydi. Shu bilan birga, yaratilgan diaxron korpus o‘zbek tilining mustaqillik davridagi badiiy matnlarini ilmiy tahlil qilish, tilning tarixiy va stilistik o‘zgarishlarini aniqlash hamda NLP sohasida amaliy tadqiqotlar olib borish uchun ishonchli manba sifatida xizmat qiladi.

Adabiyotlar/Литература/References:

1. Atabayeva N. B. *“Mediamatnlar diaxronik korpusida til rivojining empiric tahlil tamoyillari”* monografiya. Buxoro. 2024. 67-68.
2. Elov B.B., KHamroeva Sh.M., Xusainova Z.Y. *NLP (tabiiy tilga ishlov berish) ning Pipeline konveyeri*. Muhammad al-xorazmiy avlodlari ilmiy-amaliy va axborot tahliliy jurnal. 2023. 181-182.
3. Elov B. B., Amirkulov M. *Uzbeki-English Parallel Corpus Algorithm and Alignment Problem*. Central Asian Studies. 2023. 71-76.
4. Xusainova Z.Y., Yangibayeva S.G. *“Diaxron korpus yaratish bosqichlari”* maqola. Toshkent. 2025. 165-166.
5. Xusainova Z., Yangibayeva S. *Mustaqillik davri nashrlariga asoslangan diaxron korpus yaratishning lingvistik ta’minoti*. International scientific-practical conference: *Contemporary Technologies of Computational Linguistics – CTCL*. 2025. 270-273.
6. Xusainova Z.Y., Yangibayeva S.G. *“Diaxron korpus arxitekturasini”* maqola. Qo‘qon. 2025. 1073-1078.
7. <https://uznatcorpara.uz>
8. <https://ruscorpora.ru/>