QO‘QON DAVLAT PEDAGOGIKA INSTITUTI ILMIY XABARLARI
(2025-yil 2-son)

FILOLOGIYA

PHILOLOGY

DIAXRON KORPUS VA UNING ARXITEKTURASI

Zilola Xusainova Yuldash qizi
ToshDO‘TAU filologiya fanlari falsafa doktori (PhD).
E-pochta: xusainovazilola@navoiy-uni.uz
Yangibayeva Surayyo Gulimboy qizi
ToshDO‘TAU, Kompyuter lingvistikasi mutaxassisligi
I bosqich magistranti.
E-pochta: surayyoyangibayeva4@gmail.com

Annotatsiya. Ushbu maqolada til tarixini o‘rganish uchun maxsus yaratilgan diaxron korpuslar va ularning arxitekturasi atroflicha yoritilgan. Diaxron korpus til o‘zgarishlarini vaqt bo‘ylab tadqiq qilish uchun muhim manba bo‘lib, tilning tabiati, tuzilishi va lingvistik tadqiqotlardagi rolini tahlil qilishga xizmat qiladi. Shuningdek, maqolada diaxron korpusning boshqa turdagi korpuslardan farqli jihatlari, uning ma’lumotlar bazasi, annotatsiya tizimi va texnologik asoslari haqida batafsil ma’lumot berilgan. Qolaversa turli tillarda mavjud bo‘lgan diaxron korpuslar tahlil qilinib, ularning lingvistik tadqiqotlar uchun ahamiyati misollar bilan tushuntiriladi.

Kalit so‘zlar: diaxron korpus, NLP, OCR, interfeys, ma’lumotlar bazasi, annotatsiya, klassifikatsiya, tarixiy tilshunoslik, funksiya.

ДИАХРОНИЧЕСКИЙ КОРПУС И ЕГО АРХИТЕКТУРА

Аннотация. В данной статье подробно освещены диахронные корпуса, специально созданные для изучения истории языка, а также их архитектура. Диахронный корпус является важным ресурсом для исследования языковых изменений во времени, служит для анализа природы языка, его структуры и роли в лингвистических исследованиях. Кроме того, в статье представлена детальная информация о различиях диахронного корпуса от других типов корпусов, его базе данных, системе аннотации и технологических основах. Также рассматриваются диахронные корпуса на разных языках, и на примерах объясняется их значение для лингвистических исследований.

Ключевые слова: диахронный корпус, NLP, OCR, интерфейс, база данных, аннотация, классификация, историческая лингвистика, функция.

DIACHRONIC CORPUS AND ITS ARCHITECTURE

Annotation. This article provides a comprehensive analysis of the concept of a diachronic corpus, which is specifically designed for studying language history, and examines its

architecture in detail. The diachronic corpus serves as a crucial resource for investigating language changes over time, with a focus on its nature, structure, and role in linguistic research. Additionally, the article explores the distinguishing features of the diachronic corpus compared to other types of corpora, detailing its database, annotation system, and technological foundations. Furthermore, existing diachronic corpora in various languages are analyzed, illustrating their significance for linguistic studies with relevant examples. The prospects of developing a diachronic corpus for Uzbek linguistics are also discussed.

Keywords: diachronic corpus, NLP, OCR, interface, database, annotation, classification, historical linguistics.

Kirish

Til – bu insoniyat tarixining ajralmas qismi bo‘lib, uning rivojlanishi va o‘zgarishlari davrlar osha kuzatiladi. Til hodisalarini tadqiq qilish uchun maxsus vositalar talab etiladi. Tabiiy tilni qayta ishlash – bu sun'iy intellektning kichik sohasi bo'lib, u mashinalarga inson tilini tushunish va qayta ishlashga yordam beradi. Bu borada mavjud korpuslarni kengaytirish hamda yangilarini yaratish tilshunoslikni rivojlantirishdagi talablardan biridir [Elov B., Hamroyeva Sh., Alayev R., Xusainova Z., Yodgorov U., 2023]. Diaxron korpuslar ham tilshunoslikning muhim yo‘nalishlaridan biri hisoblanib, tilning vaqt bo‘ylab rivojlanishini tadqiq etish imkonini beradi [Smith, 2004]. Lingvistik tadqiqotlarda til hodisalarining tarixi va dinamikasini o‘rganishda diaxron korpuslar asosiy manba sifatida xizmat qiladi. Shu sababli, diaxron korpuslar ma’lum vaqt oralig‘idagi turli janr va tillardagi matnlar to‘plamidan iborat bo‘lib, lingvistika sohasida quyidagi maqsadlarda ishlatiladi:

1. Tilning tarixiy rivojlanishini o‘rganish.

2. Leksik va grammatik birliklarning o‘zgaruvchanligini kuzatish.

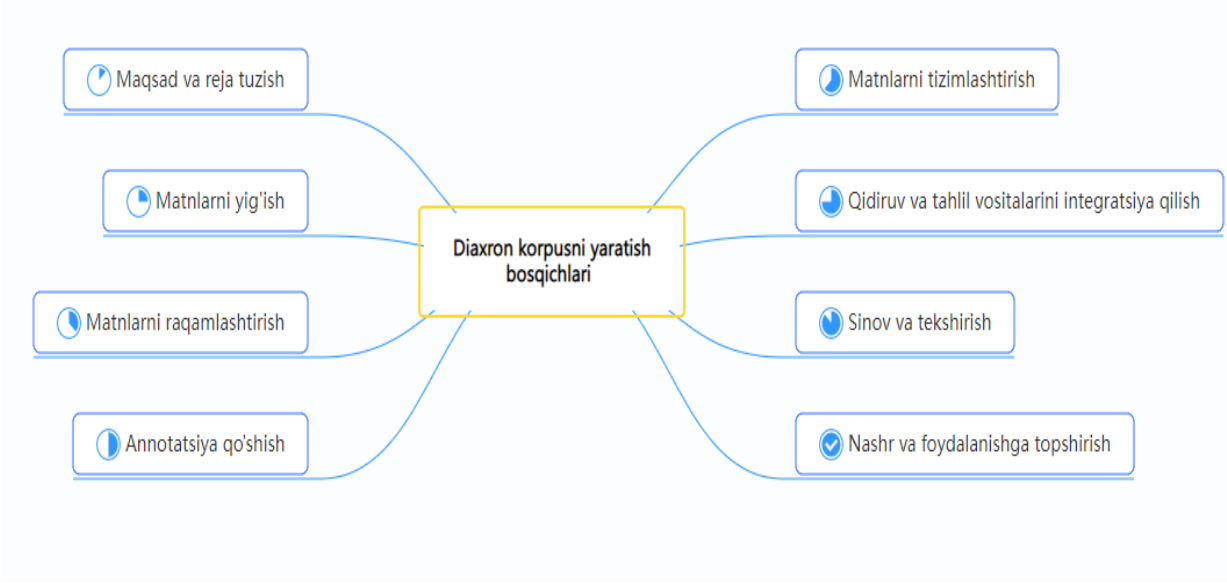
3. Tildagi stilistik va fonetik o‘zgarishlarni tahlil qilish.

Diaxron korpus davr (vaqt) bo‘yicha aniq segmentlarga bo‘linadi va har bir segment o‘ziga xos til xususiyatlarini aks ettiradi. Masalan, eski o‘zbekcha so‘zlar va zamonaviy o‘zbekcha so‘zlar o‘rtasidagi farqlarni tadqiq qilish uchun diaxron korpusdan foydalanish tavsiya qilinadi.

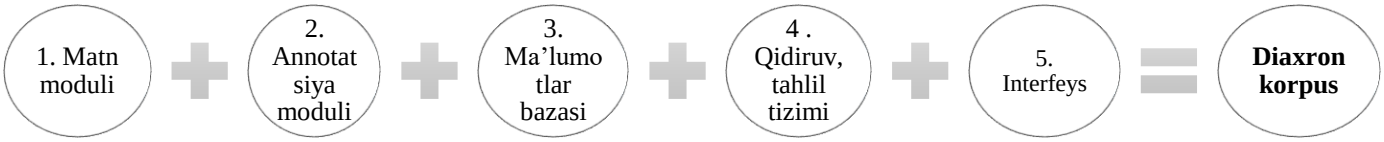
Diaxron korpusning arxitekturasi

Diaxron korpusni yaratish murakkab jarayon bo‘lib, lekin shunday bo‘lsa ham tilshunoslik uchun juda muhimdir. Ushbu korpusni yaratish jarayoni bir qancha bosqichlarni o‘z ichiga oladi. Bu bosqichlar korpusning maqsadi, tarkibi va texnologik asoslarini belgilab, uning samarali va ilmiy tadqiqotlar uchun mos bo‘lishini ta’minlaydi. Quyidagi 1-rasmda ushbu bosqichlar ketma-ketligi keltirilgan.

1-rasm. Diaxron korpusning yaratilish bosqichlari



Ushbu bosqichlarni bajarish orqali ma’lum bir tilning tarixiy bosqichlarini tahlil qiladi va xususiyatlarini aniqlashga xizmat qiladi. Diaxron korpus arxitekturasi ma’lumotlar bazasi, annotatsiya tizimi va foydalanuvchi interfeysi kabi komponentlardan tashkil topadi [Ide & Suderman, 2007]. Yuqoridagi bosqichlarni o‘z ichiga olgan holda korpus arxitekturasi bir necha qismlardan tashkil topgan holda shakllantiriladi (2-rasm).



2-rasm. Diaxron korpus arxitekturasi

1. Matn moduli.

Ushbu modulda korpusdagi **matnlar** ma’lum bir vaqtning **davrlari, janrlar** va **uslublar** bo‘yicha tasniflanadi. Har bir matnga **til, davr, geografik hudud** va boshqa kontekstual ma’lumotlar biriktiriladi.

2. Annotatsiya moduli

Bu modulda matnlarga lingvistik annotatsiyalar (teglar) biriktiriladi. Ya’ni leksik, morfologik va sintaktik teglar asosida matnlar teglanadi. Odatda, teglangan matnni avtomatik tahlil qilish imkonini taqdim etadi. Korpuslarni teglash orqali foydalanuvchi muayyan so‘zni korpus asosida lingvistik tahlil qilishi mumkin [Leech, 2011].

3. Ma’lumotlar bazasi

Korpusda dastlab matnli ma’lumotlarni o‘zida jamlagan baza yaratiladi. Ushbu bazada ma’lumotlar strukturlangan tarzda saqlanadi. Natijada, matnli ma’lumotardan izlash jarayoni optimallashtiriladi. Matnlarning metama’lumotlari (muallif, yil, janr) va lingvistik tahlillari ma’lumotlar bazasida saqlanadi.

4. Qidiruv va tahlil tizimi

Ushbu modulda foydalanuvchilar uchun qulay qidiruv mexanizmlari ishlab chiqiladi. Lingvistik birliklar va ularning o‘zgarishini kuzatish uchun maxsus grafik va statistik vositalar taqdim etiladi. Bu jarayonlar korpusning annatotsiya moduli yaratilgandan keyin ishga tushiriladi.

5. Interfeys

Grafik interfeys foydalanuvchiga kerakli ma’lumotlarni izlash va tahlil qilish imkonini beradi. Interfeysdan foydalanuvchi tilga oid turli ko‘rsatkich (parametr)larni birlashtirgan holda MBdan ma’lumotlarni aniqlab, tadqiqot olib borishi mumkin.

Yuqorida keltirilgan modullar bir-biri bilan bog‘liq bo‘lib, muayyan modeldagi kamchiliklar boshqa modullarga ham uzviy tasir qilishi mumkin. Modullarni o‘zaro birlashtirgan holda diaxron korpusni ma’lum bir til asosida yaratish jarayoni bir qancha texnik yondashuvlarni talab qiladi:

1-jadval. Diaxron korpusning texnik yondashuvlari

Tabiiy tilni qayta ishlash (NLP)	**Mashinali o‘qitish**	**Optik belgilarni tanish (OCR)**
Matnlarni avtomatik qayta ishlash va tahlil qilish uchun ishlatiladi.	Annotatsiya va klassifikatsiya jarayonlarini avtomatlashtirish uchun zarur.	Qog‘oz shaklidagi qadimiy matnlarni raqamli formatga o‘tkazish uchun ishlatiladi.

Ko‘rib turganimizdek diaxron korpuslar o‘ziga xos struktura va annotatsiya tizimiga ega bo‘lib, tahlil va qidiruvni amalga oshirish uchun maxsus dasturiy vositalarni talab qiladi [Biber et al., 1998]. Masalan, Britaniya milliy korpusining diaxron moduli turli davrlarga oid yozma materiallarni o‘z ichiga olgan [Burnard, 2007]. Til tarixini o‘rganish maqsadida yaratilgan diaxron korpuslar murakkab bosqichlardan tarkib topgan holda yaratiladi. Bugungi kunda jahon tajribasida ko‘plab diaxronik korpuslar ishlab chiqilgan bo‘lib, o‘z vazifasiga ko‘ra ma’lum bir maqsad asosida shakllantirilgan (2-jadval).

2-jadval. Diaxron korpuslar

Ingliz tilidagi diaxron korpuslar	
Helsinki Corpus of English Texts (HCET)	750-1700-yillar oralig‘idagi matnlarni o‘z ichiga oladi.
	Ingliz tilining leksik va grammatik o‘zgarishlarini tahlil qilish uchun ishlatiladi.
Corpus of Historical American English (COHA)	Amerika ingliz tilisining 1810-2000-yillar oralig‘idagi rivojlanishini aks ettiruvchi korpus

	Amerika ingliz tilisining semantik va sintaktik o‘zgarishlarini o‘rganish imkonini beradi.
Penn-Helsinki Parsed Corpus of Middle English (PPCME)	O‘rta asr ingliz tiliga oid tahliliy matnlar.
	Sintaktik annotatsiya va davrlarga oid taqqoslash imkoniyati mavjud

Nemis tilidagi diaxron korpuslar

Deutsches Textarchiv (DTA)	XVI asrdan XX asrgacha bo‘lgan matnlarni o‘z ichiga oladi.
	Nemis tilining tarixiy grammatikasi va leksikasini o‘rganish.
Bonner Frühneuhochdeutsch korpus (BFNHC)	XVI asr nemis tiliga oid matnlar bilan ishlash uchun mo‘ljallangan korpus.
	Erta zamonaviy nemis tilining grammatik tizimini o‘rganish uchun ishlatiladi.

Fransuz va ispan tillaridagi diaxron korpuslar

FRANTEXT	Fransuz tilining XIII asrdan XX asrgacha bo‘lgan rivojlanishini o‘z ichiga oluvchi korpus.
	Fransuz tilidagi leksik va grammatik o‘zgarishlarni tahlil qilish.
CORDE (Corpus Diacrónico del Español)	Ispan tilining dastlabki davrlaridan zamonaviy davrgacha.
	Ispan tilining tarixiy rivojlanishini o‘rganish uchun Real Academia Española tomonidan yaratilgan korpus.

Lotin tiliga oid diaxron korpuslar

LatinISE	Lotin tilining grammatik va leksik rivojlanishini tadqiq qilish imkonini beradi.

O‘zbek va boshqa turkiy tillar diaxron korpuslari

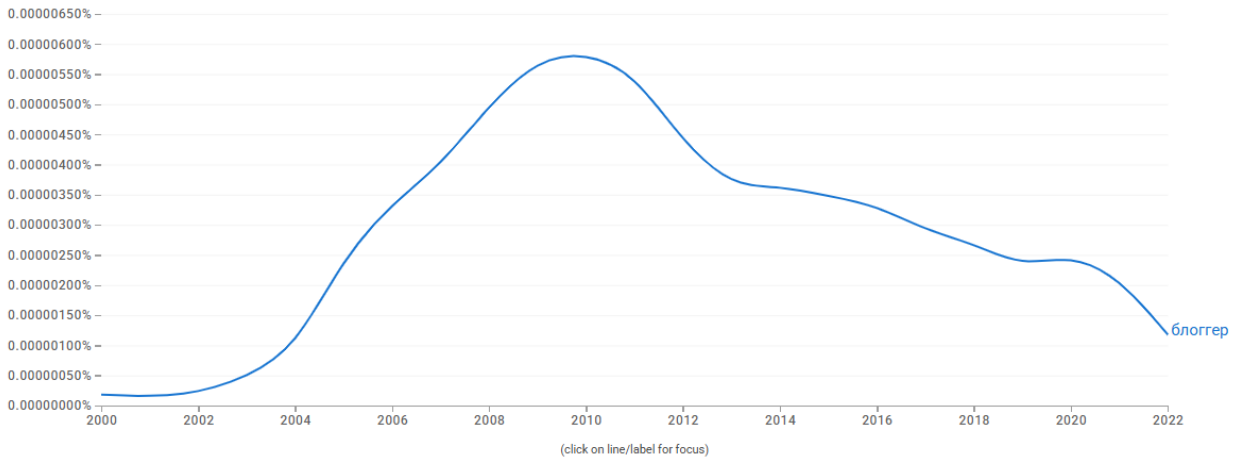
Turkish National Corpus (TNC)	Zamonaviy turk tilining rivojlanishini kuzatish uchun yaratilgan, ammo eski turk tiliga oid ma’lumotlar ham mavjud.
Altay til oilasiga oid korpuslar	Eski uyg‘ur, qadimiy turk, chag‘atoy va boshqa turkiy tillarning tarixiy matnlarini o‘z ichiga oladi.
Orhon-Yenisey yozuvlari korpusi:	Qadimiy turk yozuvlarini tadqiq qilish uchun maxsus yaratilgan.
Chag‘atoy korpusi	Alisher Navoiy davridagi tilni o‘rganish uchun mo‘ljallangan matnlar bazasi.
O‘zbek tiliga oid ishlar	Mediamatnlar diaxronik korpusi. O‘zbek va ingliz gazetalari mediamatnlarini o‘rganish uchun mo‘ljallangan.

Diaxron korpus boshqa turdagi korpuslardan bir qancha farqli funksiyalari bilan farqlanadi. Jumladan, diaxron korpus tilning vaqt bo‘ylab o‘zgarishni inobatga olgan holda

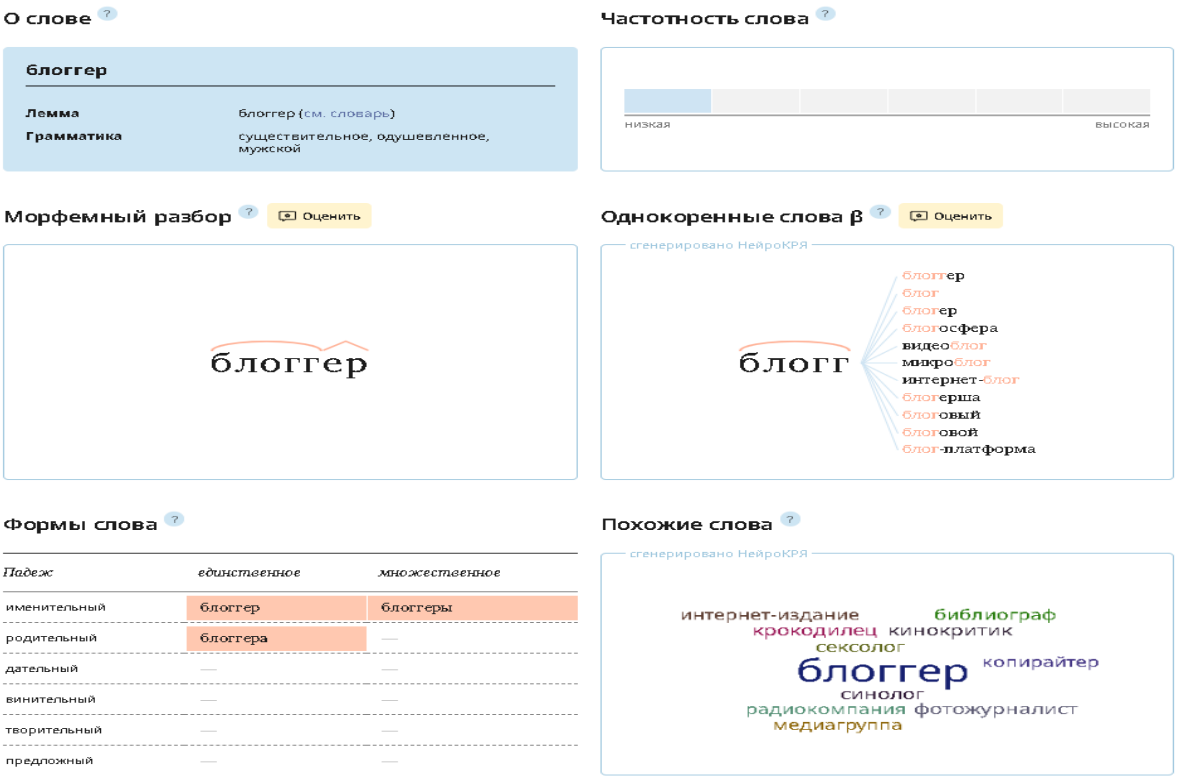
tahlil qilsa, sinxron korpus tilning bir vaqt ichidagi holatini tahlil qiladi. Ingliz korpuslari misolida korpuslarni solishtiradigan bo‘lsak, sinxron korpuslar zamonaviy ingliz tilidagi matnlarni o‘z ichiga olgan bo‘ladi (Masalan, gazetalar, bloglar). *Qadimiy (Old English), o‘rta (Middle English)* va *zamonaviy (Modern English)* kabi ingliz tilidagi diaxron korpuslar turli davrlarga oid matnlarni o‘zida jamlagan. Bundan tashqari, ba’zi diaxron korpuslar o‘z ichida *sinxron, parallel, diaxron, dialektal* kabi bir qancha korpuslarni o‘zida jamlagan. Bu turdagi korpuslarga rus tili milliy korpusini keltirish mumkin [https://ruscorpora]. Ushbu korpus so‘zlarni turli xil jihatdan tahlil qilib bera oladigan korpus hisonlanib, xususan, uning tarkibidagi diaxron korpusda *“blogger”* so‘zining ishlatilish chastotasi vaqt davomida qanday o‘zgarganligini tasvirlab beriladi (3-rasm). Sinxron korpusda *“blogger”* so‘zining hozirgi zamonda morfemik, grammatik jihatdan qaysi kontekstlarda, qanday birliklar bilan bog‘langan holda ishlatilishi, lemmasi ko‘rsatib berilgan (4-rasm). Shu jihatlarga ko‘ra, korpuslar bir qancha funksiyalar bilan farqlanadi (3-jadval).

3-jadval. Diaxron va sinxron korpus funksiyalari

№	Diaxron korpus funksiyalari	Sinxron korpus funksiyalari
	Ma’lum bir tarixiy davr oralig‘ida	*Hozirgi zamonda*
1.	Til o‘zgarishlarini kuzatish	Tarjimada ekvivalentlarni topish
2.	Semantik o‘zgarishlarni tahlil qilish	Kollokatsiyalar (so‘z bog‘lanishlarini) tahlil qilish
3.	Grammatik o‘zgarishlarni kuzatish	Sintaktik va grammatik strukturalarni o‘rganish
4.	Murakkab iboralar va jumlalar evolyutsiyasini kuzatish	So‘z va iboralar chastotasini aniqlash
5.	Dialekt va regional farqlarni o‘rganish	Dialekt va uslub farqlarini tahlil qilish
6.	Yozuv shakllarining o‘zgarishini kuzatish	Terminologiyani o‘rganish va standartlashtirish



3-rasm. “blogger” so‘zining yillar kesimida diaxron korpusda tasvirlanish



4-rasm. “blogger” so‘zining sinxron korpuslarda ifodalanishi

Xulosa

Diaxron korpusni yaratish jarayoni murakkab va ko‘p bosqichli bo‘lib, u lingvistik tadqiqotlar uchun muhim vosita hisoblanadi. Diaxron korpus til o‘zgarishlarini chuqur tahlil qilish va tarixiy jarayonlarni tushunish imkonini beradi. Tadqiqot natijalari shuni ko‘rsatadiki, mukammal arxitekturaga ega diaxron korpuslar ma’lumotlarni samarali jamlash, tahlil qilish va foydalanuvchilarga qulay interfeys orqali taqdim etish imkonini beradi. Ushbu maqolada diaxron korpusning texnik va ilmiy asoslari yoritilib, uning NLP, OCR va mashinali o‘qitish texnologiyalari bilan integratsiyalashuvi muhim ekani keltirildi. Shuningdek, mavjud diaxron korpuslar tajribasi keltirilib, o‘zbek tiliga oid diaxron korpus yaratish zarurati ochib berildi. Tilshunoslik sohasida bu kabi resurslarning mavjudligi til tarixini, grammatik va leksik o‘zgarishlarni keng miqyosda tahlil qilish imkonini beradi. Bugungi kunda diaxron korpus texnologiyalarining rivojlanishi tilshunoslik tadqiqotlarini yangi bosqichga olib chiqishi kutilmoqda. Shuning uchun ham o‘zbek tiliga oid tarixiy matnlarni to‘plash va raqamlashtirish bo‘yicha izlanishlarni davom ettirish muhim ahamiyatga ega.

ADABIYOTLAR

1. Baker, P., & McEnery, T. (2015). Corpus Linguistics and 20th-Century Historical English. Cambridge University Press.

2. Biber, D., Conrad, S., & Reppen, R. (1998). Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press.

3. Elov B., Hamroyeva Sh., Alayev R., Xusainova Z., Yodgorov U. O`zbek tili korpusi matnlarini qayta ishlash usullari. Maqola. Toshkent: 2023.

4. EAGLES Project (1996). Recommendations for Corpus Annotation.

5. https://ruscorpora.

6. Irma Taavitsainen, Andreas H. Jucker & Jukka Tuominen (Eds.), John Benjamins, Amsterdam (2014). Diachronic Corpus Pragmatics.

7. JOST Gippert & Manana Tandashvili (2015). Structuring a diachronic corpus The Georgian National Corpus project.

8. McEnery, T., & Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. Cambridge University Press.

9. Rissanen, M. (2008). Diachronic Corpus Studies: Progress and Prospects. Oxford University Press.

10. Sketch Engine Documentation (2023). [Online Resource].