



Principles of working with text sources and sorting chronology in creating a diachronic corpus of media texts

Nozimjon ATABOYEV¹

Bukhara State University

ARTICLE INFO

Article history:

Received December 2023

Received in revised form

10 December 2023

Accepted 25 January 2024

Available online

25 February 2024

Keywords:

corpus linguistics,
linguistic corpus,
language development,
media text,
newspaper texts,
diachronic corpus.

ABSTRACT

The development of corpus linguistics is closely related to the widespread use of language corpora in finding solutions to linguistic problems. It is known that corpus linguistics has a set of methods focused mainly on lexicography but they can be used by stylistics, syntax and sentence construction, language teaching methodology, and other branches of science. In the article, the problem of collecting and sorting lexical units in the creation of a diachronic corpus of media texts aimed at the study of language development was raised, and detailed solutions have been proposed.

2181-3663/© 2024 in Science LLC.

DOI: <https://doi.org/10.47689/2181-3701-vol2-iss1-pp46-52>

This is an open-access article under the Attribution 4.0 International (CC BY 4.0) license (<https://creativecommons.org/licenses/by/4.0/deed.ru>)

Mediamatnlar diaxronik korpusini yaratishda matniy manbalar bilan ishlash asoslari va saralash xronologiyasi

ANNOTATSIYA

Kalit so'zlar:

Korpus lingvistikasi,
lisoniy korpus,
til rivoji,
mediamatn,
gazeta matnlari,
diaxronik korpus.

Korpus lingvistikasi rivoji til korpuslarining lingvistik muammolarga yechim topishda keng qo'llanilishi bilan chambarchas bog'liqdir. Ma'lumki, korpus lingvistikasi asosan lug'atshunoslik, ya'ni, leksikografikaga oid tadqiqotlarda yo'naltirilgan metodlar jamlanmasiga ega bo'lsada, ularni stilistika, sintaksis va gap qurilishi, til o'qitish metodikasi hamda boshqa fan tarmoqlariga mos ravishda ishlatish mumkin. Maqolada, til rivojini tadqiq qilishga yo'naltirilgan mediamatnlar diaxronik korpusini yaratishda mantiy birliklarni yig'ish va saralash muammosi ko'tarilgan hamda atroflicha yechimlar mulohaza shaklida taklif etilgan.

¹ Doctor of Philosophy (PhD) in Philology, Associate Professor, Dean, Foreign Languages Faculty, Bukhara State University. E-mail: ataboevnozim@gmail.com

Принципы работы с текстовыми источниками и разборки хронологии при создании диахронического корпуса медиатекстов

АННОТАЦИЯ

Ключевые слова:

корпусная лингвистика,
лингвистический корпус,
развитие языка,
медиатекст,
газетные тексты,
диахронический корпус.

Развитие корпусной лингвистики тесно связано с широким использованием языковых корпусов при решении языковых задач. Известно, что корпусная лингвистика применяет комплекс методов, в основном ориентированных на лексикографию, но также актуальных в стилистике, синтаксисе, построении предложений, методике преподавания языка и в других научных областях. В статье рассматривается проблема сбора и сортировки лексических единиц при формировании диахронического корпуса медиатекстов, цель которого — изучение развития языка, и предлагаются подробные решения указанной проблемы.

O'zbek tilshunosligi, xususan, leksikografiyada ham korpus lingvistikasi metodlaridan foydalanilganligini kuzatish mumkin, biroq, ular aynan korpus tahlillariga asoslanganligi qayd etilmaydi. Masalan, *O'zbek xalq dostonlarining chastotali izohli frazeologik lug'atini* yaratish borasida qilingan ishlar sirasiga matniy birliklarni yig'ilganligi, ularni kompyuter xotirasiga kiritib, alfavit-chastotali lug'atlar olinganligi aytiladi: "... lug'atni tuzish jarayonida dastlab «Alpomish», «Ravshan», «Rustam» dostonlarining matnlari kompyuter xotirasiga kiritildi. Kompyuter xotirasida to'plangan ana shu matnlar asosida kompyuter yordamida 28499 so'z shaklidan iborat, uchta chastotali va uchta alfavit-chastotali lug'atlar olindi: «Alpomish» dostoni umumiy matn hajmi – 14029, ya'ni – 49,2 % ni; «Ravshan» dostoni – 7059, ya'ni 24,7 %ni, «Rustam» dostoni 7411 so'z shakli 26,2% ini tashkil qiladi. Matnlarni qayta tekshirish jarayonida mazkur matnlarda jami 1041 FB qo'llanganligi ma'lum bo'ldi, ular xalq dostonlari matnida 1641 marta qaytarilgani aniqlandi". Ko'rinib turganidek, biror o'rinda qaysi elektron yoki kompyuter dasturidan qo'llanilganligi eslatib o'tilmaydi. Vaholanki, bu jarayon korpus lingvistikasining asosiy maqsadi bo'lgan *korpuslar yaratish* orqali lisoniy bazaga oid statistik axborotni tez fursatda olish imkonini yaratishga xos birlamchi vazifalar turkumiga kiradi.

Bizningcha, yuqoridagi holatni inobatga olganda, o'zbek lingvistika ilmi korpusga asoslangan nazariy tadqiqotlarning amaliy jabhalarda qo'llashga oid yangi yondashuvlarni talab qilmoqda. Demak, yaratilishi kerak bo'lgan korpuslar qatorida turuvchi *O'zbek va ingliz gazetalari mediamatnlari diaxronik korpusini yaratish* aynan shu turdagi natijaviy ilmiy izlanishlar sirasiga kiradi.

J. Sinclair va F. Douglas kabi olimlarning fikricha, korpusga kiritiladigan matnlarning ba'zilar uzun bo'lgani tufayli, ularning to'liq hajmini jalb etmasdan, balki matn bo'laklaridan foydalanish ba'zan matnning faqatgina bir turi ustun bo'lmasligi uchun korpusda qo'llaniladi. Biroq, to'liq matnlardan foydalanishning ustunligiga doir ko'plab yondashuvlar mavjud deya xulosa qilishadi. Bizning nazarimizda olimlarning fikrlari o'rinli va matnlar hajman teng bo'lishi amaliyotda kamdan-kam uchragani bois, korpusga kiritiladigan matnning to'liq shaklini kiritish uning mazmunan va mantiqan tog'ri tahlil qilinishiga asos bo'ladi.

M. Dosenna, S. Mackay, E. Moreton, Morton, J. Pennebaker, M. Mehl va K. Niederhoffer, L. Sowada, L. Anthony va boshqa bir qator olimlar korpus bazasiga kiritiladigan matniy birliklarining xususiyatlari to'g'rida to'xtalib, ularni saralashga, maqsadli ravishda jamlashga alohida e'tibor qaratilishini ta'kidlashgan.

Bugungi kunda yaratilayotgan ko'plab korpuslar, xususan, o'zbek tilining zamonaviy korpuslari asosan onlayn, ya'ni internet materiallaridan shakllantirilayotganligini kuzatish mumkin. Bizningcha, korpus yaratish maqsad qilinganda, undan kutiladigan natijani inobatga olgan holda, imkon doirasida manbaning ishonchliligi, autentik ekanligi hamda eng muhimi uning ertangi kunga saqlanishiga e'tibor qaratilishi joiz. Fikrimizning isboti sifatida N. Koteykoning quyidagi fikrini keltirish mumkin: "onlayn ma'lumotlarning mavjudligi yanada xilma-xil korpuslarni yaratishga va ularda topilgan lingvistik birliklar haqida ko'proq umumlashtirilgan xulosalar berishga imkon yaratadi. Biroq, internet materiallariga asoslangan ma'lumotlar to'plash va hujjatlashtirish nuqtai nazaridan ham muayyan qiyinchiliklarni keltirib chiqaradi. Internetdan olingan materiallar bilan bog'liq ikkita keng tarqalgan muammo mavjud. Bu manba haqidagi qo'shimcha ma'lumotlarning yetishmasligi yoki to'liq emasligi va ma'lumotlarning beqaror tabiati. Ya'ni, bugun tahlil qilish mumkin bo'lgan matnlar ertaga ogohlantirilmasdan olib tashlanishi mumkin. Shu sababli, chop etilgan, onlayn ma'lumotlar bazasida kiritilgan va/yoki saqlangan gazeta matnlari va siyosiy chiqishlar odatda boshqa onlayn ma'lumotlar turlariga qaraganda korpusni to'plash uchun ishonchliroq manba hisoblanadi" [Koteyko N., 2014].

Yuqoridagi fikrlardan kelib chiqib, tadqiqotimiz ob'yekti sifatida "Oila va jamiyat" va "Anaheim Bulletin" gazetalarining 1991-2020 – yillar oralig'ida chop etilgan nashr sonlari tanlandi. Bu orqali ishonchilik, reprezentativlik va haqqoniylik tamoyillariga asos bo'la oladigan materiallar sifatida e'tirof etilishi mumkin bo'lgan baza shakllantirildi. Bunda tanlab olingan o'zbek tilining "Oila va jamiyat" va ingliz tilining "Anaheim Bulletin" gazetalarining 30 yillik nashrlaridan olingan matnlar manba sifatida tanlanganligiga asos sifatida quyidagilarni sanab o'tish mumkin:

Birinchidan, yuqorida aytilgan 30 yillik muddat, ya'ni, 1991–2020-yillar oralig'i o'zbek tili uchun 1989-yil 21-oktabrdagi "O'zbekiston Respublikasining davlat tili to'g'risida" gi qonuniga muvofiq unga davlat tili maqomi berilganidan keyingi revoliutsiya, ya'ni, tillarning eng samarali rivojiga olib keluvchi hisoblangan davlat mustaqillik davrining ilk uchta o'n yilligi sifatida dolzarblik kasb etadi. Ingliz tilining ham ayana shu davrlarda chop etilgan gazeta nashrlaridan yig'ilgan matnlar bilan ishlash maqsad qilindi.

Ikkinchidan, "Oila va jamiyat" va "Anaheim Bulletin" gazetolari nisbatan kam o'rganilgan, haftalik chiqadigan, nashrlar chop ettirish boshlanish davri bir biriga yaqin bo'lgan hamda eng muhimi ijtimoiy gazetalar sirasiga kirishini inobatga olib tanlab olingan. Yanada aniqroq aytganda, "Anaheim Bulletin" gazetasi 1966-yildan mustaqil gazeta sifatida Amerika Qo'shma Shtatlari Kaliforniya shtatida joylashgan eng ko'p aholi istiqomat qiluvchi Anaheim shahrida chiqarib kelinmoqda. Gazeta o'zining tarixida 1992-yilning 6-avgust sanasidan boshlab haftada bir marta chop etiladigan nashrlar sirasiga o'tgan, bu esa 1991-yil 5-sentabr sanasidan "Oila va jamiyat" gazetasining har haftaning payshanba kuni chop etilishi yo'lga qo'yilgan. Davriy yaqinlik mavjudligi hamda ular ko'taradigan ilm-fan, madaniyat, san'at, ishlab chiqarish, tadbirkorlik, qishloq xo'jaligi va yana bir nechta sohalarga bag'ishlangan ruknlari o'zaro qiyoslashga mos gazetalar ekanligini ko'rsatdi.

Uchunchidan, “Oila va jamiyat” va “Anaheim Bulletin” gazetalarining tanlangan davr oralig’idagi nashrlari arxivining mavjudligi, garchi har ikkala gazetalarning eski nashrlarida kompyuter o’qiy oladigan imkoniyatning mavjud emasligi kabi maummoning uchraginini inobatga olmaganda, aynan shu manbalar tanlanganiga asos bo’ladi.

Har ikkala baza o’zida gazetalarning skanerlangan holatini salovchi <https://newspaperarchive.com/> hamda <http://press.natlib.uz/ru/publications> rasmiy saytlardan olindi. Ushbu saytlarda barcha ma’lumotlar foydalanuvchilar uchun ochiq ekanligi va cheklovlarning mavjud emasligi tadqiqotda tug’ilishi mumkin bo’lgan mualliflik huquqi (copyright) tamoyillariga zid harakat bajarilmagini ko’rsatadi. Matnlar bilan ishlash me’yorlari quyidagi tartibda uch bosqichda amalga oshirildi:

1-bosqich: <https://newspaperarchive.com> va <http://press.natlib.uz> rasmiy saytlaridan mavjud skanerlangan gazeta nashrlari to’plandi. Bunda, ularning katta hajmli jildlarda yillar kesimida, mavjud nashrlar ixtiyoriy tarzda, matnlarni o’qiy olinishidagi sifat ko’rsatkichiga tayangan holda saralab tartiblandi;

2-bosqich: Jamlangan va xronologik tartiblangan baza Microsoft Word dasturi orqali qo’l mehnati yordamida birma-bir qayta terildi va *.doc kengaytmali hujjatlar shaklida saqlandi. Bu turdagi hujjatlarda har ikkala tildan olingan ma’lumotlar maxsus ishlab chiqilgan jadvallarga joylashtirildi (qarang 1-jadval), hamda ushbu jadvallarni umumlashtirgan alohida jildlar yaratilib saqlandi. Xronologik yig’ilgan manbalarga yillar kesimida +/- 5000 so’zli bazalar shakllantirish talabini qo’yishni lozim deb topdik. Zero, miqdoriy me’orning bo’lishi qidirilayotgan so’zning qo’llanishiga oid yillar kesimida faollik darajasini ko’rsatuvchi statistik ma’lumotning ishonchliligini isbotlaydi.

1-jadval

Gazeta matnlarini kiritish uchun mo’ljallangan maxsus jadval namunasi

<i>Gazeta nomi:</i> Oila va jamiyat
<i>Titul:</i> Har bir shaxs mehnat qilish huquqiga ega....
<i>Matn:</i> “Jamiyat hayotining mezoni bo’lgan Konstitutsiyamizni chuqur va har tomonlama o’rganishimiz, uning ma’nosiga yetib borib, amaliy hayotimizning doimiy qo’llanmasiga aylantirishimiz darkor” – Islom Karimov. O’zbekiston Respublikasi Konstitutsiyasining 37-moddasida belgilanishicha: “Har bir shaxs mehnat qilish, erkin kasb tanlash, adolatli mexnat sharoitlarida ishlash va qonunda ko’rsatilgan tartibda ishsizlikdan himoyalaniş xuquqiga egadir. Sud hukmi bilan tayinlangan jazoni o’tash tartibidan yoki qonunda ko’rsatilgan boshqa hollardan tashqari majburiy mehnat ta’qiqlanadi”. Ushbu moddaga asosan har bir so’z manfaatlariga ko’ra qanday faoliyat bilan shug’ullanishni qanday kasb va mutaxassislikni tanalashni o’zi xal etadi. Mexnat to’g’risidagi kodeksga binoan, ishga qabul qilishni rad etish taqiqlangan Konstitutsiyada ko’rsatilgan adolatli mehnat sharoitlarida ishlash huquqi fuqarolar mehnat qilish huquqining tarkibiy qismidir. Mehnat sharoitlari bevosita korxonaning o’zida va ma’muriyat bilan o’rtasidagi kelishuvga ko’ra qonun asosida belgilanishi mumkin. Mehnat to’g’risidagi qonun bo’yicha har bir fuqaroga mehnat qilish huquqini himoya qilish, jumladan, sud orqali himoya qilish imkoni beriladi. Xodimni ishga tiklash to’g’risida sudning qonuniy kuchga kirgan qarori ma’muriyat uchun majburiy bo’lib, u darhol ijro etilishi shart. Bunda xodimning buzilgan huquqi tiklanish bilan birga, majburan ishga chiqqan kunlari uchun haq to’lanadi.

“Aholini ish bilan ta’minlash to’g’risida”gi O‘zbekiston Respublikasi qonunida fuqarolarni ishsizlikdan himoyalashning konstitusiyaviy huquqi belgilab berilgan. Ushbu qonun bozor iqtisodiyoti va mulkchilikning turli shakllariga o’tish sharoitida insonning ish bilan taminlanishga oid huquqlarini ro’yobga chiqarishning tashkiliy, nazariy ijtimoiy-iqtisodiy kafolatlarini belgilab beradi. Bu qonun mushkul keladigan ishni tanlash ishga joylashda haq olmay ko’mak- pishishni, haq olmay yangi kasb o’rgatishni, ishsizlik bo’yicha nafaqa to’lashni, ishidan mahrum bo’lganlarga beriladigan boshqa kafolatlarni ko’zda tutadi. O‘zbekiston Respublikasining 1992 – yil 2 – iyulda qabul qilingan “Kasaba uyushmalari, ularning huquqlari va kafolatlari to’g’risida”gi Qonuniga muvofiq, ma’muriyat tashabbusi xodimni ishdan bo’shatishga final kasaba uyushmasi komiteti roziligi bilan yo’l qo’yiladi. Fuqarolarning mehnat qilish huquqini buzganlikda aybdor bo’lgan mansabdor shaxslar qonun bo’yicha intizomiy, moddiy mamuriy javobgarlikka, muayyan hollarda esa jinoiy javobgarlikka tortilishlari mumkin.

Muallif: X. Tuxtamurod

Havola: <http://press.natlib.uz/ru/editions/59043>

Nashr yili va soni: 2011-yil 11-son

So’zlar soni: 294

Ushbu bosqichda, ya’ni matnlarni qayta terib kiritishda ikkinchi bobda o’zbek tilining diaxronik korpuslar yaratishdagi muammolari sirasiga kiruvchi yozuv grafikasidagi haligacha saqlanib kelinayotgan har xillikni yechish maqsadida matn to’liq lotin grafikasida kiritildi (qarang 2.2-fasl). Bunda, jarayon transliteratsiya qoidalari asosida amalga oshirilgan bo’lib, so’zlarning yozilishidagi orfografik uslubidagi barcha holatlar aslida qoldirildi, tuzatish va o’zgartirishlar joriy etilmadi. Misol tariqasida, Oila va jamiyat gazetasining 2020-yil 26-avgust sanasida chiqarilgan №27-28 (1479) nashr raqamli sonida maqola sarlavhasi lotin grafikasida kiritilgan bo’lsa, uning matni krill grafikasida ekanligi ana shunday muammolar sirasiga kiradi. Muammoligi shundaki, gazeta ilk bora nashr qilingan kezlar misolida 1991-yil 26-sentabrda chiqarilgan 3-4 (04) nashr raqamli sonida maqolalar to’liq krill grafikasida bo’lganligi ko’rinadi (qarang 1 va 2-rasmlar).



1-rasm. “Oila va jamiyat” gazetasining 1991-yil 26-sentabr sanasida chiqarilgan soni



2-rasm. “Oila va jamiyat” gazetasining 2020-yil 26-avgust sanasida chiqarilgan soni

Afsuski, bu holat yuqorida aytilgan misoldagi 30 yillik muddat oralig'ida ham o'nglanmaganligi, o'zbek tili uchun ulkan fojea sifatida qaralishi mumkin. Korpus asosan kompyuter vositasida bajariladigan amalga suyanishini inobatga olganda, korpus ikki xil yozuv grafikasidan kiritiladigan bir so'zni ikki turli so'zlar sifatida talqin qilishini nazarda tutgan holda, shu yuqoridagi transliteratsiya usulini taklif etdik. Ingliz tili bu rivojlanish bosqichidan korpusga kiritilishi ko'zda tutilgan davr oralig'idan oldin o'tib bo'lganligi bois “Anaheim Bulletin” gazetasidan olingan ma'lumotlar o'zgarishsiz manba tarkibiga kiritildi.

FOYDALANILGAN ADABIYOTLAR RO'YXATI

1. Anaheim Bulletin (Anaheim, Calif.) 1966-Current Available at <https://www.loc.gov/item/sn95061766/> [retrieved on 08.02.2023]
2. Anaheim Bulletin <https://newspaperarchive.com/>
3. Anthony L. AntFileConverter (Version 1.2.1) [Computer Software]. – Waseda University, Tokyo, Japan. 2017. Available at. <http://www.laurenceanthony.net>
4. Dossena M. Towards a corpus of nineteenth-century Scottish Correspondence. // *Linguistica e Filologia*, 18, 2004. – pp. 195-214.
5. Douglas F. M. The Scottish corpus of texts and speech: Problems of corpus design // *Literary and Linguistic Computing*, 18(1). 2003. – pp. 23-37.
6. <https://lex.uz/ru/docs/-109394> [Murojaat sanasi 09.02.2023]
7. Koteyko N. Compilation of Specialized Corpora // *Language and Politics in Post-Soviet Russia*. – University of Leicester, UK. 2014. – PP. 48-49 Available at https://link.springer.com/chapter/10.1057/9781137314093_4
8. Mackay S. Records of the Cape Colony (1821–1822) (Vol. 14). Available at <https://www.eggsa.org/1820-settlers/index.php/records-of-the-cape-colony>
9. Moreton E. Profiling the female emigrant: A method of linguistic inquiry for examining correspondence collections. // *Gender and History*, 24(3), 2012. – pp. 617-646.
10. Morton J. G. The Indian army on the Western Front; 1914–1915: A portrait of collaboration. // *War in History*, 13(3), 2006. – pp. 329-362.

11. O'zbek xalq dostonlarining chastotali izohli frazeologik lug'ati /B.Yo'ldoshev, D.O'rinboyeva; Mas'ul muharrir B.O'rinboyev; O'zbekiston Respublikasi oily va o'rta maxsus ta'lim vazirligi, Alisher Navoiy nomidagi Samarqand Davlat universiteti. – Samarqand, 2013. – 102 b. <https://elib.buxdupi.uz/books/34891.pdf>
12. Oila va jamiyat gazetasining №27-28 (1479) Nashr raqamli soni <http://press.natlib.uz/uz/editions/30092>
13. Oila va jamiyat <http://press.natlib.uz/ru/publications>
14. Oila va jamiyat <https://milliycha.uz/oila-va-jamiyat/> [murojaat sanasi]
15. Pennebaker J., Mehl M., and Niederhoffer K. Psychological aspects of natural language use: Our words, our selves. // Annual Review of Psychology, 54, 2003. – pp. 547-577. Available at <https://www.annualreviews.org/doi/pdf/10.1146/annurev.psych.54.101601.145041>
16. Sinclair J. Corpus and text – basic principles. In M. Wynne (Ed.) // Developing linguistic corpora: A guide to good practice. Oxford: Oxbow Books. 2005. – pp. 1-16
17. Sowada L. Writing in the First World War: French letters and diaries from inexperienced writers from the Franco-German border region (Vol. 456). – Heidelberg University. 2021. Available at <https://www.degruyter.com/document/doi/10.1515/9783110719161/html>