O‘ZBEKISTON MILLIY UNIVERSITETI XABARLARI, 2024, [1/10/1]
ISSN 2181-7324

FILOLOGIYA
http://journals.nuu.uz
Social sciences

UDK: 8 81 811 81`32

Sayyora ABDURAHMANOVA,
Alisher Navoiy nomidagi ToshDO‘TAU tayanch doktoranti
E-mail: sayyoraabdurahman@gmail.com

ToshDO‘TAU professori, f.f.d Z.Xolmanova taqrizi asosida

THEORETICAL FOUNDATIONS OF THE CREATION OF THE SHEVA CORPUS (FORMATION AND DEVELOPMENT OF UZBEK CORPUS LINGUISTICS)

Annotation

The formation and development of Uzbek corpus linguistics. This article touches upon formation and development of Uzbek corpus linguistics and the results of scientific and practical research. The practical works carried out are presented analytically and critically using comparative methods, and the issue of development of Uzbek corpus linguistics is considered. At the same time, there are opinions and theoretical information on the creation of a dialectal corpus of the Uzbek language on the basis of corpora created for the Uzbek language.
Key words: corpus, corpus linguistics, national corpus, educational corpus, Uzbek language corpus, dialectal corpus, subcorpus.

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ СОЗДАНИЯ ДИАЛЕКТНОГО КОРПУСА (СТАНОВЛЕНИЕ И РАЗВИТИЕ УЗБЕКСКОГО КОРПУСНОГО ЯЗЫКОЗНАНИЯ)

Аннотация

В статье рассматриваются вопросы становления и развития узбекского корпусного языкознания и результаты научных и практических исследований. Проведенные практические работы представлены аналитически и критически с использованием сопоставительных методов, а также рассмотрен вопрос развития корпусной лингвистики узбекского языка. Вместе с тем, имеются мнения и теоретические сведения о создании диалектного корпуса узбекского языка на основе корпусов, созданных для узбекского языка.
Ключевые слова: корпус, корпусная лингвистика, национальный корпус, учебный корпус, корпус узбекского языка, диалектный корпус, подкорпус.

SHEVALAR KORPUSINI YARATISHNING NAZARIY ASOSLARI (O‘ZBEK KORPUS LINGVISTIKASINING SHAKLLANISHI VA TARAQQIYOTI)

Annotatsiya

Mazkur maqolada o‘zbek korpus lingvistikasining shakllanishi va taraqqiyoti va bu borada olib borilayotgan ilmiy va amaliy izlanishlarning natijasiga to‘xtalib o‘tilgan. Bajarilgan amaliy ishlar qiyosiy metoddan foydalanilgan holda tahliliy va tanqidiy fiklar bayon qilinib o‘zbek korpus lingvistikasini rivojlantirish masalasiga to‘xtalib o‘tilgan. Shu bilan bir qatorda o‘zbek tili uchun yaratilgan korpuslar asosida o‘zbek tili dialektal korpusini yaratish borasidagi fikrlar, nazariy ma’lumotlar mavjud.
Kalit so‘zlar: korpus, korpus lingvistikasi, milliy korpus, ta’limiy korpus, o‘zbek tili korpusi, dialektal korpus, subkorpus.

Kirish. Mamlakatimizda tilga e’tibor ma’naviyatga e’tiborning ustuvor yo‘nalishlaridan biri darajasiga ko‘tarildi. Shu bois ona tilimizni avaylab-asrash, boyitish, undan amaliy foydalanish samaradorligini oshirish bilan birga, o‘zbek tilining zamonaviy axborot-kommunikatsiya tizimida keng qo‘llanishiga erishish kechiktirib bo‘lmaydigan dolzarb vazifaga aylandi. Chunki ona tilimizning jahonga chiqishiga erishish milliy ma’naviyatni takomillashtirish va yuksaltirishning asosiy yo‘llaridandir[1].

XIX asr oxiri XX asr boshlaridan boshlab tilni tadqiq etishning yangi bosqich va nazariyalari yaratila boshladi. Aynan mana shunday yangi qarashlar amaliy tilshunoslik sari tashlanayotgan ilk qadam hisoblangan. Tilshunoslik fanida ham integratsiya natijasida kompyuter lingvistikasi fani rivojlandi. Aynan mana shu fan zamirida amaliy tilshunoslik masalalarini yechishga qaratilgan avtomatik tahrir, avtomarik tarjima korpus lingvistikasi kabi yo‘nalishlar yangi taraqqiyot bosqichiga ko‘tarildi. Xususan, kompyuter lingvistikasi va u asosida yuzaga kelgan korpus lingvistikasi mana shunday amaliy ahamiyatga ega yo‘nalishlardan sanalgan. Bu borada dastlabki tadqiqotlar injener lingvistikasi nomi ostida birlashgan bo‘lsa bugungi kunga kelib kompyuter lingvistikasi tarkibiga kiruvchi har bir yo‘nalish mustaqil fan darajasiga ko‘tarila oldi. Bu boradagi ilk tadqiqotlar sun’iy intellekt haqidagi farazlardan avvalroq boshlanganini inobatga olsak, dastlabki kompyuter lingvistikasiga oid ishlar inson mehnatiga tayanilgan holda bajarilgani va bu jarayon juda katta kuch talab etgani shubhasiz.

Zamonaviy axborot texnologiyalari tilning funksional imkoniyatlaridan foydalanish borasida cheksiz imkoniyatlar eshigini ochdi. Xususan, tilning imkoniyatlarini namoyon qilish va egallash borasida dunyo miqyosida tez sur’atlarda yaratilayotgan til korpuslarining roli beqiyos[2]. Korpus lingvistikasi ham kompyuter lingvistikasi tarkibidagi etalon yo‘nalishlardan biri bo‘lib, tabiiy tilni saqlab qolish, uni qayta ishlashga xizmat qiladi. Bu boradagi ilk tadqiqotlar sun’iy intellekt haqidagi farazlardan avvalroq boshlanganini inobatga olsak, dastlabki kompyuter lingvistikasiga oid ishlar inson mehnatiga tayanilgan holda bajarilgani va bu jarayon juda katta kuch talab etgani shubhasiz.

Mavzuga oid adabiyotlar tahlili. Korpus (korpus) lotincha so‘z bo‘lib, «tana» degan ma’noni bildiradi. “Korpus so‘z, so‘z birikmasi, grammatik shakllarni, so‘z ma’nosini muayyan qidiruv tizimi orqali topishni anglatuvchi electron ko‘rinishdagi matnlar jamlanmasidir”[3]. Korpus tushunchasi bilan yonma-yon “matnlar korpusi” atamasi ishlatilmoqda. Matnlar korpusi elektron holda saqlanadigan fonema, grafema, morfemalar, leksema, gap va matnlardan tashkil topishi mumkin bo‘lgan yaxlit butunlikdir. Korpuslar aslida ma’lumotlar bazasi sifatida shakllantiriladigan, tilshunoslik masalalarini hal etish maqsadida va turli yo‘nalishdagi tadqiqotlarni amalga oshirish uchun material sifatida xizmat qiladigan jamlanmadir[4]. Bir so‘z bilan aytganda korpus bu – til birliklarining xususiyatlarini aniqlash maqsadida qidiruv dasturiga bo‘ysundirilgan matnlar majmui, tabiiy tildagi elektron shaklda saqlanadigan yozma yoki og‘zaki, kompyuterlashtirilgan qidiruv tizimiga dasturiy ta’minot asosida joylashtirilgan onlayn yoki oflayn tizimda ishlaydigan matnlar jamlanmasi hisoblanadi. Bu yo‘nalish 1960-yillarda Amerika Qo‘shma Shtatlarida tilshunoslikning yangi yo‘nalishi sifatida paydo bo‘lgan. “Dastlabki matnlar korpusi (Braun korpusi)” 60-yillarda AQShda yaratilgan va Amerika bosma nasrining lingvistik xususiyatlarini aks ettirish uchun mo‘ljallangan. Bu korpus magnit tashuvchi (disketa yoki qattiq disk)ga yozib olingan va umumiy hajmi millionga yaqin so‘zdan iborat AQSh bosma nasrining turli matnlarining besh yuz ikki ming (502 000) so‘zli parchalarini o‘z ichiga olgan. Bu borada Rikov quyidagi farazni ilgari suradi. Ehtimol, Braun korpusi yaratuvchilari uchun kutilmaganda:

boshqa shunga o‘xshash korpuslarni yaratish uchun o‘ziga xos standartga aylandi;

korpus tilshunosligida yangi fanning yaratilishiga turtki bo‘ldi;

matnlar korpusi va korpus tilshunosligi usullarini qo‘llash sohasi uni yaratuvchilar kutganidan ham ancha kengroq va rang-barang bo‘lib chiqdi[5].

Darhaqiqat Braun korpusi keyinchalik yaratilgan boshqa til korpuslari uchun andoza vazifasini o‘tadi. O‘tgan asr so‘ngida mazkur jarayon jadallashdi, XXI asr boshlarida millionlab so‘zlarni aks ettirgan yuzlab til korpuslari paydo bo‘ldi. Sun’iy intellektning avtomatik tarjima, kompyuter tahlili, tezaurus, elektron lug‘at singari imkonioyatlari kengaydi, ilmiy, nazariy asoslari yaratildi, amaliyotda qo‘llash mumkin bo‘lgan ilk namunalari qo‘llanila boshladi[6]. Bu jarayon tabiiy tilni qayta ishlash borasida tashlangan katta qadam edi. Dunyo tilshunosligida korpus lingvistikasi fan sifatida o‘qitiladi va bu yo‘nalishning obekti korpus yaratish nazariyasi va amaliyotini o‘z ichiga olsa, fan sifatida dasturlash, dasturiy ta’minot yaratish ustuvorlik qiladi. Korpus lingvistikasi kompyuter lingvistikasining tarkibiy qismi bo‘lib, til korpusini yaratish kompyuter texnologiyasi yordamida ulardan foydalanishning umumiy nazariyasi va amaliyoti bilan shug‘ullanadi[7]. Fanning obekti va maqsadidan kelib chiqqan holda korpus lingvistikasining predmeti til korpusi soblanadi. Bugungi kunga kelib korpus tushunchasini ifodalashga xizmat qiluvchi ko‘plab ta’riflar mavjud. E. Finegan: “Korpus – bu odatda, kompyuter o‘qiy oladigan formatda bo‘lgan va bizga matn ishlab chiqilgan vaziyat, informatsiya beruvchi, muallif, adresat yoki auditoriya haqidagi ma’lumotni o‘z ichiga olgan matnlar to‘plamidir”, - deydi. Turli umumiy ma’lumot beruvchi ijtimoiy saytlar korpusni statistik analiz hamda farazlar tekshiruvi asosida anglangan sohalarda uchrovchi til qoidalari va hodisalarini asoslay oladigan katta hajmdagi, tizimli matnlar to‘plami (endilikda odatiy elektron shaklda) sifatida ta’riflaydi[8]. Bundan kelib chiqib aytishimiz mumkinki korpus yaratish tamoyillari umumiy bo‘lsa, har bir tilning tabiatidan kelib chiqqan holda korpuslar xususiylik kasb etadi.

O‘zbek korpus lingvistikasi ham ayni vaqtda taraqqiyot bosqichiga ko‘tarilib ulgurdi. O‘zbek tili milliy korpusi[9], O‘zbek tilining ta’limiy korpusi[10], O‘zbek tili korpusi[11], Sahxlo Hamroyevaning O‘zbek tili mualliflik korpusini tuzishning lingvistik asoslari monografiyasi, Oqila Abdullayevaning O‘zbektilining internet axborot matnlari korpusini shakllantirishning nazariy va amaliy asoslari dissertatsiyasi hamda “Tug‘ro korpusi[12], Nargiza G‘ulomovaning Alisher Navoiy mualliflik korpusi[13], Dialektal korpus[14], Aziza Rahmonovaning O‘zbek tili milliy korpusini yaratishda kompyuter usullari nomli f.f.f.d. PhD dissrtatsiyasi, Anorxon Ahmedovaning Parallel korpusda o‘xshatishlarning leksik-semantik munosabatlari tadqiqi monografiyasi kabilar bu borada amalga oshirilgan ishlar ko‘lamidan dalolat beradi.

Tahlil va natijalar. Dastlabki tadqiqotlar asosan o‘zbek tili tabiatidan kelib chiqqan holda korpus yaratishning nazariy masalalariga qaratilgan bo‘lsa, keyingi tadqiqotlarda amaliy masalalarga e’tibor qaratildi. O‘zbek tilining milliy korpusi, O‘zbek tilining ta’limiy korpusi va O‘zbek tili elektron korpusini yaratish borasioda borasidagi amaliy ishlar bugungi kunga qadar davom etmoqda. Korpusshunoslikda, xususan, dialectal korpuslarni yaratish borasida olib borilgan amaliy va nazariy tadqiqotlarni ikki guruhga[15] ajratishimiz mumkin. Bulardan birinchisis o‘zga til korpuslari asosida o‘z tili va shevalari xususiyatlarini inobatga olib yaratilgan korpuslar: Janub ovozi[16], Portugaliya korpusi[17] – The syntax-oriented Corpus of Portuguese dialects, 1972–1974 yillarda Finlandiyaning Xelsinki tumani aholisining so‘zlashuviga asoslangan korpus[18] Bolgar multimedia korpusi[19], Polsha korpusi (parallel korpus)[20] – Dialecty I gwary polskie Kompendium internetowe, Parijning og‘zaki so‘zlashuv korpusi[21], Ispan korpusi[22] – Corpus Oral y Sonoro del Español Rural, Bolqor korpusi[22] – Bulgarian dialectology as Living Tradition, Gurjiston korpusi (2 054 705 ta so‘z, 3400 ta matndan iborat)[22], Tuva dialektal korpusi[25], Alban korpusi[26] va hokazolar, Xitoy tili Mandarin korpusi (bu dialekt xitoy tilidagi turlilik va o‘zgaruvchanlikni o‘zida namoyon etadi) va hokazo.

Ma’lum bir hududning shevasi asosida yaratilgan o‘sha til xususiyatlarini inobatga olib yaratilgan korpuslar: Arxangelsk viloyatining Ustya tumani sheva korpusi[27], Kuban dialektal korpusi[28], Rus milliy shevalarining elektron kutubxonasi[29], Rus milliy shevalarining elektron bazasi[30], Bolgar multimedia korpusi[31], Saratov dialektal korpusi va hokazolar. Dialektlar korpusi yaratilishida anchagina ishlar amalga oshirilgan. Dialeklat korpuslarni yaratishda bir qancha me’zonlar mavjud ekanligi ham dunyo tilshunoslari e’tiboridan chetda qolmagan. O‘z tili shevalarining fonetik, leksik, grammatik xususiyatlarini o‘zida aks ettira oladigan korpus yaratish masalasi birinchi o‘rinda turgan. Milliy korpus bilan bir qatorda dialectal kjorpus yaratish borasidagi amaliy ishlar dastlab ingliz dialektlari bilan bog‘liq edi. Xelsinki korpusi (HD) – Sharqiy Angliya, Janubi-g‘arbiy Lankashir orollaridan aholisi bilan suhbatlashib orfografik transkripsiyalangan audioyozuvli nutq to‘plami bo‘lib, ushbu yozuvlar 1970–1980 yillar oralig‘ida tadqiqotchilar tomonidan olib borilgan tadqiqotlar natijasidir[32]. Ushbu dialektologik korpusning maqsadi nafaqat dialektologiya, balki nutq madaniyati, leksikologiya, sotsiolingvistika, grammatika va fonetika, fonologiya sohalarida lingvistik tadqiqotlar uchun material berishdir. Binobarin, korpus etnografiyasi, mahalliy aholining urf-odatlari, qisman tarix va boshqa ko‘plab lingvistik tadqiqotlar uchun tadqiqot maydoni vazifasini o‘taydi. Bu dialektologik korpus uchun Essek Lankashirga Rita Kerman 1988 yil o‘zining (Pro Gradu) diplom ishi uchun yig‘gan ma’lumotlari asos bo‘ladi. U tadqiqotlari uchun mezon sifatida unga ma’lumot beruvchi shaxslar uncha ham ko‘p vaqtni maktabda o‘tkazmagan, ya’ni o‘rta ma’lumotli shaxs bo‘lishi kerakligini va aynan ular tomonidan berilgan ma’lumotlar o‘z qiymatini yo‘qotmagan, adabiy til bilan sayqallanmagan holda berilishini lozimligini ta’kidlaydi[33]. Bundan ko‘rinadiki tadqiqotchi mahalliy aholi madaniy boyligini tushunish uchun sheva vakillari orrasidan ijtimoiy kelib chiqishi, ma’lumoti, yoshini inobatga olishi lozim. Etnolingvistik ekspeditsiya nafaqat “axborotni o‘rganish va materiallarni yig‘ish”, balki qishloq aholi yashovchi hududlarda yashash tadqiqotchida an’anaviy madaniyat bilan “nafas olish”, “oziqlanish” imkonini beradi[34] va shu bilan birga dialektal korpus yaratish jarayonida soflik va xolislikni taminlaydi. Yana bir dialektal korpus o‘z ichiga Daniya, Islandiya, Farer, Norvegiya va Shvedsiya kabi mamlakatlarning og‘zaki so‘zlashuv tilini qamrab olgan Norvegiya shevalar korpusi yani Nordic Dialect corpus (NDC) hisoblanadi[35]. U o‘z ichiga Shimoliy o‘lkada istiqomat qiluvchi aholining shimoliy German dialekti aralash shevalarni olgan va teglar (so‘zlar) bilan belgilangan adabiy tilga oid birliklar, og‘zaki yozuvlar hamda transkriptlar korpusning tarkibiy qismi sifatida tilga olinadi. Korpus uchun yig‘ilgan ma’lumotlar Daniya, Norvegiya, va Islandiya, yozuvlari milliy tadqiqot kengashlari moliyaviy ko‘magi asosida loyiha sifatida bajarilgan amaliy ish hisoblanadi[36].

Mamlakat	Ma’lumot beruvchilar	Hududlar	Identifikatsion belgilar
Daniya	81	15	220,360
Farer	20	5	64,803
Islandiya	48	8	94,338
Norvegiya	438	111	1,997,920
Shvesiya (shu jumladan Ovdalian)	150	44	376,868

Yevropa dialektal korpuslari. 1

Dialektal korpusni yaratish bo‘yicha jahon tajribasiga nazar tashlaydigan bo‘lsak sharq mamlakatlarida bu borada qanday ishlarning amalga oshirilganligini ko‘rib chiqish joiz bo‘ladi. Shevasiga ega bo‘lmagan tillarda ham bu jarayon yani korpusshunoslik bo‘yicha qilingan ishlar natijasini Arab shevalari korpusi (MSA) misolida ko‘rish mumkin. Arab tilshunosligidagi

rasman sheva qayd etilmagan bo‘lsada onlayn elektron xatlar, chatlar, o‘zaro suhbatlar, bloglar va sms shaklidagi yozishmalarda qayd etilib, sekin astalik bilan adabiy tilga aylanib bormoqda va mana shu jarayon arab tiulshunoslarini arab dialektlarining keng qo‘llanishini hududlarning kengayishi bilan arab standart tilidan fonologik, leksik va morfologik sathda o‘zgarishlarda kuzatish mumkin[37].

Xulosa va takliflar. Korpus lingvistikasi shakllanishi bilan bog‘liq nazariyalar XX asrning ikkinchi yarmidan boshlangan bo‘lksa, mamlakatimizda o‘zbek tilini avtomatlashtirish, o‘zbek tili xususiyatlaridan kelib chiqib korpus yaratish nazariyalarini ishlab chiqish 2000-yillardan boshlangan. Bugungi kunga kelib o‘zbek tilida mavjud bir qancha korpus til imkoniyatlarini ochib berishga xizmat qilmoqda. Shu alohida aytib o‘tish joizki til korpuslari tilning saqlab qolish, kelajak avlodga sof holatda yetkazish uchun zarur bo‘lsa, dialektal korpuslar tilni ichki manbaa hisobiga boyishini taminlaydi.

ADABIYOTLAR

1. Mengliyev B. Globallashuv : tillar taraqqiyoti va tanazzuli. “Ma’rifat” gazetasi , 2017-yil, 14-oktabr, 82-son.
2. Mengliyev B. Tilshunoslikning amaliy masalalari Monografiya Globeedit Toshkent 2020
3. http:/ruskorpora.ru
4. Z.Xolmanova. Korpuslarning lingvistik tadqiqot materiali sifatidagi ahamiyati.52-bet (Баранов А.Н. Введение в прикладную лингвистику. – М.: Эдиториал УРСС, 2001)
5. Rykov, V. V. Text corpus as an implementation of the objekt-oriented paradigm In Prosedings of the international seminar “Dialogue-2002” . Moscow: Nauka. PP. 124-129
6. Xamrayeva Sh. O‘zbek tili mualliflik korpusi.-Toshkent: Globe edit, 2020. – 168 b.
7. Захаров В.П. Корпусная лингвистика: Учебно-метод. пособие. – СПб., 2005. -48 с, - С. 27
8. Захаров В.П. Корпусная лингвистика. – Иркутск, 2011.С. 7.
9. http://uzbekcorpora.uz/ijrochi
10. https://uzschoolcorpara.uz/
11. https://uzbekcorpus.uz/
12. Abdullayeva O. O‘zbek tilining internet axborot matnlari korpusini shakllantirishning nazariya va amaliy asoslari. f.f.f.d – Toshkent, 2022. – b. 158. https://uzkorpus.uz/
13. G‘ulomova N. Alisher Navoiy mualliflik korpusi va uninng semantik teglari bazasini yaratish. f.f.f.d – Toshkent, 2022. – b. 189. https://navoiykorpusi.uz/
14. https://dialect.uz/
15. Kawai Chui., Huei-Ling Lai., The NCCU Corpus of Spoken Chinese: Mandarin, Hakka, and Southern\\ Min Article in Taiwan Journal of Linguistics. 2008. Vol. №4. Is. №9. – P. 119-144.; Вадеев С.Йе. Лингвистические принсипы построения и исползование корпуса текстов для исследования официально-делового стиля современного немеского языка: Автореф. дис…канд. фил. наук – М.: МГУ. 2005.; http://rusling.narod.ru/qqq_corp_nonslav_other.htm; http://dialekt.corpus.tatar/ https://ucnk.ff.cuni.cz/cs/; https://cyberleninka.ru/article/n/skandinavskiy-ostrov-v-slavyanskoy-yazykovoy- srede-dialekt-selastaro-shvedskoe-imya-suschestvitelnoe/viewer; http://www.ling.helsinki. fi/projects/hanco/; http://www.tekstlab.uio.no/scandiasyn/;; Диалектная култура Кубани в свете етнолингвистического анализа (по данным електронного корпуса диалектной культуры кубани)\ Трегубова Йе.Н., Сфинко О.С., Баласенко Н.С., Литус Йе.В. – Краснодар: Екоинвест, монография, 2017. – S. 204.
16. http://newsouthvoices.uncc.edu/nsv,
17. http://rusling.narod.ru/qqq_corp_nonslav_other.htm
18. http://www.ling.helsinki.fi/uhlcs/readme-all/README-uralic-lgs.html#C34
19. https://sarteorlingv.narod.ru/dialekt/kru4kova-goldin.html
20. https://ruscorpora.ru/new/search-para-pl.html
21. http://cfpp2000.univ-paris3.fr/Corpus.html
22. O‘sha manba. https://ruscorpora.ru/.
23. https://www.tuvancorpus.ru/?q=content/dialektnyy-korpus-tuvinskogo-yazyka
24. http://web-corpora.net/AlbanianCorpus/search/.
25. http://www.slavist.de/Pushkino/
26. https://ethnolex.ru/?page_id=421
27. http://dialekt.corpus.tatar/
28. http://www.ruslang.ru/krylov_dialect
29. https://sarteorlingv.narod.ru/dialekt/kru4kova-goldin.html
30. Xolova M. O‘zbek milliy shevalari korpusi tadqiqi (Boysun tumani “j”lovchi shevalari misolida) Monografiya. Termiz 2022 144-b.
31. http://www.helsinki.fi/varieng/CoRD/corpora/Dialects/fieldwork.html.
32. Xolova M. O‘zbek milliy shevalari korpusi tadqiqi (Boysun tumani “j”lovchi shevalari misolida) Monografiya. Termiz 2022 144-b.
33. http://www.tekstlab.uio.no/nota/scandiasyn/
34. Janne Bondi Johannessen., Kristin Hagen., Anders Noklestad., Joel Priestley. Tour de CLARIN: The Nordic Dialect Corpus\\ edited by Darja Fišer, Elisa Gorgaini, and Jakob Lenardič. 2020. №5 (17), – P. 178-b
35. Building Dialectal Arabic Corpora (acl-bg.org)