



“O‘ZBEK FOLKLORI VA SHEVALARI TADQIQOTLARI: AMALIYOT, METODOLOGIYA, YANGICHA YONDASHUV”

mavzusidagi xalqaro ilmiy-nazariy
konferensiya materiallari

2026-yil 23-may

*(Professor Samixon Ashirboyev tavalludining
80 yilligi munosabati bilan tashkil etilgan)*



“O‘zbek folklori va shevalari tadqiqotlari: *amaliyot, metodologiya, yangicha yondashuv*” mavzusidagi xalqaro ilmiy-nazariy konferensiya materiallari
/ Mas’ul muharrir: Sh.Sirojiddinov. – Toshkent, 2026. bet.

Mas’ul muharrir:
Shuhrat Sirojiddinov
filologiya fanlari doktori, akademik

Tahrir hay’ati:

Normurodova N.Z. filologiya fanlari doktori, professor;
Jabborov N.A. filologiya fanlari doktori, professor;
Ashirboyev S. filologiya fanlari doktori, professor;
Tilavov A.X. filologiya fanlari nomzodi, dotsent;
Xidraliyeva Z.R. filol. fanlari bo‘yicha falsafa doktori (PhD), dotsent v.b.;
Aliqulova H.Q. filol. fanlari bo‘yicha falsafa doktori (PhD), katta o‘qituvchi;
Turdialiyev A.X. filol. fanlari bo‘yicha falsafa doktori (PhD), dotsent v.b.;
Alimardanov E.I. filol.fanlari bo‘yicha falsafa doktori (PhD), kattao‘qituvchi;
Qodirov U.A., katta o‘qituvchi.

Ushbu to‘plamda Alisher Navoiy nomidagi Toshkent davlat O‘zbek tili va adabiyoti universitetida 2026-yil 23-mayda o‘tkazilgan **“O‘zbek folklori va shevalari tadqiqotlari: *amaliyot, metodologiya, yangicha yondashuv*”** mavzusidagi xalqaro ilmiy-nazariy konferensiya materiallari hamda filologiya fanlari doktori, professor Samixon Ashirboyevning tabarruk yoshi munosabati bilan yo‘llangan tabriklar o‘rin olgan.

To‘plam materiallaridan filolog mutaxassislar, ilmiy tadqiqotchilar, magistratura va bakalavriat talabalari, shuningdek, keng o‘quvchilar ommasi foydalanishlari mumkin. Maqolalar sifati, mazmuni va haqiqiyliги uchun muallifning o‘zi javobgar hisoblanadi.



	Nurafshon shahar shevasining ayrim dialektal xususiyatlari	
15	Sadoqat RAJABOVA, Ishtixonda yashovchi etnik arablar	246
16	Umarjon QODIROV, Shonazar Shoabdurahmonov – lingvofolkloristika asoschilaridan	249
17	Gulchexra AHATOVA, Toyloq shevasidagi iboralar va ularning ingvomadaniy tahlili	253
18	Hayotxon SOTIMOVA, Xorazm o‘zbek shevalarida oila konseptining urf-odat va marosimlardagi lingvomadaniy ifodasi	260
19	Ramazon ORTIQOV, “Ikkilamchi nominatsiya” tushunchasining ilmiy adabiyotlardagi talqini	264

3-SHO‘BA. DIALEKTAL LEKSIKOGRAFIYA, DIALEKTAL KORPUS, TOPONIMIYA VA LINGVOKULTUROLOGIYA

1	Hamidulla DADABOYEV, XI–XII asr turkiy til tibbiy terminologiyasi tizimining “devonu lug‘otit turk”da aks etishi 10.52773/tsuull.conf.2026.23.05/PTYD5470	271
2	Nishonboy HUSANOV, “Lug‘ati chig‘atoyi va turkiy” – o‘zbek shevalarini o‘rganishda muhim tarixiy manba 10.52773/tsuull.conf.2026.23.05/MAKE1648	276
3	Baxtiyor MENGLIYEV, Dilrabo ELOVA Lingvo-mantiqiy ekspertiza: yuridik lingvistikada integrativ yondashuv masalasi	281
5	Muhayyo HAKIMOVA, Milliy tafakkurning abstrkat leksika shakllanishiga ta’siri 10.52773/tsuull.conf.2026.23.05/MBYM8287	284
6	Nilufar ABDURAXMONOVA, Nasima ABDULLAYEVA, Dialektal birliklarning raqamli korpus modelini yaratish masalalari 10.52773/tsuull.conf.2026.23.05/TYZS7723	292
7	Zamira DJURABAYEVA, Fitonimlarning dialektal tabiati haqida ba’zi mulohazalar	297
8	Muyassar XOLOVA, Dialektal korpusda olmosh so‘z turkumining teglanishi	300
9	Anvarbek TURDIALIYEV, Ma’lumotlar bazasini shakllantirish va qidiruv tizimiga joylashtirish masalalari	304
10	Gulxayo ISABAYEVA, Sardor JO‘RAQULOV Tog‘ay Murod nasrida dialektizmning o‘rni (Otamdan qolgan dalalar romani misolida)	313
11	Feruza AMINOVA, Badiiy matnda nurota shevasiga xos morfologik shakllarning qo‘llanishi va uslubiy vazifasi	316
12	Oynisa XOJIYEVA, O‘zbek va turk tillarida paremiologik birliklarning struktur-semantik va funktsional munosabati	320
13	Dildora AMIROVA, “Bobur” antroponimining badiiy matnlarda allyuziv nom sifatida qo‘llanilishi	325
14	Kamola NOSIROVA, Dialectological features and lexical layer of alla texts in tashkent region	328

MA'LUMOTLAR BAZASINI SHAKLLANTIRISH VA QIDIRUV TIZIMIGA JOYLASHTIRISH MASALALARI

ISSUES OF DATABASE FORMATION AND PLACEMENT IN A SEARCH ENGINE

Anvarbek TURDIALIYEV,

ToshDO'TAU dotsent v.b

10.52773/tsuull.conf.2026.23.05/NLRI9494

Annotation. *In today's linguistics, with the rapid development of digitalization, it is important to create a database of corpora. This article discusses the methods of filling the database and the corpora that have been created so far.*

Key words. *Annotation, corpus, database, NLP, dialect, dialect, audio, video.*

Dunyo tilshunosligi zamon bilan hamnafas rivolanish bosqichida bo'lib, zamon bilan hisoblashmagan til egalari shu tilning kelajagini yo'qotib qo'yadi. Buni dunyo tilshunosligida qilinadigan tadqiqotlardan ham kuzatish mumkin. Har qanday tilning kelib chiqish asosi va uning milliyligini saqlab turgan lingvistik bazasi mavjud bo'ladi. Lingvistik bazalar esa bu har bir tilning o'z shevasi (shevalari) hisoblanib, sheva sofligi qanchalik yuqori saqlansa, tilning ham yashovchanligi o'sib boradi. Hozirgi axborot texnologiyalar rivojlangan dunyoda tillar xilma-xil bo'lishiga qaramay u doimiy kamayib bormoqda, chunki mahalliy nutq shakllari kichik til oilalari, rivojlanmagan davlatlar tili asosiy muloqot vositasi bo'lgan rivojlangan (davlatlar) tillar ta'sirida o'lib bormoqda. Shu nuqtayi nazardan Amerika tilshunoslik maktabi (LSA) saytida tillarning yo'qolishi haqida quyidagicha jumlar keltiriladi, "*Tilni yosh bolalar o'rganishdan to'xtata, uning (o'lim)kuni aniq sanaladi*" [Abdullayeva 2022; 14]. Demak, bundan shunday xulosaga kelish mumkinki, har qanday tilni asrab qolish uchun til ma'lumotlarini bir joyga to'plash uni qayta ishlash tahlil qilish va uni keng omma foydalanishi uchun qulay tizimga tushirish kerak.

Til korpuslari har qanday tilning qulay sistemalashtirilgan, foydalanish uchun qulaylashtirilgan dasturlar to'plami hisoblanadi. Dunyo tilshunosligida zamonaviy yo'nalish sanalganda korpus lingvistikasi va uning turlari, korpus yaratish, undan foydalanish usullari alohida tilga olinadi. Korpus lingvistikasi – bu lingvistik hodisalar yoki lingvistik o'zgarishlar haqida ma'lumot olish uchun tabiiy, mashina-da o'qiladigan matnlarning tizimli to'plamlariga tayanadigan metodologiya, – deb ta'riflanadi. Tilshunoslikda korpus haqida turli xil ta'riflar birlashib quyidagicha sharhlanadi: til birliklarining xususiyatlarini aniqlash maqsadida qidiruv dasturiga bo'ysundirilgan matnlar majmui, tabiiy tildagi elektron saqlanadigan yozma yoki og'zaki kompyuterlashtirilgan qidiruv tizimiga dasturiy ta'minot asosida joylashtirilgan matnlar jamlanmasidir [Abdullayeva 2022; 14]. Jahon tilshunosligida dastlabki korpus nemis olimi Kading nomi bilan bog'lanadi. Olim 1890-yillarning oxirida 11 mln nemis so'zlaridan iborat korpus yaratgan. Korpus nemis tilida harflar chastotasini aniqlash va harflar ketma-ketligini taqsimlash uchun ishlatilgan [Abdullayeva 2022; 14]. Korpus lingvistikasining rivojlanishi manbalarda 3 davrga bo'linadi: 1) Korpus lingvistikasining ilk davri – bu davr 1950-yilgacha bo'lgan davr bo'lib, bu davr kompyuter bo'lmagan vaqtlarga to'g'ri keladi har qanday korpus qo'lda ko'p kuch va vaqt sarflab tayyorlangan; 2) Xomskiy davri – 1950-60 - yillar oralig'i. Bu davr kichik bo'lsa ham korpus rivojlanishi uchun Xomskiyning qarashlari va ta'limoti yetakchi rol o'ynagan; 3) uchinchi davr 1960-yildan hozirgacha bo'lgan zamonaviy korpus lingvistikasi davridir. Korpus va korpus lingvistikasi nazariyalarining yaratilishi shu davrga to'g'ri keladi. Korpus so'zining

etimologiyasiga ham e'tibor qaratishni lozim topdik. Korpus – lotin tilida “*tana*” degan soʻzdan kelib chiqqan boʻlib, VI asrda “Corpus Juris Civilis” huquqiy matnlarini tavsiflash uchun ishlatilgan [Abdullayeva 2022; 23]. Oksford ingliz tili lugʻatida keltirilgan besh taʼrifning biriga koʻra, korpus bu “*lingvistik tahlilga asoslangan yozma yoki ogʻzaki materiallarning tanasi*”dir. Demak korpus faqatgina matnlar toʻplami emas, balki, lingvistik tahlil uchun foydalaniladigan maʼlum bir til, dialekt, boshqa tilning vakili deb taxmin qilingan matnlar toʻplamidir [Francis, Johansson 1982]. Korpus lingvistikasi (CL) “tadqiqotlarni korpusdan olingan maʼlumotlarga asoslaydigan tilshunoslikning bir tarmogʻi” deb taʼriflanishi mumkin. Buni lingvistik intizom emas, balki koʻplab fanlarda qoʻllanilishi mumkin boʻlgan uslubiy yondashuv deb qarash mumkin. Ushbu fanlararo bogʻliqlik tufayli u haqiqatan ham hozirgi tilshunoslik tadqiqotlarida tobora muhim ahamiyat kasb etmoqda. CL – lingvistik korpuslarni ishlab chiqish va tahlil qilishga asoslangan. Korpus – bu tilning haqiqiy matn namunalari (romanlar, pyesalar, ssenariylar, yangiliklar, insholar, radio va televideniye dasturlari stenogrammasi, suhbatlar yoki nutqlardan olingan) toʻplami. Ular odatda onlayn yoki elektron formatda paydo boʻladi, chunki ularning oʻlchamlari koʻpincha katta boʻladi. Oldindan aniqlangan obyektiv mezonlar boʻyicha tanlangan namunalar oʻrganilayotgan tilning ifodasini taʼminlashga qaratilgan. Xulosa qilib aytganda, korpuslar tilshunoslikning muhim vositasidir, chunki ular tilning turli tomonlarini yoki uning xilma-xilligini tadqiq qilish imkonini beradi.

Tadqiqotlarda matnlarni tanlashda qoʻllaniladigan mezonlar korpusning reprezentativligini belgilaydi. Koʻpincha quyidagi tasnif amalga oshiriladi: Umumiy korpus: ular tilning turli navlari, tuzilmalari va lugʻat boyligining namunaviy va toʻliq namunasini taqdim etadi, shuning uchun ular katta hajmga ega; korpusning bu turi rivojlanishning turli bosqichlarida tilning xususiyatlarini oʻrganish imkonini beradi.

Ixtisoslashgan korpus: ularning taʼrifi atrofidagi qarama-qarshiliklarga qaramay, ular koʻpincha aniqroq maqsadni koʻzlaydi va umumiy korpuslarga qaraganda kichikroq hajmda boʻladi; u kichik tilning vakili boʻlishni yoki umumiy til xulq-atvoriga ega boʻlgan maʼruzachilar guruhi tomonidan tildan foydalanishni maqsad qiladi.

Ikki tilli yoki koʻp tilli korpus: uning ichida ikkita turni topish mumkin: parallel korpus: asl matnlar va ularning bir yoki bir nechta tillarga tarjimalarini oʻz ichiga oladi; taqqoslanadigan korpus: bir nechta tillardagi oʻxshash matnlarni oʻz ichiga oladi, bu ularni taqqoslash imkonini beradi.

Korpusshunoslik rivojlanib borish jarayonida juda koʻp yutuqlarga erishdi. Hozirgi kungacha boʻlgan davrda elektron korpuslar quyidagi bosqichlarni bosib oʻtgan [Abdurahmonova 2021; 286]:

1. 1960-yillar: Brown korpusi AQSh, 1 mln soʻz va 500ga yaqin namuna.
2. 1970-yillar: LOB – Lancaster – Oslo/Bergan korpusi, Buyuk Britaniya, Norvegiya 1 mln soʻz.
3. Rus tilining mashina fondi, Upsala rus tilining korpusi (Shvetsiya), 1 mln soʻz.
4. 1990-yillar: British National Corpus milliy korpus 100mln soʻz (venger, italyan, xorvat, chesh, yapon tili).
5. 1990-yillar: Penn korpusi. Sintaktik razmetkaga asoslangan korpus, 4,5 mln soʻz.
6. Ingliz tili banki Birmingham (Collins Cobuild) 600 mln soʻz.
7. 2000-yillar: Amerika milliy korpusi, 100 mln soʻz.
8. Zamonaviy Amerika ingliz tili korpusi, 400 mln soʻz.
9. Rus tili milliy korpusi, 140 mln soʻz.
10. Gigaword korpusi: arab, ingliz, xitoy tili 2 mlrd soʻz.
11. Oxford English korpusi, 2 mlrd soʻz.
12. Portugaliya va Galitsiyatillari uchun CRPC, NILC, CORGA.
13. Ispan tili uchun CREA.

14. Qozoq tili uchun Kazakh-KTB.

15. Tatar tilining "Tugan Tel" korpusi, 180 mln so'z.

Turkiy tillar elektron korpuslari: Turk tilining milliy korpusi, 40 mln so'z; Qozoq tilining Almaty Corpus (NCKL) 40 mln so'z; Oltoy tilining korpusi; Boshqird tilining milliy korpusi; Xakas tilining korpusi; Tatar tilining yozma matnlar korpusi; Yoqut tilining korpusi; Sibir ozchilik tillarning raqamli korpusi. Bundan tashqari Turk tili bo'yicha 423 mln so'z, 491 tokendan iborat BOUN korpusi, 2 mln so'zdan iborat METU, 50 mln so'zdan iborat veb korpus 34 Mbdan iborat tillar doirasida yaratilgan turk tilli parallel korpuslari mavjud [Abdurahmonova 2021; 21]. O'zbek tili milliy korpusi (2020-2022), O'zbek tili korpusi (2021), O'zbek tilining ta'limiy korpusi (2021-2022)⁷. Tadqiqotlardan ko'rinadiki ko'plab turkiy tillar korpusi yaratilgan. Dunyo tilshunosligida korpusdan foydalanilgan asosiy tadqiqotlar deb Frensis [Abdurahmonova 2021] quyidagilarni ko'rsatadi: leksikografik tadqiqotlar, dialektologik tadqiqotlar va til grammatikasi ustida olib borilgan tadqiqotlar.

Bugungi kunda yaratilgan til korpuslari yuzasidan qilingan tadqiqot ishlarini kuzatganimizda ularning tarkibida dialektologik tadqiqotlar ham o'rin olganligini ko'rishimiz mumkin. Birinchi navbatda dialektal korpuslar haqida to'xtalsak. Dialektal korpuslar aynan bir til yuzasidan qilingan til korpusi va Milliy til korpuslari tarkibidagi dialektal korpus bo'limlaridan iborat. MAROKO: *Moldaviya va Ruminya dialektal korpusi*, Arab tilining ko'p tilli dialektal korpusi, Nordik dialektal korpusi, Zamonaviy Katalon dialektal korpusi, Ispan tilining og'zaki *va ovozli korpusi* (COSER), *Xelsinki Britaniya ingliz dialektlarining korpusi (HD) Bolgariya dialektal korpusi, Portugaliya dialektal korpusi, Mandarin Xitoy-materik shevasi (CALLFRIEND), Albaniya milliy korpusi, Arab dialekti parallel korpusi (PADIC), Rus milliy korpusi (dialektologiya bo'limi), Tatar dialektal korpusi, Kuban dialekti madaniyatining elektron korpusi, Saratov dialektal korpusi, Janubiy Gollandiya lahjalarining tahlil qilingan korpusi (GCND)*⁸ va boshqalar. Yuqorida sanalgan dialektal korpuslar aholi yashash joylaridan to'plangan dala ma'lumotlari, telefon so'zlashuvlari, internet ma'lumotlari, xalq og'zaki ijodi namunalari, tarixiy, diniy, badiiy asarlar, televideniye va radio eshittirish, intervyular va talabalarning ilmiy tadqiqot ishlaridagi ma'lumotlar asosida tuzilgan, transkripsiyalangan va izohli yozuvlaridan iborat lingvistik korpusdir. Dialektal korpuslar mundarijasida tilshunoslar va kompyuter dasturchilari bilan kelishilgan holda quyidagi sahifalar o'rin olgan:

Kirish;	Boshqa korpuslarga yo'nalma;
Baza tizimi;	Tarix va fon;
Tahrir qoidalari;	Tadqiqotlar;
Ma'lumotlar bazasi;	Video va audio yozuvlar;

Dialektologik korpuslarning maqsadi dialektologiya, sotsiolingvistika, nutq madaniyati, fonetika-fonologiya, leksikologiya, grammatikaga oid tadqiqotlar hamda mahalliy aholining urf-odatlarini, tarixi va boshqa ko'plab sohalar uchun tadqiqotlar maydonini qulaylashtirish, kelajak avlodga qimmatli ma'lumotlarni yetkazish vazifasini bajarishdir.

Dialektologik korpuslarni tuzish ahamiyati shundaki, dunyoda tan olingan tillardan tashqari bir necha 10 ming so'zlovchiga ega bo'lsada iqtisodiy rivojlanishning pastligi sababli yo'qolib borayotgan til va ularning shevasini saqlab qolish hamda tadqiq qilishdir. Dialektlarni aniqlash va tasniflashning turli usullari mavjud bo'lib, bugungi kunda dunyo aholisi 7000 ga yaqin

⁷ <https://ila.uca.es/linguistica-de-corpus/?lang=en>; <https://www.tnc.org.tr/tr/derlem-hakkinda/amac/>; http://web-corpora.net/KazakhCorpus/search/?interface_language=kz; <https://github.com/nevmenandr/bashkir-language-resources>; <https://khakas.altai.ru/dictionary/>; <https://www.corpus.tatar/>; <http://adictsakha.nsu.ru/corpora/corp/>; <https://corpora.iea.ras.ru/corpora/index.php>; <http://uzbekcorpora.uz/>

⁸ <https://www.ghentcdh.ugent.be/projects/corpus-southern-dutch-dialects-gcnd>

tillarda so'zlashayotgani taxmin qilinmoqda⁹. Ushbu tillarning har biri ijtimoiy guruh yoki boshqa omillarga qarab farq qilishi mumkin bo'lgan bir nechta dialekt (sheva)lar bo'lishi mumkin.

Dunyoda qancha dialekt borligini aniq aytish qiyin. Bu qiyinchilik qisman tilshunoslarning dialekt nimadan iborat ekanligi haqidagi kelishmovchiligidan kelib chiqadi – Ba'zi tilshunoslar tilning ikki shaklini turli dialekt deb bilishadi, boshqalari esa bunday fikrda emas. Shu o'rinda aytish joizki, dunyodagi dialektlarning miqdori juda katta bo'lib ana shu taxmin qilingan 7000 tildan faqat bittasi bo'lgan ingliz tilida 160ga yaqin lahjalar borligi taxmin qilinadi. Ba'zi ma'lumotlarga ko'ra, dialektlarning umumiy soni 10 000 dan ortiq. Ammo shuni ta'kidlash kerakki, bu lahjalarning aksariyati rasman tan olinmagan yoki ro'yxatga olinmagan. Bunga misol sifatida Hindiston davlatini olishimiz mumkin. 2001-yildagi aholini hisoblash davrida Hindistonda 1 mln gacha so'zlashuvchiga ega bo'lgan 30ta til, 10 000 nafargacha so'zlashuvchiga ega bo'lgan 122ta til, 1599ta kam sonli so'zlashuvchiga ega tillar borligi qayd qilinadi. 2011-yildagi hisob kitoblarda esa 121ta til borligi aniqlanadi va shu tillardan 22tasi Hindiston konstitutsiyasida til sifatida rasman tan olinadi¹⁰. Dialekt va lahjalarning miqdorini aniqlashning imkoni yo'qligi birgina hind tili misolida ham ko'rishimiz mumkin. Turkiy tillarda mavjud lahja va dialektlar ham shular sirasiga kiradi, chunki aholi migratsiyasi, iqtisodiy va ijtimoiy omillar bu farqlarni o'sishida hamda tillarni yo'qolishiga sabab bo'ladi. Korpus dialektologiyasining ikkinchi bir ahamiyati shu yo'qolish arafasidagi til(sheva)larni saqlab qolishdan iboratdir.

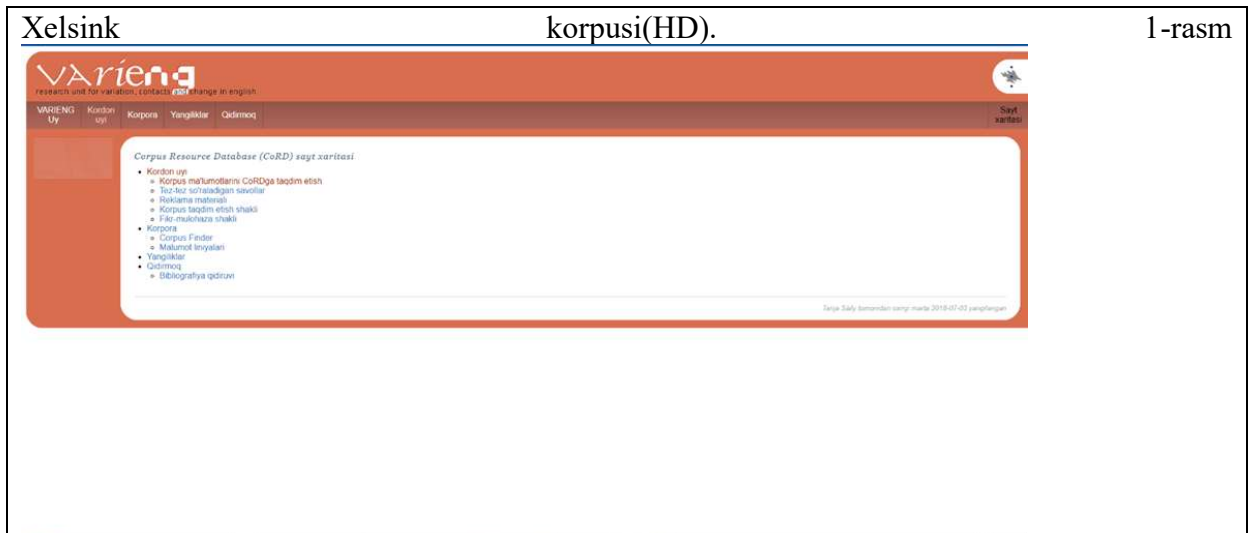
Jahon tilshunosligida yaratilgan dialektal korpuslarning yuzaga kelishiga nazar tashlaydigan bo'lsak: Xelsinki Britaniya ingliz dialekti korpusi (HD) Sharqiy Angliya, Janubiy-g'arbiy Lankashir orollari aholisidan to'plangan orfografik transkripsiyalangan audio yozuvlar to'plamini uchratishimiz mumkin. Yozuvlar 1970-1980-yillardagi Finlyandiya aspirantlari tomonidan qilingan. Hajmi: 1.008.641 so'zdan iborat jami 187ta fayl; Vaqt davrlari: 1970-1980-yillar (CAM, DEV, ELY, SOM, SUF), 1980-yillar (ESS, LAN). Birinchi bosqich 2006-yilda yakunlangan, tadqiqot uchun korpus materiallari biriktirilgan. Ikkinchi bosqich davom etmoqda, korpus ma'lumotlar bazasini ilgari nashr etilmagan korpus ma'lumotlar bilan kengaytirish davom etmoqda. Korpusning asosiy ma'lumotlar ba'zasi shevalardan yozib olingan audio yozuvlardir. Korpusning tuzilishi (1-rasm):

Kirish;
Asosiy tuzilma;
Konventsionalarni tahrirlash;
Ma'lumotlarni yig'ish;
Kembrijshir korpusi;
Devon korpusi;
Essex va Lancashire korporatsiyasi;

Eli(Ely) oroli korpusi;
Somerset korpusi;
Suffolk korpusi;
Korpus haqida;
Tadqiqot;
Video va audio namunalari;

⁹ <https://www.quora.com/How-many-languages-and-dialects-are-there>; <https://aclanthology.org/P19-1068/>; <https://aclanthology.org/W14-3601/> qarang: <https://hal.science/hal-01261587>; <https://ethnolex.ru/> <https://varieng.helsinki.fi/CoRD/corpora/Dialects/>; <https://www.corpusdoportugues.org/now/>; <http://www.tekstlab.uio.no/scandiasyn/>; <https://www.ub.edu/ccub/variacion-dialectal-cod/>; <http://www.corpusrural.es/descripcion.php>; <https://catalog.ldc.upenn.edu/LDC96S55>; <http://albanian.web-corpora.net/>; <https://smart.loria.fr/corpora/>; <https://ruscorpora.ru/en/search?search=CgQyAggIMAE%3D>; <https://dialekt.corpus.tatar.dialect/link.html>; <https://www.translateday.com/how-many-dialects-does-english-have/>

¹⁰ https://en.wikipedia.org/wiki/Languages_of_India; <https://varieng.helsinki.fi/CoRD/corpora/Dialects/> qarang: https://en.wikipedia.org/wiki/Eighth_Schedule_to_the_Constitution_of_India; <https://varieng.helsinki.fi/CoRD/sitemap/index.html>; <https://ruscorpora.ru/en/corpora/dialect?search=CgQyAggIMAE=>

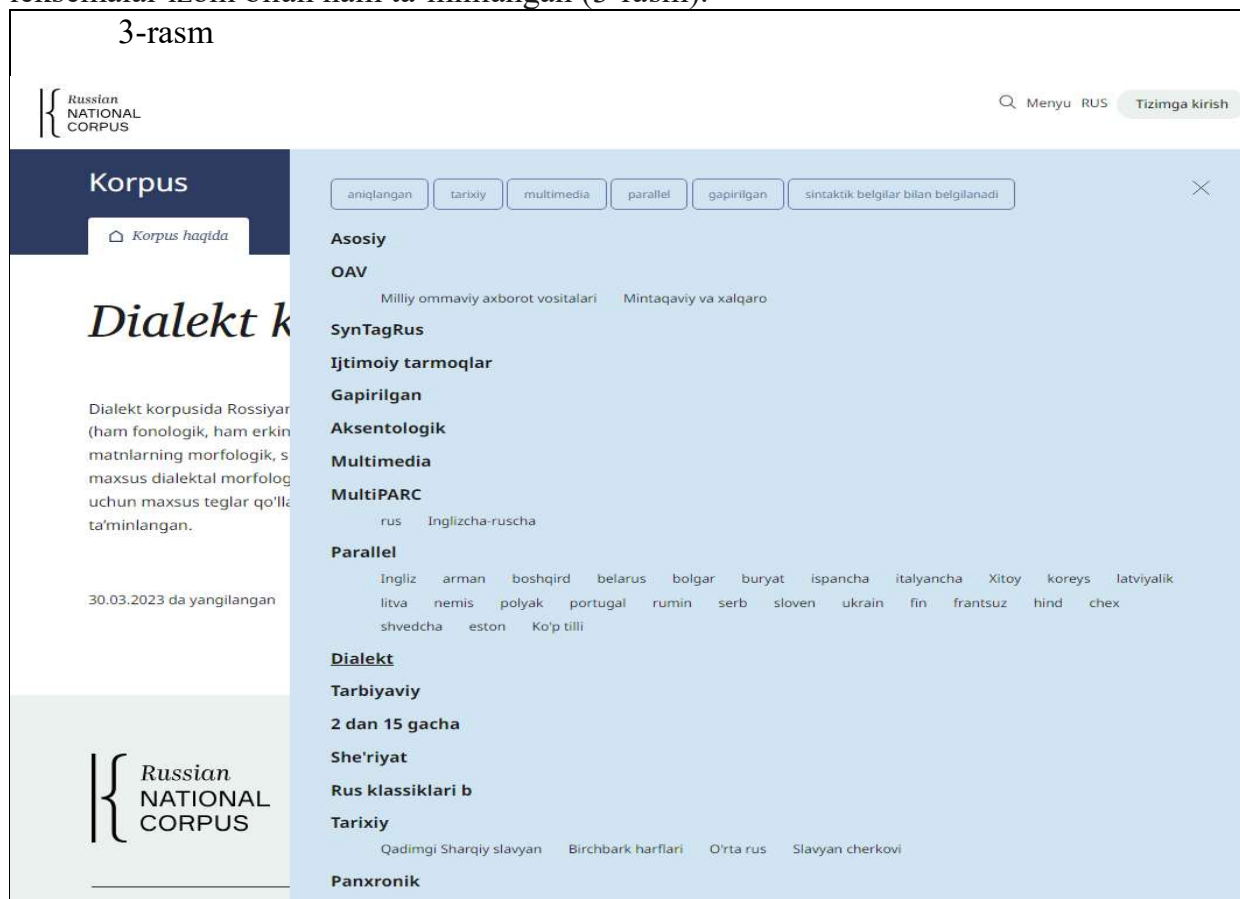


Tadqiqotda ma'lumot beruvchi shaxslar maktabda yaxshi o'qimagan, o'rta ma'lumot egasi, qishloqdan chetga chiqmagan shaxslar bo'lishi kerakligi mezon sifatida qabul qilingan. Shundagina ma'lumot beruvchi o'z qiymatini yo'qotmagan, adabiy til ta'siriga uchramagan manbani taqdim qila oladi, deb hisoblaydi korpus mualliflari.

PADIC - parallel arab dialekti korpusi Mag'rib va Yaqin Sharq dialektlarining taxminan 6400 ta jumlasidan iborat. Har bir dialekt zamonaviy standart arab tili (MSA) bilan moslashtirilgan. PADIC Mag'ribdan to'rtta dialektni o'z ichiga oladi; bular: Jazoir, Tunis, Marokash, Yaqin Sharq (Suriya va Falastindan). PADIC noldan boshlab SMarT tadqiqoti a'zolari bo'lgan Salima Xarrat, Karima Meftouh va K.Smayli va M.Abbos ishtirokida yaratilgan korpusdir. Korpus arab tilining turli lahjalaridan yig'ilgan ijtimoiy tarmoqdagi audio va videolar jamlanmasidan tuzilgan. Bundan tashqari arab milliy musiqasi hamda adabiy til elementlaridan ham foydalanilgan. Ushbu modellashtirish tarjima va nutqni aniqlash uchun ikki yoki undan ortiq tillar o'rtasidagi assotsiatsiyalarni aniqlash, ajratib olish va taklif qilish uchun matematik usullardan foydalanishni o'z ichiga oladi. Tillar kam manbali tillar (arab dialektlari), arab-fransuz va arab-ingliz tillari uchun bir tilli, parallel yoki taqqoslanadigan korpuslar orqali o'rganiladi. Dastur boshida kam sonli lahjalar ustida ish olib borilgan bo'lib, korpus asosi yaratilgandan keyin umumiy yechim va izohlarga ega 25ta arab tili (lahja) dan tashkil topgan korpus tuzilgan (2-rasm).



Rossiya milliy korpusi. Dialektal korpusda Rossiyaning turli mintaqalarida yashovchi xalqlarning til va sheva xususiyatlari bilan yozib olingan audio matnlar, xalq og'zaki ijodi namunalari turli davrga oid dialektal yozuvlar (erkin standartlashtirilgan orfografiyada taqdim etilgan) mavjud. Bu matnlarning morfologik, sintaktik va leksik xususiyatlari saqlanib qolgan. Subkorpusda maxsus dialektal morfologik belgilar (standart tilda mavjud bo'lmaganlar) uchun maxsus teglar qo'llaniladi. Bundan tashqari sof dialektal leksemalar izohi bilan ham ta'minlangan (3-rasm).



Dialektal korpuslarni tuzishda sheva vakillarining kelib chiqishi, ism sharifi, yashash joyi tarixi, kasbi, ma'lumoti haqidagi ma'lumotlar bilan to'ldirilganligini ko'rishimiz mumkin. Barcha yaratilgan dialektal korpuslarda shu tartib aks etgan.

Korpus lingvistikasining rivojlanishi o'zbek tilshunosligida yangi yo'nalishlardan bo'lib, korpus va uning tuzish tamoyillari to'g'risida qator tadqiqotlar olib borilgan. Bunga misol sifatida O'zbek tilining milliy korpusi, O'zbek tili ta'limiy korpusi, O'zbek tili korpusi, O'zbek milliy shevalari korpusi va ular tarkibidan o'rin olgan mualliflik va parallel korpuslar shular sirasidandir. O'zbek tilshunosligida korpusshunoslikni o'rganish va tadqiq qilish 2018-yildan e'tiboran keng rivojlandi. Bu sohaning rivojlanishida S.Rizayev, S.Muhamedov, A.Po'latov, S.Karimov, B.Mengliyev, L.Raupova, Z.Xolmanova, S.Muhammedova, N.Abdurahmonova, Sh. Hamroyeva, B.Elov,

M.Abjalova¹¹, O.Abdullayeva¹², O'.Xoliyorov, D.Ahmedova, D.Elova va boshqa ko'plab tilshunos va dasturchilarning xizmatlari salmoqli [Холиёров 2021, Ахмедова 2020, Элова 2022]. Bizning tadqiqotimiz ham korpusshunoslik an'anasini davom ettiradi. Til korpuslarida Surxondaryo viloyati Denov tuman shevasi ma'lumotlar ba'zasini shakllantirishda yuqorida sanalgan tadqiqotlar va tuzilgan korpus ma'lumotlaridan foydalanishga harakat qildik. O'z o'rnida ta'kidlashimiz kerakki, "O'zbek milliy shevalar korpusi" bu o'zbek shevalari korpusi uchun debocha vazifasini bajaradi. Tadqiqotchi M.Xolova Boysun tuman shevasini tadqiq qilib (<https://dialect.uz/>) dialektal korpusni tuzishga muvaffaq bo'lgan. Tadqiqotchi dialektologik korpus tuzishni ikki bosqichda amalga oshirishni ko'rsatadi: 1-bosqichda korpusning texnik va ma'lumotlar bazasini shakllantirish, bunda dialektal korpusning pasporti, mavzuviy guruhlari, janr va matnlar bo'yicha metarazmetkasini tuzish; 2-bosqichda to'plangan ma'lumotlar asosida fonetik va morfologik metarazmetkasini ishlab chiqish [Холова 2022]. Tadqiqotchining "Onlayn shevalar bazasi"da 10ta maydon orqali qidiruv amalga oshirish imkoniyati mavjud bo'lib, adabiy til, transliteratsiya, transkripsiya, turkum, dialekt izohi, dialekt qatnashgan gap tarkibi, sinonim, omonim, antonim, o'zlashgan qatlam, rus va ingliz tilidagi tarjimalari kabilarni ichiga olgan 5 sahifadan iborat (4-rasm):

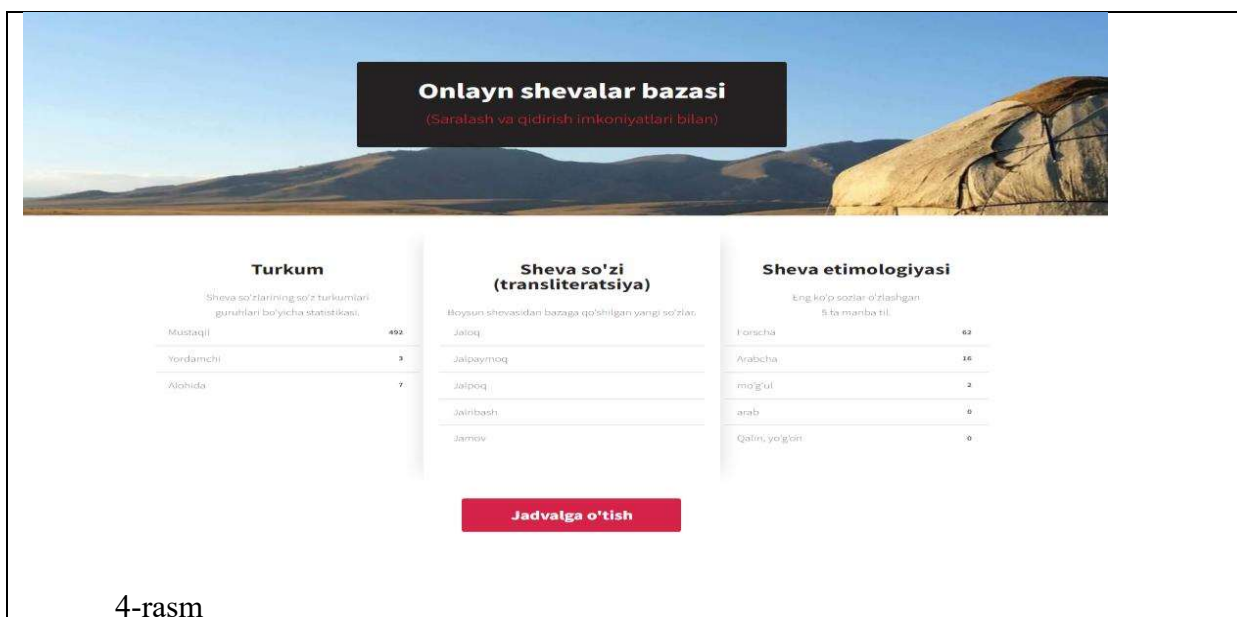
Bosh sahifa;

So'zlar jadvali;

Loyiha haqida;

Transkripsiya;

Aloqa;



4-rasm

Shevalar korpusini yaratish bosqichida shevashunosning tadqiqot maydoni doirasida matnlarni fonetik, leksik, grammatik jihatdan razmetkalash bosqichi dastlabki

¹¹ https://www.researchgate.net/publication/372447469_Monografiya_PhD_Abjalova_Manzura;
<http://uzschoolcorpara.uz/> ; <http://uzschoolcorpara.uz/>

<https://dialect.uz/>; <http://uzbekcorpara.uz/>; <http://uzschoolcorpara.uz/>

¹² [file:///C:/Users/admin/Desktop/Abdullayeva%20O.%20Dissertatsiya_last%20\(1\).pdf](file:///C:/Users/admin/Desktop/Abdullayeva%20O.%20Dissertatsiya_last%20(1).pdf)

vazifa hisoblanadi. Razmetkalash deganda matn va uning komponentlariga maxsus teglarni biriktirib chiqish nazarda tutiladi. Maxsus teglar ikki xil bo'ladi: lingvistik teglar va ekstralingvistik (tashqi) teglar. Lingvistik teglar matn elementlarining leksik, grammatik va shunga o'xshash boshqa xususiyatlarini tavsiflaydigan ma'lumotlardan iboratdir. Ekstralingvistik teglar esa muallif haqidagi va matn haqidagi ma'lumotlarni (muallif, nomlanishi, nashr yili va joyi, janr, tematika va hakoza) tavsiflaydi¹³. Yangi tuzilgan korpuslar uchun razmetkalash qo'l mehnati asosida bajariladigan ish hisoblanadi. Razmetkalash uch yo'nalishda olib boriladi: kuzatish, fonetik razmetkalash, shevaga oid (dialektal) matnlarni razmetkalash. Kuzatish bosqichida matn olinadigan shahar, qishloq, aholining etnik tarkibi, respondentning kelib chiqishi, kasbi, ma'lumoti (o'rta, oliy) kabilar o'rin olsa, fonetik razmetkalashda unli (old qator, orqa qatorlik, singarmonizm, urg'u olish, olmaslik), undosh (portlovchi-sirg'aluvchi, jarangli-jarangsiz) tovushlar doirasidagi o'zgarishlar; dialektal matnlarni razmetkalash¹⁴da sheva materiallarining leksik metarazmetkasi ham muhim hisoblanib har bir shevadagi so'zlar alohida teglanib adabiy til bilan bog'lanib chiqiladi, dastlabki dialektal korpuslarda har qanday tildagi korpuslarda bo'lganidek har bir so'z (variantlari) sinonimi, omonimi, antonimi, paronimlari alohida teglanadi. Shevalar o'rtasidagi so'zlarni leksik-semantik teglash murakkab jarayon hisoblanadi. Grammatik razmetkalash yagona standartlashtirilgan grammatik birliklar mavjudligi shevalararo bog'liqliklar sabab ozmi ko'pmi bir-birini tushunishga olib keladi va bu jarayon yuqoridagi uch bosqishdan yengilroq sanaladi. Bundan tashqari, avtomatik teglash (razmetkalash) ham mavjud bo'lib bunda teglashni "tegger (teglovchi dastur)" bajaradi [Abdullayeva 2022; 103]. Yuqorida sanalgan korpuslarda qisman razmetkalash amalga oshirilganligi bizning ishimizda ham o'z o'rnida foydalanildi. Til korpuslarida sheva ma'lumotlar bazasi shakllantirildi. Bunda avval Denov shevalari haqidagi ma'lumotlar jamlanib tahlil qilindi. Denov tumanida o'zbek shevalarining qipchoq, qarluiq lahjasida so'zlashuvchi aholi va tojik tilida so'zlashuvchi aholi vakillari yashaydi. Denov tumanida joylashgan 98ta (MFY) aholi yashash joylarini kuzatib, yashaydigan odamlar bilan suhbatlashish chog'ida yig'ilgan audio matnlar tahlil qilindi. To'plangan dialektologik materiallarni 2020-yilda yaratilgan "O'zbek tili korpusi"¹⁵da joylashtirish metodologiyasi ishlab chiqildi hamda dialektologik materiallar korpusning dialektologik subkorpusiga joylashtirildi. uzbekcorpus.uz dasturi N.Abduraxmonovning O'zbek tili elektron korpusining kompyuter modellari (Fil.f.d.DSc. 2021.) dissertatsiyasi doirasida ishlab chiqilgan. Korpusda o'zbek tilida ilk bor *FST* (finite state transducer) – chekli avtomat transyuteri vositasida korpusni morfologik teglashning avtomatik usuli yordamida lemmatizatsiyalash va tokenizatsiyalash jarayonlarining lingvistik tahlil bosqichlari ishlab chiqilgan morfologik ma'lumotlar bazasi hamda morfotaktik qoidalar tizimi shakllantirilgan.

Korpusni sintaktik teglash va annotatsiyalashda universal tobelik nazariyasining UdPipe metodi orqali *CONLLU* format namunasi o'zbek tilining sintaktik modellari aniqlangan.

¹³ <http://compling.navoiy-uni.uz/index.php/conferences2/article/view/167/160>

¹⁴ Matnlarni rametkalashda: nasriy matn, she'riy matn va aralash (she'riy-nasriy) matnlar ko'proq tadqiq qilingan bo'lib Dialektal korpusda tabiiy va ishonchligini ta'minlash uchun dialog va monolog (audio) matnlarni tanlab oldik va shularni izohladik.

¹⁵ <https://uzbekcorpus.uz/newIndex>

Olib borilgan izlanishlarning amaliy ifodasi sifatida <http://uzbekcorpus.uz> sayti yaratilgan bo'lib, o'zbek tili korpus menejeri (qidiruv tizimi)ning lemma, token bo'yicha hamda *n-gram* modeli asosida konkordanslarning qidiruv tizimi ishlab chiqilgan.

Tarjima qilish algoritmi va o'zbek tili parallel korpusining tarjima xotirasini yaratishda *WordFast* texnologiyasidan foydalanilgan hamda o'zbek tilining ingliz va rus tillarida tarjima birliklari muqobillarining ma'lumotlar bazasi tuzilgan.

Korpusni grammatik teglash bosqichida *Protege* dasturi orqali ontologik modellashtirish usulining ilmiy asoslari belgilangan. Korpusda qidirilayotgan so'zning imloviy xatolarini aniqlash va unga yaqin bo'lgan so'zlar koeffitsientini belgilashda qidiruv interfeysida ilova qilishda Djaro *Winkler* algoritmini qo'llashning maqsadga muvofiq ekanligi belgilangan.

Xulosa o'rnida shuni aytish mumkinki, Denov shevasi nisbatan murakkab arealni hosil qilganligi bilan xarakterlanadi va bu arealga xos xususiyatlarni korpusga joylashtirish o'zbek shevalari korpusining mukammalligini ta'minlashga xizmat qiladi. Denov shevalarini ham mavjud dialektal korpuslar sahifalari asosida joylashtirish mumkin. Korpusni sintaktik teglash va annotatsiyalashda universal tobelik nazariyasi metodologiyasi asosida o'zbek tilining sintaktik modellari aniqlanganligini hisobga olish zarur. Shuningdek, dialektal zona materiallari asosida subkorpus ma'lumotlar bazasi shakllantirilib, qidiruv oynasida natijalarni aks ettirilishi tajribalarda isbotlandi.

Foydalanilgan adabiyotlar

1. Turdialiyev A. Denov shevasini areal o'rganish va til korpusiga joylashtirish. – Toshkent, 2025.
2. Ashirboyev S. O'zbek dialektologiyasi. – Toshkent, 2021.
3. Abdurahmonova N. O'zbek tilini elektron korpusining kompyuter modellari. – Toshkent. 2021. – 199 b.
4. Khidralieva Z. R. Vowel harmony features in ikan dialect. // *ACADEMICIA: An International Multidisciplinary Research Journal*, Vol. 10, Issue 12, December 2020. pp. 1177-1182.

