JAHON KORPUS LINGVISTIKASIDA DIALEKTAL KORPUSLAR

Vosiljonov Azizbek
Farg‘ona davlat universiteti
Ijtimoiy-gumanitar fanlar kafedrasi o‘qituvchisi.

Annotatsiya: Ushbu maqolada Britaniya ingliz dialektlarining Xelsinki korpusi (HD) hamda Chexiya milliy korpusining chex dialektlari korpusi misolida dialekt korpuslar konsepsiyasi zamonaviy til fanining yutuqlari va dialektologlar uchun mavjud bo‘lgan materiallarning tabiati bilan bog‘liq bo‘lgan xalq nutqini o‘rganish bo‘yicha maktabning asosiy yo‘nalishlarini hisobga olgan holda ishlab chiqilganligi yoritilgan.
Kalit so‘zlar: dialektologiya, dialekt korpus, intervyu, axborotchilar, geografik joylashuv, SED, yunikod.

Annotation: In this paper, using the examples of the Helsinki Corpus of British English Dialects (HD) and the Czech National Corpus of the Czech Dialects, the concept of dialect corpora relates to the progress of modern linguistics and the nature of the materials available to dialectologists in the school of vernacular studies. It is explained that it was developed taking into account the main directions.
Keywords: dialectology, dialect corpus, interview, informants, geographic location, SED, unicode.

Аннотация: В данной статье на примере Хельсинского корпуса британских английских диалектов (HD) и Чешского национального корпуса чешских диалектов концепция диалектных корпусов представляет собой исследование народной речи в связи с достижениями современной лингвистики и характер материала, доступного диалектологам, поясняется, что он был разработан с учетом основных направлений школы.
Ключевые слова: диалектология, диалектный корпус, интервью, информаторы, географическое положение, СЭД, юникод.

Milliy til va uning shevalarini informatsion dunyoga olib chiqish, uni elektronlashtirish nafaqat tilshunoslik, balki shevashunos tadqiqotchilar oldida turgan muhim masaladir. Muloqotning turli sohalarida milliy tillarning ishlashini ifodalovchi korpuslar bilan bir qatorda, milliy til doirasida ajralib turadigan alohida lingvistik jamoalardagi muloqotni taqlid qiluvchi korpuslar ham kerak. Bu tipdagi eng muhim til shakllanishlari shevalardir. Dialekt matnlarining mashina fondini yaratish zarurligi haqidagi g‘oyalar dastavval 1980-yillarda A.S.Gerd va V.E.Goldin tomonidan ilgari surildi[Goldin, 1990; 17]. Hozirgi vaqtda dunyo tilshunosligida dialektal nutqning alohida elementlarini ifodalovchi bir qator korpuslar mavjud: dialektal matnlarning xorijiy korpusi (masalan, ingliz dialektlarining Xelsinki korpusi, Kirkning Shimoliy Irlandiya transkripsiyalangan nutq korpusi (NITCS), IViE (ingliz tilidagi intonatsion o‘zgarishlar) korpusi, BBC Voices); rus tili milliy korpusining (NCRL) dialekt subkorpusi; Pustosha – Moskva viloyati, Shaturskiy tumani, shu jumladan matnlar – dialektdagi ma'ruzachilar nutqining namunalari shular jumlasidandir.

Britaniya ingliz dialektlarining Xelsinki korpusi (HD) asosan Sharqiy Angliya va Janubi-G‘arbiy hududlardan, Lankashirdan kichik to‘plamga ega bo‘lgan orfografik tarzda yozib olingan audio yozuvlar to‘plamidir.

Korpusning maqsadi dialektologiya, sotsiolingvistika, nutq tahlili, morfologiya, sintaksis va fonologiya sohalarida lingvistik tadqiqotlar uchun material berishdir. Korpus shuningdek, aloqa etnografiyasi, mahalliy urf-odatlar va tarix kabi tilga oid bo‘lmagan, ko‘p tarmoqli tadqiqotlar uchun material beradi. Korpusning asosiy ma’lumotlari og‘zaki dialekt nutqining audio yozuvlaridir Hajmi 1 008 641 so‘zdan iborat bo‘lib, jami 187 ta fayldan tashkil topgan. Birinchi bosqich 2006-yilda yakunlandi, ikkinchi bosqichi esa korpus materiallarini ilgari nashr etilmagan korpus ma’lumotlari orqali kengaytirish rejalari bilan davom etmoqda [Abdurakhmonova, N. 2021,2022]. Yozuvlar 1970 va 1980-yillarda Finlyandiya aspirantlari tomonidan qilingan, bu ishda professor Garold Orton va boshqa Lids olimlari maslahat bergan. Korpusda og‘zaki dialekt nutqining orfografik transkripsiyalari mavjud.

Namuna olish jarayonining tamoyillari ma’lum darajada mahalliy va ma’lumot beruvchilarni tanlash uchun *Lids ingliz dialektlari tadqiqoti (SED)* mezonlariga asoslangan. Ma’lumot to‘plash usuli sifatida, asosan, so‘rovnomadan foydalanish, ma’lumot beruvchining fikrlash oqimini to‘xtatib qo‘yish va nutqning uzoq cho‘zilgan qismini yozib olish usullari orqali ifoda etilgan. Ma’lumot beruvchilarning asosiy qismi qishloq aholisining “dala ishchilari” hisoblanadi va ularning yosh va jins kategoriyalarida ham turli xil farqlanishlar mavjud. Jumladan, boshqa dialektal korpuslardan farqli ravishda erkak informatorlar ayol informatorlarga nisbatan ko‘pchilikni tashkil qiladi. Yosh munosabati esa, asosan, nafaqa yoshidagi shaxslar hisoblanadi. Ma’lumot to‘plash birinchi navbatda tekshiriladigan joylar tarmog‘ini o‘rnatish va keyin ushbu joylarda tegishli ma’lumot beruvchilarni topishdan boshlanadi. Qishloq pochta bo‘limlari ma’lumot beruvchilarni topishda katta yordam beradi, chunki ular potentsial ma’lumot beruvchi bo‘lgan nafaqaxo‘rlar ro‘yxatini saqlaydilar. Do‘konlar ham yaxshi ma’lumot manbai sanaladi, chunki do‘kondorlar odatda qishloq aholisini yaxshi bilishadi. Bundan tashqari, qishloq hayotiga qiziquvchi mahalliy aholi bilan uchrashish ma’lumot to‘plashda katta yordam berdi, chunki ular dala ishchisiga potentsial ma’lumot beruvchilar haqida qo‘shimcha ma’lumot berishlari va hatto o‘zlari ham suhbatdosh sifatida yordam berishlari mumkin[Abdurakhmonova, N. 2021,2022]. Ular axborot beruvchilar nutqidagi notanish dialekt shakllarini farqlashda va izohlab berishda yordam bergan.

Har bir intervyu Britaniya ingliz dialektlarining Xelsinki korpusi (HD) da moslashtirilgan beshta parametr kodlari bilan ta’minlangan . Ular quyidagilar:

B: HD fayl kodi (<B DICAM13>)

N: tuman, qishloq, ma’lumot beruvchi ma’lumoti (|N_CAM_LANDBEACH_SJ)

Y: ma’lumot beruvchi(lar)ning yoshi(lar)i (<Y 86>)

X: ma’lumot beruvchi(lar)ning jinsi (<X WOMEN>)

H: ma’lumot beruvchi(lar)ning kasbi(lar)i (<H HOUSEWIFE>)

Asosiy parametr kodlaridan keyin quyidagi tafsilotlarni ham ko‘rish mumkin:

- raqamli arxiv kodi (Dig. CAM32A)
- tasma uzunligi daqiqalarda (47:37 min)
- dala ishchisining bosh harflari bilan yozilgan yil yoki sana (Rec 1974, AO tomonidan)
- transkriptorning bosh harflari (AO)
- jami va olib tashlangan so‘zlar soni (bizning WC 6,075, tozalangan WC 5,796)

● arxiv sahifalari soni (41 arxiv sahifasi)
● yakuniy kompyuter qo‘lyozmasi sanasi (CMS 9.1.2001)

Umuman olganda, Britaniya ingliz dialektlarining Xelsinki korpusi (HD) Britaniya ingliz tili dialektlarini ilmiy, ta’limiy maqsadlarda o‘rganishda maqbul qo‘llanma hisoblanadi.

Chexiya milliy korpusining chex dialektlari korpusi. 2017-yilda Chexiya Milliy Korpusi instituti o‘zining “Og‘zaki korpus” bo‘limi tomonidan tayyorlangan Chexiya milliy korpusining yangi ixtisoslashtirilgan chex dialektlari korpusini taqdim etdi[Goláňová H. Waclawičová M. 2018; 79]. Korpus Chexiya Respublikasi hududida mavjud bo‘lgan an'anaviy mintaqaviy dialektlarni qamrab oladi va taqdim etadi. Uning birinchi ommaviy versiyasida korpusning o‘lchami taxminan 100 000 so‘zdan iborat, lekin ko‘proq ma’lumotlar doimiy ravishda to‘plangan va kengaytirilgan versiyalari kelajakda nashr etiladi.

Auditoriya nuqtai nazaridan, dialekt korpusi til mutaxassislari (dialektologlar, boshqa tilshunoslar va tegishli sohalardagi tadqiqotchilar), shuningdek, keng ommaning havaskorlarini qamrab olishni maqsad qilgan. Shuningdek, u barcha ta’lim darajalarida o‘quv manbasi bo‘lib xizmat qiladi. Dialekt korpusi Chexiya Respublikasining barcha an’anaviy dialekt hududlarida qilingan nutq yozuvlari va ularning transkriptlaridan iborat. Joriy versiyada umumiy uzunligi 13 soat bo‘lgan 324 ta yozuv mavjud bo‘lib, ular 178 ta noyob dinamikni o‘z ichiga oladi. Ma’ruzachilar soniga dialektologik yozuvlar nisbatan qisqa (taxminan 1–6 daqiqa, odatda 2 daqiqa) bo‘lishi ta’sir qiladi, chunki ular asosan bitta mavzu bo‘yicha bir kishining hisobiga qaratilgan. Hududiy qamrov nuqtai nazaridan korpusda Chexiya Respublikasining barcha dialekt mintaqalari, shu jumladan Polshadagi chex tilida so‘zlashuvchilarning yozuvlari mavjud. Hozircha u Chexiya, Moraviya va Sileziya chegaralaridan olingan yozuvlarni o‘z ichiga olmaydi, ammo ular an'anaviy dialekt mintaqalariga tegishli emas. Ikkinchi jahon urushidan keyin aholining ko‘p ko‘chirilishi sababli, bu asosan nemis tilida so‘zlashuvchi hududlarda an'anaviy dialekt substrati yo‘q[Abdurakhmonova, N. 2021,2022].

Korpusda [http://www.korpus.cz] to‘plangan yozuvlar turli manbalardan olingan, buning natijasida ular juda uzoq vaqtni qamrab oladi. Ular eski va yangi bo‘lgan ikki vaqt qatlamiga bo‘lingan va bu ma’lumotlardan qidiruvlarni yoki korpus bilan bog‘liq boshqa operatsiyalarni cheklash uchun foydalanish mumkin. Qadimgi qatlam 1950-yillarning oxiridan 1980-yillargacha yozilgan yozuvlardan iborat. Ushbu materialning bir qismi Chexiya Fanlar akademiyasining Chex tili institutining dialektologiya bo‘limi tomonidan to‘plangan va Chexiya lingvistik atlasiga qo‘shimcha sifatida nashr etilgan; qolganlari shaxsiy to‘plamlardan olingan bo‘lib, ular asosan ilgari nashr etilgan. Yangi qatlam esa 1990-yillardan hozirgi kungacha bo‘lgan yozuvlarni qamrab oladi. Ushbu yangi qatlam turli universitetlardagi tadqiqot faoliyati davomida, mustaqil shaxslar tomonidan, eng muhimi, Chexiya Milliy Korpusi institutining o‘zi tomonidan olib borilgan izlanishlarning yozuvlarini o‘z ichiga oladi.

Iloji boricha ko‘proq asl sheva xususiyatlarini yozib olish uchun ma’lumotchi sifatida eng qadimgi avlod vakillarigina ishtirok etgan. Axborotchilar qishloqdagi mahalliy aholi orasidan tanlab olingan, ular aholining o‘troq qatlamiga mansub bo‘lib, asosan shu yerda butun umrini o‘tkazgan, qishloq xo‘jaligi turmush tarziga yoki ma’lum bir hunarga bog‘langan. Ularning barchasi 60+ yosh guruhiga to‘g‘ri keladi, Yozib olish bilan bog‘liq metama’lumotlar yozib olish joyi haqida batafsil ma’lumotni o‘z ichiga oladi, masalan, mahalliy joy turi (shahar, qishloq) va uning o‘lchami, geografik joylashuvi - mamlakat, mintaqa (Bogemiya, Moraviya, Sileziya), dialekt guruhi *(skupina),* kichik guruh *(podskupina),* bo‘linish *(úsek)* va tur (tip) kabi. Aloqa

holatining qo‘shimcha xususiyatlariga yozuv manbai, yozib olish sanasi, vaqt qatlamining a’zoligi, asosiy mavzu, nutq turi (monolog, dialog va ularning kombinatsiyasi), ma’ruzachilar soni, va tadqiqotchining mavjudligiva hokazolar kiradi. Korpusda informatorning jinsi, yoshi, ma’lumoti, bolalik davridagi yashash joyi va eng uzoq yashash joyidan tortib, ma’ruzachining eng uzoq professional kasbigacha bo‘lgan bir qator metama’lumotlar kuzatib boriladi. Ushbu ma’lumotlarning barchasi korpus interfeysida ko‘rsatilishi va chastota statistikasini olish uchun ishlatilishi mumkin[Abdurakhmonova, N. 2021,2022]. Agar bizni shimoli-sharqiy Bogemiyadagi dialekt ovozlari darajasidagi hodisalar qiziqtirsa, biz matnning qidiruvni ***cheklash*** funksiyasidan foydalanib, boshqa hududlar qidiruvini cheklashimiz orqali biz istagan mintaqadagi dialektni topishimiz mumkin yoki ushbu dialekt mintaqasi asosida o‘z doimiy subkorpusimizni yaratishimiz mumkin. Matn tinish belgilari yozma chex tilining standart qoidalariga amal qiladi, lekin jumlalar bosh harf bilan boshlanmaydi.

Shunday qilib dialektal korpus yaratish jarayonida og‘zaki folklorlik xususiyatlariga ega bo‘lmagan matnlar, og‘zaki folklorlik xususiyatlariga ega bo‘lgan matnlar, yozma folklorlik xususiyatlariga ega bo‘lmagan matnlar va yozma folklorlik xususiyatlariga ega bo‘lgan matnlarning barchasini to‘laqonli kiritish dialektal bazaning shakllanishi uchun o‘ziga xos bosqich bo‘lib xizmat qiladi.

Foydalanilgan adabiyotlar:

1. Goldin. Mashina fondi, 1986-1990.
2. Абдурахмонова, Н., & Абдувахобов, Г. (2021). O ‘quv lug ‘atini tuzishning nazariy metodologik asoslari. *Международный журнал искусство слова*, *4*(6).
3. Abdurakhmonova, N. (2021). Formal-Functional Models of The Uzbek Electron Corpus. *ANGLISTICUM. Journal of the Association-Institute for English Language and American Studies*, *10*(8), 59-66.
4. Abdurakhmonova, N., Alisher, I., & Toirova, G. (2022, September). Applying Web Crawler Technologies for Compiling Parallel Corpora as one Stage of Natural Language Processing. In *2022 7th International Conference on Computer Science and Engineering (UBMK)* (pp. 73-75). IEEE.
5. Абдурахмонова, Н., & Абдувахобов, Г. (2021). O ‘QUV LUG ‘ATINI TUZISHNING NAZARIY METODOLOGIK ASOSLARI. *МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИСКУССТВО СЛОВА*, *4*(6).
6. Abdurakhmonova, N., Shakirovich, I. A., & O‘G‘Li, K. N. S. (2022). Morphological analyzer (morfoAnalyse) Python package for Turkic language. *Science and Education*, *3*(9), 146-156.
7. Mahmudov, M.Ə. Kompüter dilçiliyi / M.Ə. Mahmudov. – Bakı: Elm və təhsil, – 2013. – 352 s
8. Goláňová, H. Waclawičová, M. Co je v ČNK nového Ix (Zprávy z českého národního korpusu). Korpus – gramatika – axiologie, 2018 (17), pages 78–82
9. https://varieng.helsinki.fi/CoRD/corpora/Dialekts/
10. http://www.korpus.cz