

DIALEKTAL KORPUS METODOLOGIYALARINING QIYOSIY TAHLILIGA DOIR AYRIM QAYDLAR

Xolova Muyassar Abdulhakimovna

TerDU O'zbek tilshunosligi dotsenti,

filologiya fanlari falsafa doktori (PhD)

e-mail: xolovam@tersu.uz

Annotatsiya: *Ushbu tadqiqotda dunyoning ko'zga ko'ringan dialektal korpuslari: ingliz tili shevalari mobil ilovasi, NorDic korusi, Archimob korpusi, Curras arab sheva korpusi, Chex korpuslarini taqqosladik va ularning ustivorligi, cheklovi hamda qo'llanishiga oid qarashlar o'z ifodasini topgan*

Kalit so'zlar: *dialektal korpuslar, ingliz tili shevalari mobil ilovasi, NorDic korusi, Archimob korpusi, Curras arab sheva korpusi, Chex korpusi*

Dunyo amaliy lingvistikasida aksariyat leksikografik lug'atlar adabiy til va shevalarni birdan qamrab oladi. Masalan, Insburkdagi Leopold-Frensis universiteti tomonidan tayyorlangan loyihada onlayn leksikografik lug'at foydalanuvchiga oddiy (simple) va murakkab (advanced) formalarini taqdim etadi. Unda leksema va so'zlarni tadqiq etish izoh (definitions), derivatsiya, iqtiboslik, kompozitsion yasaliish, qo'shimcha fikr, birikma kombinatsiyalari, misollar va fraza ichida ko'rsatib o'tiladi hamda filtratsiyasi nafaqat adabiy til, shevalar maydoni doirasida ham faol (navbatchi) so'zlar, fonetik farqlilikka ega bo'lgan til birliklari, leksemaning etimologiyasi, oxirgi foydalanilgan manbaning betlari, umumiy manba, morfemikaga ajratish hamda vaqtlar oralig'i vositasida amalga oshiriladi¹. Biroq jahon tajribasida dialektal korpuslar turli xil maqsadlarni ko'zlagan holda tuzilgan. Masalan, ingliz lahjalari ilova korpusi (EDAC) korpus ma'lumotlari maxsus ishlab chiqilgan mobil ilova vositasida

¹ Xolova M.A. 2025. "O'zbek tili dialektal bazasini shakllantirish prognozi" O'zbekiston: til va madaniyat Folklorshunoslik. Shevashunoslik 1: 22 – 29. doi.10.52773/tsuull.uzlc.2025.1/ZAKB5324

kraudsorsing² metodologiyasi asosida to'planib, turli hududlardagi ona tili so'zlovchilariga laboratoriya sharoitidagi dialekt nutq namunalarini masofadan turib taqdim etish imkonini berdi. Korpus 47 000 dan ortiq viktorina ishtirokchilari va 3 500 dan ziyod audio yozuvlarni o'z ichiga olgan juda katta ishtirokchilar bazasiga ega bo'lib, bu statistik ishonchlilikni ta'minlaydi. Bundan tashqari, ma'lumotlar to'plami ishtirokchilarning harakatchanligi, yoshi, etnik kelib chiqishi, ta'lim darajasi va jinsi kabi keng qamrovli sotsiolingvistik metama'lumotlar bilan boyitilgan. Ommabop mobil ilova orqali amalga oshirilgan kraudsorsing metodologiyasi ishtirokchilar uchun to'siqlarni sezilarli darajada kamaytirdi. Shu bilan birga, smartfon orqali ma'lumot to'plash usuli bir qator metodologik cheklovlarni o'z ichiga oladi. Xususan, ilova orqali ishtirokchi tanlash jarayoni Buyuk Britaniya aholisining yosh, etnik tarkib va ta'lim darajasi bo'yicha haqiqiy taqsimotiga to'liq mos kelmaydi, bu esa tanlanma qiyshiqqligini (sampling bias) keltirib chiqaradi. Bundan tashqari, ilova asosidagi metodologiya ma'lum bir ijtimoiy qatlamni — ya'ni raqamli texnologiyalarga kirish imkoni mavjud va smartfondan faol foydalanadigan aholini — maqsadli qamrab oladi, bu esa natijalarning umumlashtirish qobiliyatini (generalisability) cheklashi mumkin. Ushbu korpus zamonaviy mintaqaviy o'zgaruvchanlikni xaritalashda samarali qo'llaniladi va leksik hamda fonologik patterns (uslublar) tahlillarini namoyish etadi. Xususan, u dialektologiya sohasida hududiy sheva farqlarini aniqlash va tasniflashda, shuningdek, sud fonetikasi (forensic phonetics - nutq va ovoz yozuvlarini ilmiy tahlil qilish orqali sud-huquqiy masalalarni hal etishga ko'maklashish.) sohasida notanish so'zlovchilarning mintaqaviy kelib chiqishini aniqlashda muhim manba sifatida xizmat qiladi³. Chex dialekti korpusiga kelsak, Chexiyaning an'anaviy- mintaqaviy lahjalarini qamrab, korpus puxta ishlab chiqilgan sotsiolingvistik metama'lumotlar, ikki qatlamli transkripsiya, lemmatizatsiya va morfologik teglanish bilan ta'minlangan. Korpus puxta ishlab chiqilgan sotsiolingvistik metama'lumotlar, ikki qatlamli (bosqichli) transkripsiya, lemmatizatsiya va morfologik teglanish bilan ta'minlangan. KonText veb-interfeysi orqali ochiq kirish imkoni mavjud bo'lib, takrorlanadigan versiyalar saqlab boriladi. Korpus mutaxassislar va keng jamoatchilik, jumladan, ta'lim maqsadlarida foydalanish uchun mo'ljallangan. Dastlabki hajmi nisbatan kichik

² bu biror vazifani yoki ma'lumot to'plashni keng ommaga (ko'plab odamlarga) tarqatish usuli, odatda internet yoki ilova orqali.

³ Leemann, A., Kolly, M.-J., Britain, D. The English Dialects App: The creation of a crowdsourced dialect corpus // Ampersand. — 2017. — Vol. 5. — P. 1–17. — URL: <https://doi.org/10.1016/j.amper.2017.11.001>

(taxminan 100 000 soʻz); anʼanaviy shevalar va yoʻnaltirilgan suhbat sharoitiga qaratilgan boʻlib, keng qamrovli spontan kundalik nutqni qamrab olmaydi. Kichik korpuslar (subkorpuslar), dialektga xos tovush arxivi va interaktiv xarita asosidagi vizualizatsiyani (“Mapka” ilovasi) qoʻllab-quvvatlaydi; tadqiqot, maktablar va keng jamoatchilik uchun moʻljallangan⁴. Mazkur taqqosni jadvalda koʻrib oʻtish mumkin (1-jadval)

1-jadval

Korpus	Ustivorlik	Cheklov	Qamrov	Qoʻllanishi
EDAC	47 000 dan ortiq viktorina ishtirokchilari va 3 500 dan ziyod audio yozuv	yosh, etnik tarkib va taʼlim darajasi boʻyicha haqiqiy taqsimotiga toʻliq mos kelmaslik	kraudsorsing ⁵ metodologiyasi	mintaqaviy oʻzgaruvchanlikni xaritalashda samarali
Chexiya corpora	sosiolingvistik metamaʼlumotlar, ikki qatlamli transkripsiya, lemmatizatsiya va morfologik teglanish	hajmi nisbatan kichik (taxminan 100 000 soʻz)	Chexiyaning anʼanaviy-mintaqaviy lahjalari	Kichik korpuslar (subkorpuslar), tovush arxivi va interaktiv xarita asosidagi vizualizatsiya (“Mapka” ilovasi)
NorDiC	ikki turdagi transkripsiya; matn bilan sinxronlashtirilgan multimedia (audio/video)	ixtisoslashgan bilimni talab etadi	Besh skandinav tilining dialektal xususiyatlari	Ilgʻor tadqiqot vositasi, statistik kenglik

⁴ Grieve, J., Montgomery, C., Nini, A., Murakami, A., Guo, D. Mapping Lexical Dialect Variation in British English Using Twitter // *Frontiers in Artificial Intelligence*. — 2019. — Vol. 2. — URL: <https://doi.org/10.3389/frai.2019.00011> // Smith, A. The Cambridge handbook of English corpus linguistics // *ICAME Journal*. — 2016. — Vol. 40. — P. 143–147. — URL: <https://doi.org/10.1515/icame-2016-0009>

⁵ bu biror vazifani yoki maʼlumot toʻplashni keng ommaga (koʻplab odamlarga) tarqatish usuli, odatda internet yoki ilova orqali.

	materiallari; xarita va statistik ko'rsatkichlar; bepullik			
ArchiMob	og'zaki sheva ma'lumotlar to'plami	kam uchraydigan lingvistik hodisalar uchun ma'lumotlar tanqisligi	Shveytsariya nemis shevasi	dialektlar orasidagi lingvistik masofani modellashtirish tadqiqotlar ⁶
Curras	fonetik transkripsiya, avtomatik segmentatsiya va vaqtinchalik hizalanish ⁷	Kamresurslik, faqat matn shakli, 56 000 token	Falastin arab shevasi	(NLP) sohasidagi ⁸

Muxtasar qilganda, yuqoridagi dialektal korpuslar: ingliz shevalari mobil ilovasi, NorDic, ArchiMob, Curras qiyosiy tahlilidan krausoring vositasida katta hajmli ma'lumotdan tortib an'anaviy suhbat usullariga qadar ko'rib o'tdik. Bunda ba'zi korpuslarda fonetik-prosodik annotatsiya samaradorligini oshirishda ikki bosqichli transkripsiya, morfologik teglash imkoniyatlari hamda audiomatn moslashishi, ma'lumotlar tanqisligi kabi cheklovlar, hatto NLP uchun manba bo'la olishini ta'kidlaymiz.

Foydalanilgan adabiyotlar:

⁶ Alotaibi, Y., Meftah, A. Comparative evaluation of two arabic speech corpora // Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010). — 2010. — P. 1–5. — URL: <https://doi.org/10.1109/nlpke.2010.5587819>

⁷ bu korpusdagi matn va unga mos audio/video materiallarning vaqt bo'yicha sinxronlashtirilishi

⁸ Johannessen, J., Priestley, J., Hagen, K., Nøklestad, A., Lynum, A. The Nordic Dialect Corpus // Proceedings. — 2012. — P. 3387–3391. — URL: <https://doi.org/10.63317/466d6ni7t6n3>

1. Xolova M.A. 2025. "O'zbek tili dialektal bazasini shakllantirish prognozi" O'zbekiston: til va madaniyat Folklorshunoslik. Shevashunoslik 1: 22 – 29. doi.10.52773/tsuull.uzlc.2025.1 /ZAKB5324
2. Leemann, A., Kolly, M.-J., Britain, D. The English Dialects App: The creation of a crowdsourced dialect corpus // Ampersand. — 2017. — Vol. 5. — P. 1–17. — URL: <https://doi.org/10.1016/j.amper.2017.11.001>
3. Grieve, J., Montgomery, C., Nini, A., Murakami, A., Guo, D. Mapping Lexical Dialect Variation in British English Using Twitter // Frontiers in Artificial Intelligence. — 2019. — Vol. 2. — URL: <https://doi.org/10.3389/frai.2019.00011>
4. Smith, A. The Cambridge handbook of English corpus linguistics // ICAME Journal. — 2016. — Vol. 40. — P. 143–147. — URL: <https://doi.org/10.1515/icame-2016-0009>
5. Alotaibi, Y., Meftah, A. Comparative evaluation of two arabic speech corpora // Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010). — 2010. — P. 1–5. — URL: <https://doi.org/10.1109/nlpke.2010.5587819>
6. Johannessen, J., Priestley, J., Hagen, K., Nøklestad, A., Lynum, A. The Nordic Dialect Corpus // Proceedings. — 2012. — P. 3387–3391. — URL: <https://doi.org/10.63317/466d6ni7t6n3>