LINGVISTIK KORPUSLAR TADQIQI: MULTIMEDIA KORPUSI

Nuritdinov Abrorbek Sayfiddin o‘g‘li,
University of Business and Science
Til va adabiyot ta’limi kafedrasi o‘qituvchisi
E-mail: nuritdinovabrorbek@gmail.com

DOI: https://doi.org/10.5281/zenodo.18721246

Annotatsiya. *Ushbu maqolada korpus lingvistikasining nazariy asoslari, jahon tajribasi va turkiy tillarda yaratilgan multimedia korpuslarining tahlili amalga oshirilgan. Tadqiqotda multimedia korpuslarini yaratish bosqichlari, multimodal annotatsiyalash tamoyillari hamda ularning lingvistik tadqiqotlardagi ahamiyati o‘rganilgan. ELAN, ANVIL, Praat kabi maxsus dasturiy vositalar yordamida audio va video materiallarni annotatsiyalash metodologiyasi tahlil qilingan. Turkiy tillar, jumladan, turk, tatar, qozoq, qirg‘iz va o‘zbek tillarida yaratilgan multimedia korpuslari qiyosiy o‘rganilgan. Tadqiqot natijasida multimedia korpuslarini yaratishning optimal metodologiyasi ishlab chiqilgan va o‘zbek tili multimedia korpusini yaratish bo‘yicha tavsiyalar berilgan.*

Kalit so‘zlar: *multimedia korpusi, multimodal annotatsiya, og‘zaki nutq korpusi, ELAN, ANVIL, Praat, turkiy tillar korpusi, lingvistik annotatsiya, prosodik annotatsiya.*

Аннотация. *В этой статье представлены теоретические основы корпусной лингвистики, мировой опыт и анализ мультимедийных корпусов, созданных на тюркских языках. Исследуются этапы создания мультимедийных корпусов, принципы мультимодальной аннотации и их значение в лингвистических исследованиях. Была проанализирована методология аннотирования аудио- и видеоматериалов с помощью специальных программных инструментов, таких как ELAN, ANVIL, Praat. Сравнительное изучение мультимедийного корпуса тюркских языков, включая турецкий, татарский, казахский, киргизский и узбекский языки. В результате исследований была разработана оптимальная методология создания мультимедийного корпуса и даны рекомендации по созданию мультимедийного корпуса узбекского языка.*

Ключевые слова: *мультимедийный корпус, мультимодальная аннотация, устная речь, ELAN, ANVIL, Praat, корпус тюркских языков, лингвистическая аннотация, просодическая аннотация.*

Abstract. *This article presents the theoretical foundations of corpus linguistics, world experience and analysis of multimedia corpora created in Turkic languages. The stages of creating multimedia corpora, the principles of multimodal annotation and their importance in linguistic research are investigated. The methodology for annotating audio and video materials using special software tools such as ELAN, ANVIL, Praat was analyzed. Comparative study of the multimedia corpus of Turkic languages, including Turkish, Tatar, Kazakh, Kyrgyz and Uzbek. As a result of the research, an optimal methodology for creating a multimedia corpus was developed and recommendations were given for the creation of a multimedia corpus of the Uzbek language.*

Keywords: *multimedia corpus, multimodal annotation, oral speech, ELAN, ANVIL, Praat, corpus of Turkic languages, linguistic annotation, prosodic annotation.*

Kirish va dolzarbligi. Zamonaviy tilshunoslikda axborot-kommunikatsiya texnologiyalarining jadal rivojlanishi lingvistik tadqiqotlarni yangi bosqichga ko‘tardi. Korpus lingvistikasi tilshunoslikning eng ilg‘or yo‘nalishlaridan biriga aylangan bo‘lib, u tabiiy tildagi matn korpuslarini yaratish, mavjud korpuslarni tadqiq etish va ularning funksional imkoniyatlarini o‘rganish bilan shug‘ullanadi. Ushbu soha XX asrning 60-yillarida birinchi elektron korpuslarning paydo bo‘lishi bilan shakllana boshlagan bo‘lib,

bugungi kunda jahon miqyosida katta qiziqish uyg‘otayotgan va shiddat bilan jadallashib rivojlanayotgan ham ilmiy, ham amaliy jarayon hisoblanadi.

Korpus lingvistikasining shakllanishi tarixida muhim bosqichlar alohida ahamiyat kasb etadi. 1961–1964-yillarda Braun universitetida Nelson Frensis va Genri Kusera rahbarligida yaratilgan Brown korpusi birinchi elektron lingvistik korpus sifatida e’tirof etiladi va u keyingi barcha korpus tadqiqotlari uchun metodologik asos vazifasini o‘tadi. Mazkur korpus ingliz tilining Amerika varianti yozma matnlaridan tashkil topgan bo‘lib, 1 million so‘z qo‘llash holatini qamrab olgan edi. Keyinchalik 1970–1978-yillarda Lankaster-Oslo/Bergen (LOB) korpusi, 1975-yilda London-Lund korpusi yaratilishi bilan korpus lingvistikasining metodologik apparati yanada mustahkamlandi.

XX asrning 90-yillariga kelib kompyuter texnologiyalarining taraqqiyoti korpus lingvistikasining yangi bosqichiga yo‘l ochdi. Britaniya milliy korpusi (BNC), Amerika milliy korpusi (ANC), rus tilining milliy korpusi (RTMK) kabi yirik milliy korpuslarning yaratilishi tilshunoslikda empirik tadqiqotlarning ko‘lamini sezilarli darajada kengaytirdi. Bugungi kunda korpuslar nafaqat yozma matnlar majmuasi sifatida, balki og‘zaki nutq, audio va video materiallarni ham qamrab oluvchi murakkab tizimlar sifatida rivojlanmoqda.

Multimedia korpuslari korpus lingvistikasining eng zamonaviy va istiqbolli yo‘nalishlaridan birini tashkil etadi. An’anaviy yozma va og‘zaki matn korpuslaridan farqli o‘laroq, multimedia korpuslari bir vaqtning o‘zida bir nechta axborot kanallarini – matn, audio, video, imo-ishoralar va boshqa paralingvistik vositalarni – integratsiyalashgan holda taqdim etadi. Rus tilining multimedia korpusi (MURKO) bunday korpuslarning yorqin namunasi bo‘lib, unda 1930–2000-yillardagi kinofilm lavhalaridan olingan video, audio, ovozning matnli transkripsiyasi va kadrda uchraydigan imo-ishoralar parallel parametrlar sifatida qamrab olingan.

Multimedia korpuslarining o‘ziga xos jihati shundaki, ular tilni uning tabiiy kommunikativ muhitida – ko‘p modallik sharoitida o‘rganish imkoniyatini yaratadi. Odamlar kundalik muloqotda nafaqat so‘zlar orqali, balki tovush ohangi, yuz ifodasi, qo‘l harakatlari va boshqa noverbal vositalar yordamida ham ma’lumot uzatadi. Multimedia korpuslari aynan shu murakkab kommunikativ jarayonni to‘liq aks ettirish va tadqiq etish imkoniyatini beradi.

O‘zbek tili uchun multimedia korpusini yaratish masalasi ham dolzarb ahamiyat kasb etadi. O‘zbek tili agglutinativ tillar qatoriga kiruvchi turkiy tillardan biri bo‘lib, uning morfologik tuzilishi o‘ziga xos xususiyatlarga ega. Shu bois o‘zbek tili multimedia korpusini yaratishda morfologik annotatsiyalash tamoyillarini ishlab chiqish alohida ilmiy-amaliy ahamiyatga molikdir. Bunday korpus o‘zbek tilining og‘zaki va yozma shakllarini, prosodik xususiyatlarini, noverbal muloqot vositalarini kompleks tarzda o‘rganish imkoniyatini yaratadi.

Tadqiqotning dolzarbligi bir nechta omillar bilan belgilanadi: birinchidan, o‘zbek tilshunosligida multimedia korpuslari sohasidagi tadqiqotlar hali yetarli darajada rivojlanmagan; ikkinchidan, o‘zbek tilining morfologik xususiyatlarini hisobga olgan holda multimedia kontentini annotatsiyalashning ilmiy-metodologik asoslari ishlab chiqilmagan; uchinchidan, bunday korpusning yaratilishi o‘zbek tilini chet tili sifatida o‘qitish, nutq texnologiyalarini rivojlantirish, sun’iy intellekt tizimlarini takomillashtirish kabi bir qator amaliy vazifalarni hal etishga xizmat qiladi.

Metodlar va o‘rganilish darajasi. Lingvistik korpuslar va multimedia korpusi mavzusiga oid ilmiy adabiyotlarni tahlil qilish tadqiqotning nazariy asoslarini belgilash uchun muhim ahamiyat kasb etadi. Ushbu qismda korpus lingvistikasi sohasidagi fundamental tadqiqotlar, multimedia korpuslarining nazariy asoslari va o‘zbek tili korpusshunosligiga doir ishlar ko‘rib chiqiladi. Korpus lingvistikasi nazariy jihatdan Noam Chomskiyning generativ tilshunoslik konsepsiyasi bilan uzviy bog‘liq bo‘lib, uning 1957-yilda chop etilgan “Syntactic Structures” asari tilning formal tavsifi borasidagi tadqiqotlar uchun poydevor vazifasini o‘tadi. Biroq korpus lingvistikasining mustaqil soha sifatida shakllanishi empirik yondashuvning kuchayishi bilan bog‘liq bo‘lib, bu borada J.Sinclair, D.Biber, T.McEnery, A.Wilson kabi olimlarning tadqiqotlari alohida ahamiyatga molik.

J.Sinclairning 1991-yilda nashr etilgan “Corpus, Concordance, Collocation” asari[1] korpus lingvistikasi va kollokatsiya tahlili usullarini ishlab chiqishda muhim bosqich bo‘ldi. Olim korpusni tilning xilma-xilligini ko‘rsatish va turli ko‘rinishdagi til materiallarini tavsiflash maqsadida tanlangan asl matnlar yig‘indisi sifatida ta'riflaydi. Bu yondashuv hozirgi zamonaviy tabiiy tilni qayta ishlash (NLP) texnologiyalarining rivojlanishiga katta ta’sir ko‘rsatdi. D.Biber[2] nutq va yozuv o‘rtasidagi farqlarni o‘rganish bo‘yicha olib borgan tadqiqotlarida matn janrlari va stilistik xususiyatlarini farqlashda korpus lingvistikasi usullaridan samarali foydalangan. Uning 1988-yildagi “Variation Across Speech and Writing” asari matnlarning empirik tahlili uchun metodologik asos yaratdi. Keyinchalik D.Biber va R.Reppen tomonidan nashr etilgan “The Cambridge Handbook of English Corpus Linguistics” ingliz tilidagi korpus tadqiqotlari uchun fundamental manbaga aylandi.

T.McEnery va A.Wilsonning 2001-yilda chop etilgan “Corpus Linguistics: An Introduction” asarida tilshunoslik sohasida katta hajmdagi matn korpuslaridan foydalanishning nazariy va amaliy jihatlari atroflicha yoritilgan. Keyinchalik T.McEnery va A.Hardie tomonidan 2012-yilda nashr etilgan “Corpus Linguistics: Method, Theory and Practice” asari[3] metodologik yondashuvlarni yanada rivojlantirdi. M.Stubbs “Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture” asarida[4] kompyuter yordamida matn tahlilini chuqurlashtirish, lingvistik naqshlarni tahlil qilish va til hamda madaniyat o‘rtasidagi bog‘liqlikni aniqlash usullarini ishlab chiqdi. P.Bakerning “Using Corpora in Discourse Analysis” asari diskurs tahlilini korpus lingvistikasi bilan

birlashtirish metodikasini yaratishga bag‘ishlangan. S.Th.Gries tomonidan 2009-yilda nashr etilgan “Quantitative Corpus Linguistics with R: A Practical Introduction” tadqiqotida statistik yondashuvlardan foydalangan holda lingvistik tahlilni amalga oshirish usullari ishlab chiqilgan. Bu tadqiqot korpus lingvistikasining kvantativ metodlarini rivojlantirishda muhim qadam bo‘ldi.

Rus korpus lingvistikasida V.P.Zaxarov, S.Yu.Bogdanova, A.B.Kutuzov, Ye.V.Nedoshivina, D.Ye.Gruzdev, I.O.Kuznetsov kabi olimlarning tadqiqotlari alohida diqqatga sazovor. V.P.Zaxarov va S.Yu.Bogdanovaning 2011-yilda chop etilgan “Korpusnaya lingvistika” darsligida[19, 20] korpuslarning tuzilishi, ularning amaliy ahamiyati, razmetkalash standartlari haqida qimmatli ma’lumotlar berilgan. Ushbu darslik MDB mamlakatlaridagi korpus tadqiqotlari uchun muhim manba sifatida xizmat qilmoqda. Rus tilining milliy korpusi (RTMK) 2004-yilda onlayn tarzda foydalanila boshlagan bo‘lib, XVIII asr o‘rtalaridan XXI asr boshigacha bo‘lgan davriy manbalarni qamrab oladi. Ushbu korpus semantik, morfologik va sintaktik jihatdan razmetkalangan bo‘lib, uning hajmi 900 million so‘zni tashkil etadi. Ayniqsa, rus tilining multimedia korpusi (MURKO) 1930-2000-yillardagi kinofilm lavhalaridan tashkil topgan bo‘lib, videoryad, audioryad, ovozning matnli transkripsiyasi va kadrda uchraydigan imo-ishoralar kabi parallel parametrlarni qamrab olgan.

Multimedia korpuslari bo‘yicha tadqiqotlar alohida yo‘nalish sifatida rivojlanmoqda. Multimodal korpuslar muloqot ishtirokchisi videoyozuvini ham o‘z ichiga olgan bo‘lib, mimika, qo‘l, ko‘z, qosh harakati va boshqa belgilar asosida razmetkalanadi. Bunday korpuslar o‘zaro emotsional strategiya, ixtilof, muloqot odobi, nutqiy to‘xtam va boshqa hodisalarni o‘rganish vositasi bo‘lishi mumkin.

Irkutsk davlat tilshunoslik universitetida Ta’limiy multimodal korpus (UMKO) tuzish davom etmoqda. Bu korpus materiali rus va xitoy tili egalarining ma'lum mavzu bo‘yicha oldindan tayyorlanmagan ta'limiy dialoglari videoyozuvidir. Korpus ELAN dasturi asosida razmetkalangan bo‘lib, parallel korpus shaklini olgan. Ushbu tajriba multimediali korpus asosida imo-ishora va qo‘shimcha emotsional holatlarning og‘zaki nutqqa ta'sirini o‘rganish imkoniyatini yaratadi.

Multimedia korpuslarini annotatsiyalashda maxsus dasturiy ta'minotdan foydalaniladi. ELAN (EUDICO Linguistic Annotator) multimodal korpuslardagi audio va video ma'lumotlarni annotatsiyalash uchun mo‘ljallangan bo‘lib, bir nechta qatlamli annotatsiyani amalga oshirish imkonini beradi. ANVIL video annotatsiyasi uchun mo‘ljallangan instrument bo‘lib, imo-ishoralar va noverbal elementlarni teglashda qo‘llaniladi. Praat fonetik tahlil va prosodik annotatsiya uchun keng foydalaniladigan dasturdir.

Turkologiyada korpus lingvistikasi bo‘yicha sezilarli yutuqlarga erishilgan. A.V.Dibo va A.V.Sheymovichning turkiy tillar, xususan, xakas tili korpusida avtomatik morfologik tahlilni amalga oshirish bo‘yicha tadqiqotlari, E.Adali va K.Oflazerning turk tili

ISSN: 2992-8850
www.uzresearchers.com

morfologiyasining ikki bosqichli tahlili, D.Sh.Suleymanov, A.R.Gilmullin, R.R.Gataullinlarning tatar tilining morfologik analizatorini ishlab chiqish tamoyillariga bag‘ishlangan ishlari turkiy tillar uchun avtomatik morfologik tahlil metodologiyasini rivojlantirishda muhim ahamiyatga ega.

G.G.Torotoyev va A.N.Nogovitsina turkiy tillarning agglutinativ tabiatga ega ekanligi, har bir grammatik ma'noning alohida affikslar orqali ifodalanishi teglash jarayonini avtomatik tarzda amalga oshirishda katta imkoniyat berishini ta’kidlaydi. Bu xususiyat o‘zbek tili uchun ham taalluqli bo‘lib, morfologik annotatsiyalashni avtomatlashtirish imkoniyatini kengaytiradi.

O‘zbek kompyuter lingvistikasining shakllanishida dastlab lingvostatistikaga oid N.Yoqubova, M.Ayimbetov, S.Rizayev, S.Muhamedov kabi olimlarning izlanishlari muhim rol o‘ynagan. S.Muhamedovning P.P.Piotrovskiy bilan hammualliflikda yozgan “Injenernaya lingvistika i opyt sistemno-statisticheskogo issledovaniya uzbekskix tekstov” nomli kitobida lingvistik modellar, modellashtirish tamoyillari va o‘zbekcha matnlarning kvantativ tahlillari o‘rganilgan. So‘nggi yillarda o‘zbek korpus lingvistikasi sohasida A.Po‘latov, S.Muhamedova, A.Rahimov, Z.Xolmanova, N.Abduraxmonova, Sh.Hamroyeva, A.Eshmo‘minov, O‘.Xoliyorov, M.Abjalova kabi olimlarning tadqiqotlari alohida ahamiyat kasb etadi. Sh.Hamroyevaning “O‘zbek tili morfologik analizatorining lingvistik ta’minoti” mavzusidagi doktorlik dissertatsiyasida til kategoriyalarini avtomatik morfologik tahlilini amalga oshirishning umumiy tamoyillari belgilab berilgan. N.Abduraxmonovaning “O‘zbek tili elektron korpusining kompyuter modellari” monografiyasida matnlarni morfologik teglash va tahlil qilishda FST (Finite state transducer) texnologiyasi imkoniyatlari tadqiq qilingan. Olima tomonidan o‘zbek tili korpusi (uzbekcorpus.uz) asosida fe’lga xos 26985 ta kombinatsiya hosil qilinganligi ilmiy asoslab berilgan.

V.P.Zaxarov, B.Mengliyev va Sh.Xamroyevalarning 2021-yilda chop etilgan “Korpus lingvistikasi” o‘quv qo‘llanmasi o‘zbek tilida korpus lingvistikasi bo‘yicha fundamental manba sifatida ahamiyatga molik. Qo‘llanmada korpus lingvistikasining konsepsiyasi, shakllanishi, rivojlanishi va bugungi holati yoritilgan.

Ilmiy adabiyotlar tahlili lingvistik annotatsiyaning quyidagi asosiy turlarini ajratish imkonini beradi: morfologik (POS-tagging), sintaktik, semantik, anaforik, prosodik (diskurs) va temporal annotatsiya. Har bir annotatsiya turi o‘ziga xos metodologiya va dasturiy ta'minotni talab qiladi. Morfologik annotatsiya so‘z turkumlarini teglash (POS-tagging) va grammatik kategoriyalarni belgilash bilan bog‘liq. Turkiy tillar uchun yaratilgan korpuslarning teglash tizimi turlicha bo‘lib, “Turkiy morfema” portalida turkiy tillar uchun lingvistik resurslarni birlashtiruvchi umumiy annotatsiyalash tizimi ishlab chiqilgan.

Prosodik annotatsiya og‘zaki nutq korpuslarida urg‘u, ritm, ohang va mantiqiy urg‘uni ifodalash uchun qo‘llaniladi. Bu tur multimedia korpuslari uchun ayniqsa muhim

bo‘lib, audio ma’lumotlarni lingvistik tahlil qilish imkonini beradi. Temporal annotatsiya voqealar o‘rtasidagi vaqt munosabatlarini ifodalash uchun mo‘ljallangan bo‘lib, J.Pustejovskiyning TimeML annotatsion sxemasi bu borada keng qo‘llaniladi.

Tadqiqot materiali sifatida xalqaro ilmiy ma'lumotlar bazalaridan (Web of Science, Scopus, CLARIN) 2015-2024 yillarda nashr etilgan multimedia korpus lingvistikasiga oid 150 dan ortiq ilmiy maqola va monografiyalar tahlil qilindi. Shuningdek, turkiy tillar bo‘yicha mavjud multimedia korpuslari (turk tili TNC korpusi, tatar tili “Tugan Tel” korpusi, qozoq tili Almaty Corpus) empirik tahlil uchun asos qilib olindi. Tavsifiy metod orqali multimedia korpuslarining strukturaviy xususiyatlari va annotatsiyalash tizimlari tavsiflandi. Qiyosiy metod yordamida turli tillar multimedia korpuslari o‘zaro solishtirildi. Statistik metod korpuslarning hajmiy ko‘rsatkichlarini tahlil qilishda qo‘llandi. Modellashtirish metodi multimedia korpusini yaratish bosqichlarini sxematik shaklda ifodalashda foydalanildi.

ELAN (EUDICO Linguistic Annotator) – Max Planck Psixolingvistika instituti tomonidan ishlab chiqilgan multimodal annotatsiya dasturi. Ushbu dastur vaqt bo‘yicha sinxronlashtirilgan audio va video materiallarni ko‘p qatlamli annotatsiyalash imkonini beradi. ELAN formatida morfologik, sintaktik, prosodik va imo-ishoralar qatlamlari parallel ravishda yaratiladi.

ANVIL – M. Kipp tomonidan ishlab chiqilgan video annotatsiya vositasi bo‘lib, u asosan imo-ishoralar va yuz ifodalarini kodlash uchun mo‘ljallangan. Dastur XML formatida strukturalashtirilgan ma'lumotlarni eksport qilish imkoniyatiga ega.

Praat – prosodik tahlil uchun maxsus dastur bo‘lib, u nutq signalining akustik parametrlarini (fundamental chastota, intensivlik, spektral xususiyatlar) tahlil qilish va ToBI annotatsiya sxemasiga muvofiq prosodik annotatsiya yaratish imkonini beradi. Annotatsiya sxemalari sifatida quyidagilar tahlil qilindi: MUMIN (Multimodal Utterance and Non-verbal Interaction) kodlash sxemasi, Universal Dependencies (UD) morfosintaktik annotatsiya standarti, ToBI (Tones and Break Indices) prosodik annotatsiya tizimi.

Tadqiqot natijalari. Multimedia korpuslarining strukturaviy xususiyatlari. Multimedia korpuslari yozma matn korpuslaridan tubdan farq qiladi. Ular audio va video oqimlarni, transkripsiyalarni hamda turli darajadagi annotatsiyalarni o‘z ichiga oladi. CLARIN infrastrukturasida hozirgi kunda 133 ta og‘zaki nutq korpusi mavjud bo‘lib, ulardan 122 tasi transkripsiyalar va tegishli yozuvlarni, 11 tasi esa faqat transkripsiyalarni o‘z ichiga oladi. Britaniya milliy korpusining og‘zaki matn qismi (BNC Spoken) 1991-1994 yillarda yaratilgan bo‘lib, u ikki qismdan iborat: 40 foizi spontan suhbatlar va 60 foizi ma'ruza, intervyu kabi rejalashtirilgan nutqlardan tashkil topgan. Ushbu korpus og‘zaki nutq tadqiqotlari uchun muhim manba hisoblanadi.

Rus tili milliy korpusining multimedia qismi (MURKO) 1930-2000 yillardagi kinofilm lavhalaridan tashkil topgan. U videoryad, audioryad, ovozning matnli

rasshifrovkasi va imo-ishoralar kabi parallel parametrlarni qamrab oladi. Qidiruv nafaqat talaffuz qilinayotgan matn, balki imo-ishora va nutqiy harakat turi orqali ham amalga oshirilishi mumkin. Turkiy tillarda yaratilgan korpuslar tahlili. Tadqiqot doirasida turkiy tillarda yaratilgan quyidagi korpuslar tahlil qilindi: Turk tili milliy korpusi (TNC) www.tnc.org.tr manzilida joylashgan bo‘lib, 423 million so‘z, 491 ming tokendan iborat. Korpus morfologik teglash tizimiga ega va turli uslubdagi matnlarni qamrab oladi.

Tatar tili “Tugan Tel” milliy korpusi (tugantel.tatar) 2012-yilda yaratilgan bo‘lib, u morfologik, sintaktik va semantik annotatsiyalarga ega. Korpus hajmi 116 million so‘zni tashkil etadi. 2014-yilda korpusning avtomatik morfologik razmetkasi amalga oshirilgan va xalqaro Apertium loyihasi ishlab chiqqan morfologik teglar asosida turkiy tillar teglari tizimi tayyorlangan. Qozoq tili Almaty Corpus (veb-corpora.net/KazakhCorpus) morfologik annotatsiyalangan bo‘lib, u qozoq tilining grammatik xususiyatlarini aks ettiruvchi teglash tizimiga ega. Xakas tili elektron korpusi (khakas.altaica.ru) umumiy, lug‘at, grammatika, matnlar, badiiy matnlar korpusi va dialekt podkorpuslaridan tashkil topgan. Korpus Rossiya Fanlar akademiyasi Prezidiumining “Korpus tilshunosligi” dasturi doirasida yaratilgan.

Multimodal annotatsiyalash tamoyillari. Multimedia korpuslarini annotatsiyalashda quyidagi tamoyillarga amal qilinadi: Vaqt bo‘yicha sinxronizatsiya tamoyili: barcha annotatsiya qatlamlari audio va video oqim bilan vaqt bo‘yicha sinxronlashtirilgan bo‘lishi kerak. Bu ELAN dasturida millisekund aniqligida amalga oshiriladi. Ko‘p qatlamlilik tamoyili: annotatsiya turli sathlarni qamrab olishi lozim – morfologik (so‘z turkumlari, grammatik kategoriyalar), sintaktik (gap bo‘laklari, bog‘lanish turlari), prosodik (ohang, urg‘u, pauza), semantik (ma'no munosabatlari) va pragmatik (nutqiy aktlar, diskurs markerlari). Modullilik tamoyili: har bir annotatsiya qatlami mustaqil bo‘lishi va boshqa qatlamlarga bog‘liq bo‘lmasligi kerak. Bu turli tadqiqot maqsadlari uchun alohida qatlamlardan foydalanish imkonini beradi. Reprezentativlik tamoyili: korpus materiali tilning turli uslublari, janrlari va so‘zlovchilar demografik xususiyatlarini proporsional ravishda qamrab olishi lozim. Annotatsiya sxemalari standartlari. Xalqaro miqyosda quyidagi standartlar keng qo‘llaniladi: ISO 24614-1:2010 – yozma matnlarni so‘zma-so‘z segmentlash standarti. ISO 24610-1:2006 – elementlar strukturasi standarti. ISO/DIS 24611 – morfosintaktik razmetka standarti. ISO 24613:2008 – leksik razmetka sxemasi. ISO 24615:2010 – sintaktik annotatsiyalash tizimi (SynAF). Turkiy tillar uchun UniTurk doirasida turk, tatar, qozoq va qirg‘iz tillari uchun umumiy grammatik teglash tizimi yaratilgan. Bu tizim assotsiativ shaklda munosabatlar tizimiga bog‘langan va turkiy tillarning korpus analizi uchun asos bo‘lib xizmat qiladi.

O‘zbek tili multimedia korpusini yaratish istiqbollari. O‘zbek tili elektron korpusi (uzbekcorpus.uz) hozirda yozma matnlar asosida faoliyat yuritmoqda. Korpusda og‘zaki matn namunalarini kiritish uchun og‘zaki nutq jarayonida yozib olingan audio formatdagi matnlar transkripsiya yoki transliteratsiya qilinishi lozim. Multimedia korpusining

multimodal turi uchun nafaqat audio, balki muloqot jarayoni aks etgan video lavha ham zarur. Bunda kishilarning nutq jarayonida verbal va noverbal muloqot vositalarini (ko‘z, qosh, yuz harakatlari) qo‘llashi aks etgan videolavhalar til o‘rganuvchilarga muhim ma'lumot beradi. Chekli avtomat metodi (FST) yordamida o‘zbek tilining morfologik tahlil qiluvchi dasturi yaratilgan. Uning tarkibiga qoidalar (alifbo, fonologik qoidalar va fonetik hodisaga uchraydigan maxsus fonemalar) hamda lug‘at (barcha so‘z turkumlarining o‘zak va sodda yasama shakli) kiradi.

1-jadval. Turkiy tillar multimedia korpuslarining qiyosiy tahlili

Korpus nomi	Til	Hajmi	Annotatsiya turlari
TNC (Turkish National Corpus)	Turk	423 mln so‘z	Morfologik
Tugan Tel	Tatar	116 mln so‘z	Morfologik, sintaktik, semantik
Almaty Corpus	Qozoq	50 mln+ so‘z	Morfologik
Xakas korpusi	Xakas	100 ming+ so‘z	Morfologik, leksik
O‘zbek tili korpusi	O‘zbek	10 mln+ so‘z	Morfologik

Tadqiqot natijalarini jahon ilmiy adabiyotlari kontekstida muhokama qilish quyidagi xulosalarga olib keldi: Multimedia korpuslarining rivojlanish tendensiyalari. 2015-2024 yillar oralig‘ida multimodal diskurs tahlili sohasida sezilarli o‘sish kuzatildi. Bibliometrik tadqiqotlar shuni ko‘rsatadiki, multimodal diskurs tadqiqotlari soni eksponensial ravishda o‘sdi (y = 53.46e^0.2019x, R^2 = 0.9796). Bu sohadagi tadqiqotlar uchta bosqichni bosib o‘tdi: dastlabki bosqich (1997-2001), rivojlanish bosqichi (2002-2012) va yetuklik bosqichi (2013-2023).

So‘nggi yillarda multimedia korpuslarida sun'iy intellekt texnologiyalaridan foydalanish keng tarqaldi. LLM (Large Language Model) asosidagi annotatorlar korpuslarni avtomatik teglashda samarali natijalar ko‘rsatmoqda. Shu bilan birga, “human-in-the-loop” yondashuvi annotatsiyalarni qo‘lda tekshirish va to‘g‘rilash uchun zarur bo‘lib qolmoqda. Turkiy tillar multimedia korpuslarining o‘ziga xos xususiyatlari. Turkiy tillar agglutinativ xususiyatga ega bo‘lgani uchun ularning multimedia korpuslarini yaratishda maxsus yondashuvlar talab etiladi. Morfologik tahlil algoritmi morfemalarning lisoniy model chegarasiga ko‘ra asoslanib, kichik hajmdagi lug‘atdan 70 ming hajmdagi so‘zlarning morfologik shakllarini qamrab olishi mumkin. Tatar tili korpusi boshqa turkiy korpuslarga qaraganda barcha til sathlarining annotatsiyalangani va qo‘yilgan talablarga nisbatan to‘liq javob berishi bilan ajralib turadi. Konseptual va funksional modellarning sinflari muayyan til sathining struktural va funksional tavsifi uchun zarur bo‘lgan umumiy ma'lumotlardan tashkil topadi.

O‘zbek tili multimedia korpusi uchun tavsiyalar. O‘zbek tili multimedia korpusini yaratishda quyidagi tavsiyalarga amal qilish maqsadga muvofiq: Birinchidan, UniTurk doirasidagi turkiy tillar uchun umumiy grammatik teglash tizimidan foydalanish lozim. Bu

tizim o‘zbek tili morfologik va sintaktik teglash tizimi uchun asos bo‘lib xizmat qilishi mumkin. Ikkinchidan, ELAN dasturi orqali ko‘p qatlamli annotatsiya yaratish tavsiya etiladi. Bu dastur agglutinativ tillar uchun mos keladi va turli annotatsiya qatlamlarini parallel ravishda boshqarish imkonini beradi. Uchinchidan, FST (Finite State Transducer) texnologiyasidan foydalanish morfologik tahlilda samarali natijalar beradi. Bu texnologiya agglyutinativ tillarning morfotaktik xususiyatlarini to‘liq hisobga oladi. To‘rtinchidan, korpus materiali sifatida turli uslubdagi og‘zaki nutq namunalari (spontan suhbatlar, ma'ruzalar, intervyular, media matnlar) proporsional ravishda qamrab olinishi kerak.

Tadqiqot cheklovlari va kelajak yo‘nalishlari. Ushbu tadqiqotning asosiy cheklovi shundaki, u nazariy tahlilga asoslangan va empirik korpus yaratish tajribasini qamrab olmaydi. Kelajakda o‘zbek tili multimedia korpusini amaliy yaratish va sinovdan o‘tkazish zarur. Shuningdek, turkiy tillar uchun yagona annotatsiya standartlarini ishlab chiqish dolzarb masala bo‘lib qolmoqda. Barcha turkiy tillar doirasida matnlarni lingvistik annotatsiyalash tizimi uchun umumiy grammatik teglash va annotatsiyalash tamoyillari hamda mezonlari ishlab chiqilsa, tabiiy matnlarni qayta ishlashning ko‘p tilli texnologiyalarida foydali model bo‘lib xizmat qilishi shubhasiz.

Xulosalar. Multimedia korpuslari zamonaviy tilshunoslik tadqiqotlarida muhim lingvistik resurs sifatida xizmat qiladi. Ular an'anaviy yozma matn korpuslaridan farqli ravishda, tilning og‘zaki shakli, prosodika, imo-ishoralar hamda verbal va noverbal muloqot vositalarini kompleks o‘rganish imkonini beradi. Multimodal annotatsiyalash tamoyillari vaqt bo‘yicha sinxronizatsiya, ko‘p qatlamlilik, modullilik va reprezentativlikni o‘z ichiga oladi. ELAN, ANVIL, Praat kabi maxsus dasturiy vositalar multimedia materiallarni annotatsiyalashda samarali vositalar hisoblanadi.

Turkiy tillar korpuslari, jumladan, turk tili TNC, tatar tili “Tugan Tel”, qozoq tili Almaty Corpus kabilar multimedia annotatsiyalash metodologiyasini rivojlantirishda muhim tajriba to‘plagan. Tatar tili korpusi barcha til sathlarining annotatsiyalangani bilan boshqa turkiy korpuslarga qaraganda ajralib turadi. O‘zbek tili multimedia korpusini yaratishda UniTurk doirasidagi umumiy grammatik teglash tizimi, ELAN dasturi va FST texnologiyasidan foydalanish tavsiya etiladi. Korpus materiali turli uslubdagi og‘zaki nutq namunalarini proporsional ravishda qamrab olishi lozim. Kelajakda turkiy tillar uchun yagona annotatsiya standartlarini ishlab chiqish va o‘zbek tili multimedia korpusini amaliy yaratish dolzarb vazifalar hisoblanadi. Bu tadqiqotlar agglutinativ tillarning kompyuter tahlili va tabiiy tilni qayta ishlash texnologiyalarining rivojlanishiga hissa qo‘shadi.

Foydalanilgan adabiyotlar ro‘yxati:

1. Allwood J. (2008). Multimodal corpora. In A. Lüdeling & M. Kytö (Eds.), Corpus Linguistics: An International Handbook (pp. 207-225). Berlin: Mouton de Gruyter.
2. Baldry A., & Thibault P. J. (2008). Applications of multimodal concordances. Hermes – Journal of Language and Communication Studies, 41, 11-41.

3. Baisa V., & Suchomel V. (2012). Large Corpora for Turkic Languages and Unsupervised Morphological Analysis. Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12). Istanbul, Turkey: ELRA.
4. Belcavello F., Viridiano M., Matos E., & Timponi Torrent, T. (2022). Charon: A FrameNet annotation tool for multimodal corpora. Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) (pp. 91-96). Marseille, France: ELRA.
5. Dibo A. V., Sheymovich A. V. (2011). Morfologicheskaya razmetka korpusa xakasskogo yazyka. Rossiyskaya tyurkologiya, 2(5), 48-61.
6. Hamroyeva Sh. M. (2018). Korpus lingvistikasi atamalarining qisqacha izohli lug‘ati. Toshkent: Kamalak.
7. Jewitt C. (2009). The Routledge Handbook of Multimodal Analysis. London: Routledge.
8. Knight D. (2011). The future of multimodal corpora. Revista Brasileira de Linguística Aplicada, 11(2), 491-415.
9. Liu H., Liu L., Li H. (2024). Multimodal Discourse Studies in the International Academic Community (1997-2023): A Bibliometric Analysis. SAGE Open, 14(4).
10. Lovei R., Dembryii C., Hardiei A., Brezinai V., McEneryi T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics, 22(3), 319-344.
11. Mengliev, D., Nabiyeva, D., Abdurakhmonov, A., Makhmudov, K., Nuritdinov, A., & Otemisov, A. (2025, June). Educational Text Analysis in Uzbek: Developing an NER Algorithm for Academic and Pedagogical Content. In 2025 IEEE 26th International Conference of Young Professionals in Electron Devices and Materials (EDM) (pp. 2100-2103). IEEE.
12. Nuritdinov, A. (2025). MATNLARNI LINGVOSTATISTIK TAHLIL QILISHDA KORPUS USULLARIDAN FOYDALANISH. Молодые ученые, 3(19), 93-97.
13. Nuritdinov, A. (2025). KONKORDANS–LINGVISTIK TAHLIL VOSITASI SIFATIDA. Теоретические аспекты становления педагогических наук, 4(13), 173-178.
14. Nuritdinov, A. (2025). Korpus lingvistikasida lingvostatistik tahlil metodi. MAKTABGACHA VA MAKTAB TA’LIMI JURNALI, 3(5).
15. Nuritdinov, A. (2024). JADID DAVRI ADABIY MUHITIGA DOIR ASARLARDAN KORPUSDA FOYDALANISH. COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS, 1(1).
16. Nuritdinov, A. (2022). O ‘ZBEK TILI KORPUSI UCHUN ABDURAUF FITRATNING LINGVISTIK ASARLARINI MANBA SIFATIDA OLINISHI. COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS, 1(1).
17. Nuritdinov, A. S. O. G. L. (2022). O’zbek tili milliy korpusi uchun jadid tilshunoslarining lingvistik asarlarini manba sifatida olinishi. Science and Education, 3(4), 2048-2057.
18. Suleymanov D., Gilmullin R., Gataullin R. (2011). National Corpus of the Tatar Language: Grammatical Annotation and Implementation. 5th International Conference on Corpus Linguistics (CILC2013) (pp. 68-74).
19. Zaxarov V. P., Bogdanova S. Yu. (2011). Korpusnaya lingvistika. Irkutsk: IGLU.
20. Zaxarov V. P., Azarova I. V. va b. (2019). Modelirovaniye v korpusnoy lingvistike: Spetsializirovannyye korpusy russkogo yazyka. Sankt-Peterburg: SPbGU.
21. Frontiers in Communication. (2024). Rethinking multimodal corpora from the perspective of Peircean semiotics. doi: 10.3389/fcomm.2024.1337434