MULTIMEDIA KORPUSINING FUNKSIONAL IMKONIYATLARI

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.18721236

Annotatsiya. *Ushbu maqolada multimedia korpusining funksional imkoniyatlari kompleks tarzda tahlil qilingan. Tadqiqot multimedia korpuslarning zamonaviy tilshunoslik va tabiiy tilni qayta ishlash (NLP) sohasidagi amaliy qo‘llanilishi masalalarini qamrab oladi. Maqolada multimodal ma’lumotlarni integratsiyalash, ko‘p qatlamli annotatsiyalash, qidiruv-tahlil tizimlari hamda ta’limiy va ilmiy-tadqiqot maqsadlarida foydalanish imkoniyatlari o‘rganilgan. Tadqiqot natijasida multimedia korpusining asosiy funksional kategoriyalari aniqlangan va ularning o‘zbek tilshunosligini rivojlantirishdagi ahamiyati ko‘rsatib berilgan.*

Kalit so‘zlar: *multimedia korpusi, funksional imkoniyatlar, multimodal annotatsiya, qidiruv tizimi, konkordans, morfologik tahlil, tabiiy tilni qayta ishlash, NLP.*

Аннотация. *В этой статье всесторонне анализируется функциональность мультимедийного корпуса. Исследование охватывает практическое применение мультимедийного корпуса в области современной лингвистики и обработки естественного языка (НЛП). В статье рассматривается возможность интеграции мультимодальных данных, многоуровневых систем аннотирования, поиска и анализа, а также их использования в образовательных и исследовательских целях. В результате исследования были выявлены основные функциональные категории мультимедийного корпуса и определена их роль в развитии узбекской лингвистики.*

Ключевые слова: *мультимедийный корпус, функциональность, мультимодальная аннотация, поисковая система, согласование, морфологический анализ, обработка естественного языка, НЛП.*

Abstract. *This article comprehensively analyzes the functionality of a multimedia case. The study covers the practical application of the multimedia corpus in the field of modern linguistics and natural language processing (NLP). The article discusses the possibility of integrating multimodal data, multi-level annotation, search and analysis systems, as well as their use for educational and research purposes. As a result of the study, the main functional categories of the multimedia corpus were identified and their role in the development of Uzbek linguistics was determined.*

Keywords: *multimedia corpus, functionality, multimodal annotation, search engine, agreement, morphological analysis, natural language processing, NLP.*

Kirish va dolzarbligi. XXI asrning raqamli transformatsiya davrida til korpuslarining ahamiyati tobora ortib bormoqda. An’anaviy yozma matn korpuslaridan farqli ravishda, multimedia korpuslari audio, video va grafik ma’lumotlarni birlashtirgan holda tilning verbal va noverbal aspektlarini kompleks o‘rganish imkoniyatini yaratadi. O‘zbek tilshunosligida multimedia korpuslarini yaratish va ulardan samarali foydalanish masalalari dolzarb muammolardan biri hisoblanadi.

Hozirgi kunda jahon tilshunosligida multimedia korpuslarining funksional imkoniyatlarini o‘rganish bo‘yicha keng ko‘lamli tadqiqotlar olib borilmoqda. Rus tili multimediali korpusi (MURKO), Britaniya milliy korpusining og‘zaki nutq qismi, NMMC

(National Multi-Modal Corpus) kabi loyihalar multimedia korpusshunosligining nazariy va amaliy asoslarini belgilab berdi. Ushbu tadqiqotlar multimedia korpuslarining qidiruv, tahlil va vizualizatsiya imkoniyatlarini chuqur o‘rganish zarurligini ko‘rsatadi.

Tadqiqotning maqsadi o‘zbek tili multimedia korpusining funksional imkoniyatlarini tizimli ravishda tahlil qilish va ularning lingvistik tadqiqotlar hamda ta’lim jarayonidagi ahamiyatini ochib berishdan iborat. Tadqiqot vazifalari qatoriga multimodal ma’lumotlarni integratsiyalash mexanizmlarini o‘rganish, qidiruv va tahlil funksiyalarini tavsiflash, annotatsiyalash imkoniyatlarini baholash kiradi.

Metodlar va o‘rganilish darajasi. Korpus lingvistikasining shakllanishi G.Kennedi tomonidan Braun, LOB va London-Lund korpuslari bilan bog‘lanadi. J.Sinclair “Korpus, konkordans, kollokatsiya” asarida[1] kollokatsiya tahlili usullarini nazariy asoslab, COBUILD loyihasini amalga oshirdi. N.Chomskiyning generativ nazariyasi va R.Busaning Index Thomisticus loyihasi (1949–1980) kompyuter yordamida matnlarni tahlil qilishning dastlabki namunalari bo‘ldi. G.Leech rahbarligida Lankaster universitetida CLAWS avtomatik teglash tizimi yaratildi. T.McEnery va A.Wilson, T.McEnery va A.Hardie korpus metodologiyasini rivojlantirdilar. McEnery va R.Xiao korpusning til ta’limidagi imkoniyatlarini, D.Biber esa matn tahlili va registr tadqiqotlarini o‘rgandilar.

Multimedia korpuslarining nazariy asoslari L.Burnard tomonidan ishlab chiqilgan. Rus tili multimediali korpusi (MURKO) V.P.Zaxarov[19] ta’kidlaganidek, videoryad, audioryad va imo-ishoralarni qamrab oladi. S.Johansson TEI standartining og‘zaki diskursni kodlashdagi rolini tahlil qilgan. S.Kubler va H.Zinsmeister annotatsiyalash formatlarini, X.Lu esa kompyuter metodlarini tizimlashtirgan. Turkiy tillar bo‘yicha V.Baisa va V.Suchomel Sketch Engine platformasida 18.7 mln so‘zlik o‘zbek tili korpusini yaratdilar. Tatar tilining “Tugan Tel” korpusi D.Suleymanov va boshqalar tomonidan barcha til sathlari annotatsiyalangan holda ishlab chiqildi. O‘zbekistonda Sh.M.Hamroyeva mualliflik korpusi, A.A.Eshmo‘minov sinonim so‘zlar bazasi bo‘yicha tadqiqotlar olib bordilar.

Semantik annotatsiya bo‘yicha V.Basile va boshqalar GMB loyihasini, PropBank, FrameNet, Penn Discourse TreeBank kabilar turli annotatsiya darajalarini amalga oshirdilar. Universal Dependencies loyihasida 132 dan ortiq tilda treebank yaratilgan.

Tadqiqotda qiyosiy-tahliliy, tizimli-funksional va empirik usullardan foydalanildi. Qiyosiy-tahliliy usul orqali mavjud multimedia korpuslarining funksional arxitekturasi o‘rganildi. Tizimli-funksional usul yordamida korpus komponentlarining o‘zaro bog‘liqligi va integratsiyasi tahlil qilindi. Empirik usul asosida turli multimedia korpuslarining amaliy imkoniyatlari sinovdan o‘tkazildi. Tadqiqot materiallari sifatida rus tili multimediali korpusi (MURKO), Britaniya milliy korpusining og‘zaki qismi (BNC Spoken), Xelsinki annotatsiyalangan korpusi (HANKO), Sketch Engine platformasidagi turkiy tillar korpuslari hamda o‘zbek tili elektron korpusi (uzbekcorpus.uz) ma’lumotlaridan foydalanildi. Bundan tashqari, G.Leech, J.Sinclair, V.P.Zaxarov,

A.B.Kutuzov[20] kabi olimlarning korpus lingvistikasi bo‘yicha ilmiy ishlari nazariy asos sifatida xizmat qildi.

Funksional imkoniyatlarni tasniflashda quyidagi mezonlardan foydalanildi: ma’lumotlarni kiritish va saqlash funksiyalari, qidiruv va filtrlash imkoniyatlari, annotatsiyalash darajalari, statistik tahlil vositalari, vizualizatsiya mexanizmlari, foydalanuvchi interfeysi ergonomikasi. Har bir mezon bo‘yicha miqdoriy va sifatiy ko‘rsatkichlar aniqlandi.

Tadqiqot natijalari. Tadqiqot natijasida o‘zbek tili multimedia korpusining funksional imkoniyatlari beshta asosiy kategoriyaga ajratildi: multimodal integratsiya funksiyalari, qidiruv-navigatsiya imkoniyatlari, annotatsiyalash va teglash funksiyalari, statistik-analitik vositalar, ta’limiy-tadqiqot platformasi sifatidagi imkoniyatlar.

1. Multimodal integratsiya funksiyalari. Multimedia korpusining asosiy afzalligi turli formatdagi ma’lumotlarni yagona tizimda birlashtirish imkoniyatidir. O‘zbek tili multimedia korpusi quyidagi multimodal integratsiya funksiyalariga ega:

Audio-matn sinxronizatsiyasi. Og‘zaki nutq yozuvlari matn transkripsiyasi bilan vaqt bo‘yicha sinxronlashtiriladi. Bu imkoniyat nutq ohangi, pauza, temp kabi prosodik xususiyatlarni o‘rganishda muhim ahamiyatga ega. Britaniya milliy korpusining tajribasiga ko‘ra, audio-matn integratsiyasi nutqiy aktlarni tahlil qilish aniqligini 35-40 foizga oshiradi.

Video-verbal muvofiqlashtirish. Rus tili multimediali korpusi (MURKO) tajribasi shuni ko‘rsatadiki, videoryad va nutq matni orasidagi bog‘liqlikni aniqlash imo-ishoralar, mimika va boshqa noverbal vositalarni o‘rganish imkonini beradi. Korpusda 1930-2000-yillar kinofilmlaridan olingan lavhalar verbal va noverbal komponentlar bo‘yicha teglangan.

Grafik-lingvistik aloqador taqdimot. Multimedia korpusi matnli ma’lumotlarni grafik elementlar bilan birlashtirish funksiyasiga ega. Bu xususan terminologik birliklar, maxsus belgilar va formulalarni o‘z ichiga olgan ilmiy matnlar korpusida muhim ahamiyatga ega.

2. Qidiruv-navigatsiya imkoniyatlari. Korpus menejerining asosiy vazifasi foydalanuvchiga kerakli lingvistik ma’lumotni tez va aniq topish imkoniyatini yaratishdir. V.P.Zaxarovning ta’kidlashicha, zamonaviy korpus menejerlari quyidagi talablarni qondirishi lozim: to‘liq konkordans ro‘yxatini tuzish, murakkab so‘rovlarni qayta ishlash, natijalarni ko‘p parametrli saralash, statistik ma’lumotni uzatish. O‘zbek tili multimedia korpusining qidiruv tizimi uchta asosiy rejimda ishlaydi:

Lemma bo‘yicha qidirish. So‘zning lug‘at (asosiy) shakli bo‘yicha barcha grammatik variantlarni topish. Masalan, “yoz” lemmasini qidirganda “yozmoq”, “yozgan”, “yozilgan” kabi barcha shakldosh birliklar aniqlanadi. Bu funksiya o‘zbek tilining agglyutinativ xususiyatini hisobga olgan holda morfologik bazaga tayanadi.

Token bo‘yicha qidirish. Aniq grammatik shakldagi so‘zni topish imkoniyati. [asos+grammatik kategoriya] modeli asosida so‘rovlar yaratilib, yasovchi va shakl yasovchi qo‘shimchalar bo‘yicha differensiatsiya amalga oshiriladi.

Konkordans bo‘yicha qidirish. N-gram modeli asosida so‘zning o‘ng va chap qatordoshlari bilan birga topilishi. Bu imkoniyat kollokatsialar, turg‘un birikmalar va frazeologizmlarni aniqlashda muhim ahamiyatga ega. AntConc va Sketch Engine kabi instrumentariylar tajribasi shuni ko‘rsatadiki, konkordans tahlili so‘zlarning semantik maydonini aniqlashda 60-70 foiz aniqlik beradi.

1-jadval. Qidiruv funksiyalarining qiyosiy tavsifi

Qidiruv turi	Funksional xususiyati	Qo‘llanilish sohasi
Lemma qidirish	So‘zning barcha grammatik shakllarini topish	Leksikografiya, morfologik tadqiqotlar
Token qidirish	Aniq grammatik shakldagi so‘zni aniqlash	Grammatik tadqiqotlar, korpus statistikasi
Konkordans qidirish	So‘zning kontekstdagi qo‘llanilishini aniqlash	Kollokatsiya, frazeologiya tadqiqotlari
Multimodal qidirish	Verbal va noverbal birliklar integratsiyasi	Nutq akti tahlili, pragmatika

3. Annotatsiyalash va teglash funksiyalari. Multimedia korpusining funksional imkoniyatlari orasida ko‘p qatlamli annotatsiyalash alohida o‘rin tutadi. Annotatsiyalash korpusni oddiy matnlar to‘plamidan lingvistik tadqiqot instrumentiga aylantiruvchi asosiy omil hisoblanadi. K.K.Boyarskiy tasnifiga ko‘ra, korpus annotatsiyasi ikki asosiy turga bo‘linadi: metarazmetka (ekstralingvistik) va lingvistik razmetka. O‘zbek tili multimedia korpusida quyidagi annotatsiyalash darajalari mavjud:

Morfologik annotatsiya (POS-tagging). So‘z turkumlari, grammatik kategoriyalar (kelishik, son, shaxs, zamon, nisbat) bo‘yicha teglash. O‘zbek tilining morfologik bazasi 90 mingdan ortiq leksik birlikni qamrab oladi. FST (Finite State Transducer) metodidan foydalanilgan morfologik analizator so‘zning asosini, qo‘shimchalarini va grammatik ma’lumotini avtomatik aniqlaydi.

Sintaktik annotatsiya (parsing). Gap bo‘laklari, so‘zlar orasidagi sintaktik aloqalar, gap turlari bo‘yicha teglash. CONLLU formati asosida dependency parsing amalga oshiriladi. Bu format Universal Dependencies loyihasi doirasida standartlashtirilgan bo‘lib, 100 dan ortiq til uchun qo‘llaniladi.

Semantik annotatsiya. So‘z ma’nolari, semantik rollar, tematik guruhlar bo‘yicha teglash. Groningen Meaning Bank loyihasi tajribasiga asoslangan holda semantik annotatsiya sohaviy identifikatsiya, anafora, presuppozitsiya kabi faktorlarni qamrab oladi.

Prosodik annotatsiya. Urg‘u, ohang, ritm, pauza kabi suprasegment birliklar bo‘yicha teglash. Multimedia korpusida audio fayllar bilan birga prosodik ma’lumotlar ham saqlanadi. TEI (Text Encoding Initiative) yo‘riqnomasi asosida struktural, kontekstual va temporal annotatsiya amalga oshiriladi.

4. Statistik-analitik vositalar. Multimedia korpusi lingvistik ma’lumotlarni miqdoriy jihatdan tahlil qilish uchun keng imkoniyatlar yaratadi. Statistik vositalar quyidagi funksional kategoriyalarni o‘z ichiga oladi:

Chastota tahlili. So‘z, so‘z shakli va birikmalarning qo‘llanish chastotasini aniqlash. Absolyut va nisbiy chastota ko‘rsatkichlari, chastotali lug‘atlar avtomatik generatsiya qilinadi. Bu ma’lumotlar leksikografiya va til o‘qitish metodikasi uchun muhim ahamiyatga ega.

Kollokatsiya tahlili. So‘zlarning birikuvchanlik darajasini o‘lchash. MI (Mutual Information), T-score, Log-likelihood kabi statistik metrikalari yordamida kollokatsialar aniqlanadi. Sketch Engine platformasi Word Sketch funksiyasi orqali so‘zning sintaktik pozitsiyalari bo‘yicha taqsimotini vizualizatsiya qiladi.

Korpus solishtirish. Turli subkorpuslar yoki tashqi korpuslar bilan qiyosiy tahlil. Kalit so‘zlar (keywords) tahlili orqali ma’lum matn yoki janrga xos leksik birliklar aniqlanadi. Bu funksiya uslubiy va janriy xususiyatlarni o‘rganishda qo‘llaniladi.

5. Ta’limiy-tadqiqot platformasi sifatidagi imkoniyatlar. Multimedia korpusi nafaqat lingvistik tadqiqotlar, balki til ta’limi uchun ham samarali vosita hisoblanadi. T.McEnery va R.Xiao tadqiqotlariga ko‘ra, korpusga asoslangan til o‘qitish (corpus-based language teaching) an’anaviy metodlarga nisbatan bir qator afzalliklarga ega. O‘zbek tili multimedia korpusining ta’limiy funksiyalari quyidagilardan iborat:

Autentik material manbai. Korpusdagi matnlar haqiqiy muloqot vaziyatlaridan olingan bo‘lib, til o‘rganuvchilarga tabiiy til namunalarini taqdim etadi. Turli uslub, janr va davrga oid matnlar til variativligini o‘rganish imkonini beradi.

Induktiv o‘rganish muhiti. Konkordans tahlili orqali talabalar grammatik qoidalarni mustaqil kashf etish imkoniyatiga ega bo‘ladilar. Data-Driven Learning (DDL) yondashuvi asosida til hodisalarining real qo‘llanilishi o‘rganiladi.

O‘quv materiallarini tayyorlash bazasi. O‘qituvchilar korpusdan grammatika, leksika va uslubiyat bo‘yicha mashqlar tayyorlash uchun foydalanishlari mumkin. Chastota ma’lumotlari asosida o‘quv lug‘atlari va grammatik minimumlar ishlab chiqiladi.

Xulosalar. O‘zbek tili multimedia korpusining funksional imkoniyatlari tahlili quyidagi xulosalarga olib keldi:

1. Multimedia korpusi beshta asosiy funksional kategoriyani o‘z ichiga oladi: multimodal integratsiya, qidiruv-navigatsiya, annotatsiyalash, statistik-analitik vositalar va ta’limiy platforma. Bu kategoriyalar o‘zaro bog‘liq bo‘lib, korpusning yaxlit funksional arxitekturasini tashkil etadi.

2. Multimodal integratsiya funksiyalari tilning verbal va noverbal aspektlarini kompleks o‘rganish imkonini yaratadi. Audio-matn sinxronizatsiyasi, video-verbal muvofiqlashtirish va grafik-lingvistik aloqador taqdimot zamonaviy korpusshunoslikning muhim yutuqlari hisoblanadi.

3. Qidiruv tizimi o‘zbek tilining agglyutinativ xususiyatlarini to‘liq hisobga oladi. Lemma, token va konkordans bo‘yicha qidirish funksiyalari morfologik bazaga asoslangan bo‘lib, yuqori aniqlikda natijalar beradi.

4. Ko‘p qatlamli annotatsiyalash (morfologik, sintaktik, semantik, prosodik) korpusni chuqur lingvistik tadqiqotlar uchun samarali vositaga aylantiradi. Xalqaro standartlarga muvofiqlik qiyosiy tadqiqotlar imkoniyatini kengaytiradi.

5. Multimedia korpusi til ta’limi uchun DDL metodologiyasini qo‘llash platformasi sifatida katta salohiyatga ega. Autentik materiallar, statistik ma’lumotlar va vizualizatsiya vositalari til o‘qitish samaradorligini oshiradi.

Kelgusida multimedia korpusining funksional imkoniyatlarini yanada kengaytirish, xususan, sun'iy intellekt va mashinali o‘rganish texnologiyalarini joriy etish, foydalanuvchi interfeysini takomillashtirish va ta’limiy kontentni boyitish zarur.

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