

MULTIMEDIA KORPUSLARIDA AXBOROT QIDIRUV TIZIMLARINI LOYIHALASH VA TAKOMILLASHTIRISH METODLARI

Nuritdinov Abrorbek Sayfiddin o'g'li,
University of Business and Science
Til va adabiyot ta'limi kafedrası o'qituvchisi
G-mail: nuritdinovabrorbek@gmail.com

Annotatsiya. Mazkur maqolada multimedia korpuslarida axborot qidiruv tizimlarini loyihalash va takomillashtirishning ilmiy-nazariy, metodologik va amaliy asoslari tahlil qilinadi. Multimedia korpusi matn, audio, video, tasvir, nutq, mimika, jest, vaqt belgisi, transkripsiya va metama'lumotlarni birlashtirgan murakkab axborot-lingvistik tizim sifatida qaraladi. An'anaviy matnli korpuslarda qidiruv jarayoni asosan so'z, lemma, grammatik belgi va kontekstga tayansa, multimedia korpuslarida qidiruv matniy, akustik, vizual, semantik va pragmatik belgilarni uyg'un holda qayta ishlashni talab etadi.

Kalit so'zlar: multimedia korpus, axborot qidiruv tizimi, korpus lingvistikasi, multimodal ma'lumot, semantik qidiruv, annotatsiyalash, indekrlash, sun'iy intellekt, tabiiy tilni qayta ishlash.

Аннотация. В данной статье анализируются научно-теоретические, методологические и практические основы проектирования и совершенствования информационно-поисковых систем в мультимедийных корпусах. Мультимедийный корпус рассматривается как сложная информационно-лингвистическая система, объединяющая текст, аудио, видео, изображения, речь, мимику, жесты, временные метки, транскрипцию и метаданные. Если в традиционных текстовых корпусах процесс поиска в основном опирается на слово, лемму, грамматический признак и контекст, то в мультимедийных корпусах поиск требует комплексной обработки текстовых, акустических, визуальных, семантических и прагматических признаков.

Ключевые слова: мультимедийный корпус, информационно-поисковая система, корпусная лингвистика, мультимодальные данные, семантический поиск, аннотирование, индексирование, искусственный интеллект, обработка естественного языка.

Abstract. This article analyzes the scientific-theoretical, methodological, and practical foundations for designing and improving information retrieval systems in multimedia corpora. A multimedia corpus is regarded as a complex information-linguistic system that integrates text, audio, video, images, speech, facial expressions, gestures, time stamps, transcription, and metadata. While search processes in traditional text corpora are mainly based on words, lemmas, grammatical features, and context, information retrieval in multimedia corpora requires the integrated processing of textual, acoustic, visual, semantic, and pragmatic features.

Keywords: multimedia corpus, information retrieval system, corpus linguistics, multimodal data, semantic search, annotation, indexing, artificial intelligence, natural language processing.

Kirish. Jahonda raqamli texnologiyalar, sun'iy intellekt va tabiiy tilni qayta ishlash sohalarining jadal rivojlanishi tilshunoslik, ta'lim, axborot texnologiyalari va raqamli gumanitar fanlar oldiga yangi ilmiy vazifalarni qo'ymoqda. Bugungi kunda axborot faqat yozma matn shaklida emas, balki audio, video, tasvir, nutq, ovoz, harakat, mimika, jest va turli vizual belgilar orqali ham saqlanmoqda, qayta ishlanmoqda va foydalanuvchilarga taqdim etilmoqda. Ana shu jarayon korpus lingvistikasida ham yangi bosqichni yuzaga keltirdi. Agar dastlabki korpuslar asosan yozma matnlar to'plami sifatida shakllangan bo'lsa, hozirgi bosqichda multimedia korpuslari tilning yozma va og'zaki shakllarini, nutqiy vaziyatni, ovoz ohangini, vizual harakatni, kommunikativ maqsadni va metama'lumotlarni birlashtirgan murakkab axborot muhiti sifatida rivojlanmoqda. Shu

nuqtai nazardan multimedia korpuslarida axborot qidiruv tizimlarini loyihalash va takomillashtirish zamonaviy korpus lingvistikasi, kompyuter lingvistikasi va sun'iy intellekt texnologiyalarining dolzarb yo'nalishlaridan biri hisoblanadi.

Multimedia korpuslarida axborot qidiruv tizimini loyihalash faqat dasturiy ta'minot yaratish yoki ma'lumotlar bazasini shakllantirish bilan cheklanmaydi. Bu jarayon, avvalo, til birliklarini ilmiy asosda belgilash, ularni matniy, akustik va vizual kontekstda tahlil qilish, foydalanuvchi so'rovini mazmunan talqin etish va korpusdagi turli modallikdagi ma'lumotlarni o'zaro bog'lashni talab etadi. An'anaviy matnli korpuslarda foydalanuvchi asosan so'z, ibora, lemma, grammatik shakl yoki kontekst bo'yicha qidiruv amalga oshiradi. Multimedia korpuslarida esa qidiruv birligi ancha kengayadi: nutq parchasi, audio signal, video kadr, tasviriy ob'yekt, imo-ishora, yuz ifodasi, intonatsiya, pauza va vaqt belgisi ham qidiruv ob'yekti sifatida namoyon bo'ladi. Masalan, tadqiqotchi "o'zbek tilida iltimos ma'nosini bildiruvchi nutqiy vaziyatlar", "bolalar nutqida quvonch ifodasi", "ma'ruzachi nutqida pauza va ta'kid", "videomatnda qo'l harakati bilan birga aytilgan so'zlar" kabi murakkab so'rovlarni berishi mumkin. Bunday so'rovlarni oddiy kalit so'zli qidiruv orqali to'liq bajarib bo'lmaydi. Shuning uchun multimedia korpuslarida qidiruv tizimi lingvistik tahlil, akustik qayta ishlash, vizual belgilarni ajratish, semantik bog'lash va intellektual saralashni birlashtirgan holda ishlashi zarur.

Multimedia korpusining ilmiy ahamiyati uning ko'p modalli tabiati bilan belgilanadi. Modallik deganda axborotning ifodalanish kanali tushuniladi. Matn, ovoz, tasvir, video, mimika, jest va harakat alohida modalliklar hisoblanadi. Inson muloqoti ko'p hollarda bir nechta modallikning birgalikda ishlashi orqali amalga oshadi. Og'zaki nutqda so'z ma'nosi ovoz ohangi, pauza, yuz ifodasi, qo'l harakati va kommunikativ vaziyat orqali kuchaytirilishi yoki o'zgarishi mumkin. Masalan, bir xil so'z turli intonatsiya bilan aytilganda iltimos, hayrat, kinoya, buyruq yoki rad etish ma'nosini ifodalashi mumkin. Shu sababli multimedia korpuslarida axborot qidiruv tizimi so'zning faqat yozma shaklini emas, balki uning nutqiy, akustik va vizual muhitini ham hisobga olishi lozim. Bu holat multimedia korpusini oddiy fayllar to'plami emas, balki tilning real kommunikativ jarayondagi qo'llanishini aks ettiruvchi ilmiy resurs sifatida o'rganishni talab etadi.

Adabiyotlar tahlili. Tahlil qilingan mahalliy va xorijiy adabiyotlar multimedia korpuslarida axborot qidiruv tizimlarini rivojlantirishda korpus lingvistikasi, axborot qidiruv nazariyasi va sun'iy intellekt muhim o'rin tutishini ko'rsatadi. Xorijiy olimlar (Manning, Jurafsky, Baeza-Yates) tomonidan indekslash, semantik qidiruv va multimodal ma'lumotlarni qayta ishlashning nazariy asoslari ishlab chiqilgan.

Rossetto, Schuldt va Gasser ishlarida multimedia qidiruv tizimlarining amaliy modellari, ayniqsa audio, video va tasvirlarni integratsiya qilish masalalari yoritilgan. Mahalliy tadqiqotlarda esa o'zbek tili korpuslari, nutq korpusi va morfologik tahlil yo'nalishlari o'rganilgan.

Umuman olganda, adabiyotlar multimedia korpuslarda samarali qidiruv uchun semantik va multimodal yondashuvlar hamda sun'iy intellekt texnologiyalaridan foydalanish zarurligini tasdiqlaydi.

Tadqiqot metodologiyasi. Mazkur tadqiqot metodologiyasi multimedia korpuslarida axborot qidiruv tizimlarini loyihalash va takomillashtirish jarayonini ilmiy-nazariy hamda amaliy-uslubiy jihatdan tahlil qilishga asoslanadi. Tadqiqotda, avvalo, korpus lingvistikasi, kompyuter lingvistikasi, axborot qidiruv nazariyasi, tabiiy tilni qayta ishlash va sun'iy intellekt texnologiyalariga oid ilmiy yondashuvlar o'rganildi. Shu asosda multimedia korpusi matn, audio, video, tasvir, nutq va metama'lumotlarni birlashtiruvchi ko'p modalli axborot tizimi sifatida talqin qilindi.

Tadqiqot jarayonida tahliliy, qiyosiy, tizimli va funksional yondashuvlardan foydalanildi. Tahliliy yondashuv orqali multimedia korpuslarda ma'lumotlarni modellashtirish, annotatsiyalash, indekslash va qidiruv natijalarini saralash jarayonlari o'rganildi. Qiyosiy yondashuv asosida xorijiy tadqiqotlardagi content-based retrieval, semantic search, query-by-example va multimodal indexing kabi ilg'or metodlar o'zbek tili korpuslarini shakllantirish imkoniyatlari bilan taqqoslandi. Tizimli yondashuv multimedia korpusini matn, audio, video va metama'lumotlar o'zaro bog'langan yagona axborot-lingvistik tizim sifatida qarash imkonini berdi. Funksional yondashuv esa qidiruv tizimining foydalanuvchi ehtiyojiga mos ravishda axborotni topish, saralash, semantik bog'lash va taqdim etish vazifalarini aniqlashga xizmat qildi.

Tahlillar va natijalar. Multimedia korpuslarida qidiruv tizimini loyihalashning dastlabki bosqichi ma'lumotlarni to'g'ri modellashtirishdan iborat. Ma'lumot modeli korpusdagi har bir axborot birligining qanday saqlanishi, qanday annotatsiyalanishi, qanday indekslashtirilishi va qanday qidirilishi mumkinligini belgilaydi. Matnli ma'lumotlar uchun tokenizatsiya, lemmatizatsiya, morfologik tahlil, sintaktik teglash va semantik belgilash muhim bo'lsa, audio ma'lumotlar uchun fonetik transkripsiya, nutq segmentatsiyasi, akustik parametrlar, pauza, ovoz balandligi, nutq tezligi va vaqt kodlari muhim ahamiyatga ega. Video ma'lumotlar esa kadr, sahna, harakat, ob'yekt, jest, mimika va vizual kontekst asosida belgilanadi. Shuning uchun multimedia korpuslarida ma'lumotlar alohida-alohida saqlanmasdan, yagona axborot arxitekturasi doirasida o'zaro bog'langan holda tashkil etilishi kerak. Masalan, bir nutq parchasi uning transkripsiyasi, audio yozuvi, videodagi vaqt oralig'i, gapiruvchi haqidagi ma'lumot, mavzu tegi, grammatik belgisi va vizual harakati bilan bog'langan holda saqlansa, qidiruv tizimi ushbu ma'lumotlarni birgalikda tahlil qilish imkoniyatiga ega bo'ladi.

Annotatsiyalash multimedia korpusida axborot qidiruv tizimining samaradorligini ta'minlovchi asosiy ilmiy-uslubiy jarayondir. Annotatsiya ma'lumotga qo'shimcha ilmiy belgi berish jarayoni bo'lib, u korpusni oddiy ma'lumotlar omboridan tahlil qilinadigan ilmiy manbaga aylantiradi. Matnli annotatsiyada so'z turkumi, lemma, morfologik kategoriya, sintaktik vazifa, semantik maydon va pragmatik belgi ko'rsatilishi mumkin.

Audio annotatsiyada nutq segmenti, talaffuz xususiyati, intonatsiya, pauza, ovoz balandligi, nutq tezligi va gapiruvchi xususiyatlari belgilanadi. Video annotatsiyada esa kadr, sahna, harakat, ob'yekt, yuz ifodasi, jest va vizual kontekst qayd etiladi. Qidiruv tizimi aynan shu annotatsiyalar orqali ma'lumotlarni mazmunan topadi, saralaydi va foydalanuvchiga taqdim etadi. Agar korpusdagi ma'lumotlar yetarlicha annotatsiya qilinmagan bo'lsa, qidiruv tizimi faqat yuzaki moslikni aniqlaydi, lekin ma'lumotning chuqur semantik va kommunikativ mazmunini ochib bera olmaydi.

Metama'lumotlar ham multimedia korpus qidiruv tizimida muhim o'rin tutadi. Metama'lumotlar ma'lumot haqidagi ma'lumotlar bo'lib, fayl nomi, yozib olingan sana, joy, gapiruvchining yoshi, jinsi, kasbi, hududi, nutq janri, mavzu, format, davomiylilik va mualliflik ma'lumotlarini o'z ichiga oladi. Qidiruv tizimi ushbu metama'lumotlar orqali ma'lumotlarni saralaydi va foydalanuvchi ehtiyojiga mos natijalarni tanlaydi. Masalan, tadqiqotchi "Namangan viloyatidagi yosh gapiruvchilar nutqida ishlatilgan so'zlar"ni qidirganda, tizim geografik hudud, yosh guruhi, nutq janri va gapiruvchi ma'lumotlariga tayanadi. Demak, multimedia korpuslarda axborot qidiruv tizimini takomillashtirish uchun metama'lumotlar standartlashtirilgan, to'liq va izchil shaklda kiritilishi zarur.

Indekslash axborot qidiruv tizimining texnologik asosi hisoblanadi. Indekslash ma'lumotlarni tezkor topish uchun maxsus qidiruv tuzilmasini yaratish jarayonidir. Matnli korpuslarda invert indeks, lemma indeksi, morfologik indeks, n-gramma indeksi va semantik teglar indeksi keng qo'llanadi. Multimedia korpuslarida esa ushbu indekslar audio fingerprint, vizual belgilar indeksi, vaqt-kod indeksi, ob'yekt indeksi va vektorli semantik indekslar bilan boyitiladi. Masalan, foydalanuvchi "bolalar nutqida quvonch ifodasi"ni qidirsam, tizim bir vaqtning o'zida bir nechta indeksga murojaat qilishi kerak. Matn indeksi "quvonch", "xursand", "kuldi", "tabassum" kabi so'zlarni topadi, audio indeksi ovoz ohangi va emosional tonni aniqlaydi, video indeksi yuz ifodasi va harakatni tahlil qiladi, metama'lumotlar indeksi esa gapiruvchi yoshi, nutq vaziyati va mavzuni saralaydi. Shuning uchun multimedia korpuslarda qidiruv tizimi bitta indeksga emas, balki turli indekslarni birlashtirgan gibriddan asoslanishi kerak.

Xorijiy tadqiqotlarda multimedia axborot qidiruvi uchun content-based retrieval, query-by-example, semantic search va multimodal indexing kabi yondashuvlar keng qo'llanadi. Content-based retrieval ma'lumotni uning ichki xususiyatlari asosida qidirishni anglatadi. Bunda tasvirdagi rang, shakl, tekstura, videodagi harakat, audiodagi akustik xususiyatlar yoki nutqdagi fonetik belgilar qidiruv ob'yekti sifatida ishlatiladi.[1] Query-by-example esa foydalanuvchi tomonidan berilgan namuna asosida o'xshash ma'lumotlarni topish imkonini beradi. Masalan, foydalanuvchi muayyan ovoz namunasi, tasvir yoki video parcha kiritmaga, tizim korpusdan unga o'xshash audio-video materiallarni topib beradi. Bu yondashuv ayniqsa multimedia korpuslarda samarali bo'lib, u tilshunoslik, nutq texnologiyalari, ta'lim va vizual tahlil sohalarida keng imkoniyatlar yaratadi. R. Gasser, L. Rossetto va H. Schuldt tomonidan ishlab chiqilgan vitrivr tizimi

tasvir, audio, video va 3D ob'yektlar uchun mazmunga asoslangan qidiruv arxitekturasini taklif etadi.[5] Ushbu tajriba multimedia korpuslarda axborot qidiruv tizimi turli media turlarini alohida emas, balki yagona semantik muhitda tahlil qilishi zarurligini ko'rsatadi.

Multimedia korpuslarda qidiruv tizimini takomillashtirishning eng muhim yo'nalishlaridan biri semantik qidiruvdir. An'anaviy qidiruv so'zlarning yuzaki mosligiga tayansa, semantik qidiruv foydalanuvchi so'rovining mazmuni, konteksti, maqsadi va ma'noviy yaqinligini hisobga oladi. Masalan, foydalanuvchi "shodlik ifodasi" deb qidirganda, tizim faqat "shodlik" so'zi qatnashgan materiallarni emas, balki "quvonch", "xursandchilik", "tabassum", "kulgi", "mamnunlik" kabi mazmunan yaqin birliklarni ham topishi kerak. Agar korpusda audio-video ma'lumotlar mavjud bo'lsa, tizim ushbu semantik holatni ovoz ohangi, yuz ifodasi va vizual harakatlar orqali ham aniqlashi mumkin. Bu esa leksik-semantik tarmoqlar, sinonimlar bazasi, embedding modellari, transformer arxitekturalari va sun'iy intellekt algoritmlaridan foydalanishni talab etadi.

Sun'iy intellekt multimedia korpuslarida qidiruv imkoniyatlarini sezilarli darajada kengaytiradi. Mashinali o'qitish modellari audio ma'lumotlardan nutqni tanish, video ma'lumotlardan ob'jekt va harakatni ajratish, matndan ma'noni aniqlash va foydalanuvchi so'rovini intellektual qayta ishlash imkonini beradi. Masalan, avtomatik nutqni tanish tizimi audio yozuvni matnga aylantiradi, keyin bu matn morfologik va semantik tahlil qilinadi. Kompyuter ko'rish texnologiyalari videodagi yuz ifodasi, harakat, ob'jekt va sahnani aniqlaydi. Tabiiy tilni qayta ishlash modellari esa so'rovni lemmalashtiradi, ma'no jihatdan kengaytiradi va korpusdagi tegishli ma'lumotlar bilan bog'laydi. Shu asosda multimedia korpuslarda intellektual qidiruv tizimi foydalanuvchi so'roviga faqat formal mos natijani emas, balki mazmunan yaqin, kontekstual va ilmiy ahamiyatli natijalarni ham taqdim etadi.

O'zbek tili multimedia korpuslarida axborot qidiruv tizimlarini loyihalashda milliy tilning o'ziga xos xususiyatlarini hisobga olish alohida ahamiyatga ega. O'zbek tili agglyutinativ til bo'lganligi sababli so'zlarga turli qo'shimchalar qo'shilishi orqali ko'plab grammatik va semantik shakllar hosil bo'ladi. Masalan, "kitob", "kitoblar", "kitobim", "kitobimizdan", "kitobxon", "kitobxonlik" kabi shakllar bir leksik-semantik maydonga mansub bo'lsa-da, yuzaki qidiruv tizimi ularni alohida birliklar sifatida ko'rishi mumkin. Shuning uchun o'zbek tili korpuslarida qidiruv tizimi so'zshakl va lemma o'rtasidagi munosabatni to'g'ri aniqlashi, qo'shimchalarni ajratishi, grammatik kategoriyalarni belgilashi va so'z yasalishi jarayonini hisobga olishi lozim. M. Sharipov tomonidan o'zbek tili uchun morfologik va sintaktik teglangan korpus yaratishga oid tadqiqotlar o'zbek tilida avtomatik lingvistik tahlil va qidiruv tizimlarini takomillashtirish uchun muhim asos bo'lib xizmat qiladi.[6]

O'zbek tilidagi multimedia korpuslarida nutq ma'lumotlari bilan ishlash ham alohida ahamiyatga ega. Nutq ma'lumotlari yozma matndan farqli ravishda talaffuz, intonatsiya, pauza, urg'u, nutq tezligi, emosional holat va dialektal xususiyatlarni o'z ichiga oladi.

Uzbek Speech Corpus kabi resurslar 958 nafar gapiruvchi va 105 soatlik transkripsiyalangan audio ma'lumotlarni qamrab olgani bilan o'zbek tilida avtomatik nutqni tanish, ovozli qidiruv va nutq texnologiyalari uchun muhim manba hisoblanadi.[3] Bunday resurslar multimedia korpuslarda audio so'rov, nutqni matnga aylantirish, talaffuz tahlili, dialektal farqlarni aniqlash va interaktiv ovozli interfeyslarni yaratish imkonini beradi. Ayniqsa, ta'lim, til o'rganish, sud-lingvistik ekspertiza, ommaviy axborot vositalari tahlili va madaniy merosni raqamlashtirish sohalarida nutq korpuslarining ahamiyati kattadir.

Multimedia korpuslarida axborot qidiruv tizimini loyihalashda foydalanuvchi interfeysi ham muhim omil hisoblanadi. Qidiruv tizimi natijani faqat fayl nomi yoki matn parchasi sifatida emas, balki kontekst, transkripsiya, audio-video vaqt belgisi, grammatik tahlil, tarjima, metama'lumot va vizual ko'rsatkichlar bilan birga taqdim etishi kerak. Masalan, tilshunos muayyan so'zning jonli nutqda qanday talaffuz qilinishini o'rganmoqchi bo'lsa, tizim ushbu so'z ishtirok etgan barcha audio-video parchalarni, ularning vaqt kodini, gapiruvchi ma'lumotini, nutqiy vaziyatni va transkripsiyani ko'rsatishi lozim. Ta'lim jarayonida esa o'qituvchi yoki talaba muayyan grammatik shaklning turli nutqiy vaziyatlarda qanday ishlatilishini video va audio misollar orqali ko'rish mumkin. Shu jihatdan multimedia korpus qidiruv tizimi nafaqat axborot topish vositasi, balki ilmiy tahlil va ta'limiy foydalanish platformasi sifatida xizmat qiladi.

Qidiruv tizimini takomillashtirishda gibrid yondashuv eng maqbul metod sifatida namoyon bo'ladi. Chunki faqat kalit so'zli qidiruv multimedia ma'lumotlar mazmunini to'liq ochib bera olmaydi. Faqat vektorli semantik qidiruv esa lingvistik aniqlik, morfologik moslik va grammatik belgilarni har doim ham to'liq ta'minlay olmasligi mumkin. Shuning uchun multimedia korpuslarda kalit so'zli qidiruv, morfologik tahlil, semantik qidiruv, kontentga asoslangan audio-vizual qidiruv va metama'lumotlar filtri birlashtirilgan gibrid tizimni yaratish maqsadga muvofiqdir. Bunday tizimda foydalanuvchi so'rovi dastlab lingvistik jihatdan qayta ishlanadi, keyin semantik jihatdan kengaytiriladi, so'ngra matn, audio, video va metama'lumotlar indeksleri bo'yicha izlanadi. Natijalar esa relevantlik, kontekst, vaqt belgisi va foydalanuvchi ehtiyoji asosida saralanadi.

Multimedia korpuslarida qidiruv tizimini baholash mezonlari ham ilmiy jihatdan muhimdir. Axborot qidiruv tizimlari odatda precision, recall, relevance, response time va user satisfaction kabi ko'rsatkichlar asosida baholanadi. Biroq korpus lingvistikasida bunga qo'shimcha ravishda lingvistik aniqlik, annotatsiya to'g'riligi, kontekst to'liqligi, semantik moslik va natijaning tadqiqot maqsadiga muvofiqligi ham hisobga olinishi kerak. Masalan, "bolalar nutqida buyruq mayli" degan so'rov bo'yicha tizim faqat buyruq shaklidagi so'zlarni emas, balki ularning real nutqiy vaziyatdagi vazifasini, intonatsiyasini va muloqot kontekstini ham ko'rsatishi lozim. Demak, multimedia korpus qidiruv

tizimining samaradorligi texnik tezlik bilan birga lingvistik va semantik aniqlik orqali ham baholanadi.

Multimedia korpuslarda axborot qidiruv tizimlarini takomillashtirish tilshunoslik, ta'lim, tarjimashunoslik, nutq texnologiyalari, madaniy meros, jurnalistika, soslolingvistika va psixolingvistika kabi sohalar uchun muhim amaliy natijalar beradi. Tilshunoslikda bunday tizim so'zlarning real nutqdagi qo'llanishi, intonatsiya, pauza, nutqiy aktlar, dialektal farqlar va kommunikativ strategiyalarni o'rganish imkonini beradi. Ta'limda multimedia korpus o'quvchi va talabalarga til birliklarini faqat nazariy emas, balki real audio-video kontekstda o'rganish imkoniyatini yaratadi. Tarjimashunoslikda esa iboralar, nutqiy formulalar va madaniy belgilarning kontekstual ma'nosini aniqlashga yordam beradi. Nutq texnologiyalarida bunday korpuslar avtomatik nutqni tanish, matndan nutq yaratish, ovoqli assistentlar, mashinali tarjima va emosiyani tanish tizimlari uchun ma'lumot manbai bo'lib xizmat qiladi. Madaniy merosni saqlashda multimedia korpus og'zaki ijod namunalari, tarixiy suhbatlar, marosimlar, she'rxonlik, dostonchilik va xalq og'zaki nutqini raqamli shaklda saqlash va qidirish imkonini beradi.

Xulosa.Xulosa qilib aytganda, multimedia korpuslarida axborot qidiruv tizimlarini loyihalash va takomillashtirish zamonaviy korpus lingvistikasi, kompyuter lingvistikasi va sun'iy intellekt texnologiyalarining strategik ahamiyatga ega yo'nalishlaridan biridir. Multimedia korpusi matn, audio, video, tasvir, nutq, mimika, jest va metama'lumotlarni birlashtirgani sababli undagi qidiruv tizimi ham ko'p qatlamli, gibrid va intellektual bo'lishi zarur. Bunday tizimni loyihalashda ma'lumotlarni modellashtirish, annotatsiyalash, indekslash, semantik bog'lash, kontentga asoslangan qidiruv, tabiiy tilni qayta ishlash, nutqni avtomatik tanish va vizual ma'lumotlarni tahlil qilish asosiy metodlar sifatida qaralishi lozim.

Xorijiy ilmiy tajriba multimedia axborot qidiruvida content-based retrieval, query-by-example, semantic search va multimodal indexing yondashuvlari samarali ekanini ko'rsatadi. O'zbekistonda esa o'zbek tili elektron korpusi, ta'limiy korpus, nutq korpusi va morfologik-sintaktik teglangan korpuslar milliy korpus texnologiyalarini rivojlantirish uchun muhim ilmiy asos bo'lib xizmat qilmoqda. Kelgusida o'zbek tili multimedia korpuslarida qidiruv tizimlarini takomillashtirish uchun o'zbek tilining morfologik, fonetik, semantik va pragmatik xususiyatlarini hisobga oluvchi algoritmlar ishlab chiqish, audio-video ma'lumotlarni avtomatik annotatsiyalash, milliy til resurslarini kengaytirish va sun'iy intellekt modellarini o'zbek tili materiallari asosida o'qitish zarur. Shu asosda multimedia korpuslari tilshunoslik, ta'lim, tarjimashunoslik, nutq texnologiyalari, raqamli gumanitar fanlar va madaniy merosni saqlash sohalarini uchun muhim ilmiy-texnologik platformaga aylanadi.

Foydalanilgan adabiyotlar ro'yxati:

1. Lew M.S., Sebe N., Djeraba C., Jain R. Content-Based Multimedia Information Retrieval: State of the Art and Challenges. ACM Transactions on Multimedia Computing, Communications, and Applications. 2006.

2. Abjalova M. Educational Corpus of the Uzbek Language and its Opportunities. 2024.
3. Uzbek Speech Corpus and Automatic Speech Recognition. ISSAI va Toshkent axborot texnologiyalari universiteti hamkorligida tayyorlangan o'zbek nutq korpusi.
4. Abdurakhmonova N., Rajabova K. Stages of Creating Uzbek Language Multimedia Corpus. 2025.
5. Gasser R., Rossetto L., Schuldt H. Towards an All-Purpose Content-Based Multimedia Information Retrieval System. 2019.
6. Sharipov M. Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language. CEUR Workshop Proceedings.
7. Rossetto L., Giangreco I., Tanase C., Schuldt H. vitivr: A Flexible Retrieval Stack Supporting Multiple Query Modes for Searching in Multimedia Collections. Proceedings of ACM Multimedia. 2016.
8. Heller S., Rossetto L., Gasser R., Schuldt H. Multi-stage Queries and Temporal Scoring in vitivr. 2020.
9. Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2008.
10. Jurafsky D., Martin J.H. Speech and Language Processing. Stanford University, 2023.
11. Bird S., Klein E., Loper E. Natural Language Processing with Python. O'Reilly Media, 2009.
12. Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval: The Concepts and Technology Behind Search. Addison-Wesley, 2011.

