

BOLALAR NUTQI KORPUSI UCHUN MULTIMEDIYA VA AUDIO KONTENT YARATISH MASALALARI

Maxina Anvarjon qizi UMIRZAQOVA

magistrant

Mirzo Ulugʻbek nomidagi Oʻzbekiston Milliy universiteti

Toshkent, Oʻzbekiston

mahinasabirzanov@gmail.com

[ORCID 0009-0001-6355-7645](https://orcid.org/0009-0001-6355-7645)

Annotatsiya

Ushbu maqolada bolalar nutqi korpusi uchun multimediyaga va audio kontent yaratish jarayonida yuzaga keladigan asosiy masalalar hamda ularni hal etish yoʻllari tahlil qilinadi. Tadqiqotda bolalar nutqining yoshga xos fonetik, prosodik va leksik xususiyatlarini aniq qayd etishga xizmat qiluvchi audio yozuvlar sifati, mikrofonlash texnikasi, akustik muhitni nazorat qilish, maʼlumotlarni raqamlashtirish standartlari kabi omillar koʻrib chiqiladi.

Tayanch soʻzlar: bolalar nutqi korpusi, multimediyaga kontenti, audio yozuvlar, fonetik xususiyatlar, prosodiya, annotatsiya, metamaʼlumotlar, akustik muhit, etik masalalar, raqamlashtirish standartlari.

ВОПРОСЫ СОЗДАНИЯ МУЛЬТИМЕДИЙНОГО И АУДИОКОНТЕНТА ДЛЯ КОРПУСА ДЕТСКОЙ РЕЧИ

Махина Анваржон кизи УМИРЗАКОВА

магистрант

Национальный университет Узбекистана имени Мирзо Улугбека

Ташкент, Узбекистан

Электронная почта: mahinasabirzanov@gmail.com

ORCID: 0009-0001-6355-7645

Аннотация

В данной статье рассматриваются ключевые проблемы, возникающие при создании мультимедийного и аудиоконтента для корпуса детской речи, а также предлагаются возможные пути их решения. Анализируются факторы, обеспечивающие точную фиксацию фонетических, просодических и лексических особенностей детской речи: качество аудиозаписи, техника микрофонирования, контроль акустической среды и стандарты оцифровки данных.

Ключевые слова: корпус детской речи; мультимедийный контент; аудиозаписи; фонетические особенности; просодия; аннотирование; метаданные; акустическая среда; этические вопросы; стандарты оцифровки.

Soʻnggi yillarda tilshunoslik va nutq texnologiyalari sohasida korpus lingvistikasi muhim oʻrinni egallab kelmoqda. “Dunyo kompyuter lingvistikasi integrallashgan soha sifatida shiddat bilan rivojlanib borayotgan davrda qator usullar va metodlar orqali til muammolarini hal qilishda muhim vosita boʻlib

xizmat qilmoqda. Kompyuter texnologiyalarining til rivojiga va aksincha, til texnologiyasining kompyuter texnologiyalariga ijobiy ta'siri orqali qator ilmiy yutuqlarga erishildi. Jahonda kompyuter lingvistikasining turli yo'nalishlari bo'yicha ko'plab olimlar tomonidan ilmiy izlanishlar olib borildi" [1]. Xususan, bolalar nutqini o'rganishga qaratilgan maxsus korpuslar yaratish masalasi dolzarb ahamiyatga ega bo'lib, ular nafaqat lingvistik tadqiqotlar, balki nutqni avtomatik qayta ishlash, ta'limiy dasturlarni ishlab chiqish va logopedik diagnostika jarayonlari uchun ham muhim manba hisoblanadi. Bunday korpusni yaratish jarayonida multimediya va audio kontentning sifatli hamda tizimli to'planishi alohida ahamiyat kasb etadi, chunki bolalar nutqi fonetik, prosodik va artikulyatsion xususiyatlari bo'yicha kattalar nutqidan jiddiy farq qiladi.

Multimedia va audio materiallarni yig'ish, qayta ishlash va annotatsiya qilish jarayonlari bir qator texnik, metodik va etik muammolar bilan bog'liq. Jumladan, yozuvlar sifatini ta'minlash, atrof-muhit shovqinini minimallashtirish, yosh toifalari bo'yicha balansni saqlash, shuningdek, bolalar bilan ishlashda maxfiylik va rozilik masalalariga qat'iy rioya qilish zarur.

Bolalar nutqi korpusi – turli yoshdagi bolalarning tabiiy va sun'iy sharoitlarda yozib olingan nutq namunalaridan tashkil topgan, ilmiy va amaliy tadqiqotlar uchun mo'ljallangan tizimli ma'lumotlar bazasidir. Bunday korpuslar lingvistik tahlillar, nutqni avtomatik tanish (ASR), matndan ovozga o'tkazish (TTS), pedagogik texnologiyalarni yaratish va logopedik diagnostika jarayonlarida muhim manba vazifasini bajaradi.

Bolalar nutqi kattalar nutqidan fonetik, leksik va grammatik jihatdan sezilarli farqlarga ega. Masalan, ular ko'pincha qisqartirilgan shakllardan foydalanadi, ba'zi fonemalarni barqaror talaffuz qila olmaydi, prosodiya esa rivojlanish bosqichiga ko'ra turlicha shakllanadi. Shu bois bolalar nutqi bo'yicha maxsus korpus yaratish ularning nutq rivojlanishini chuqur va tizimli o'rganish imkonini beradi hamda zamonaviy til texnologiyalarini ishlab chiqishda alohida ahamiyat kasb etadi.

Multimedia va audio kontent yig'ish metodlari. Bolalar nutqini muvaffaqiyatli yig'ish jarayoni puxta rejalashtirilgan metodikaga tayangan holda amalga oshirilishi zarur. Avvalo, sifatli yozuv uskunalaridan foydalanish, ya'ni tashqi shovqinni kamaytiruvchi mikrofonlar, mobil qurilmalar yoki portativ audio yozuvchilarni tanlash talab etiladi. Yozuv jarayoni studiya sharoitida ham, uy yoki bog'cha muhitida ham tashkil qilinishi mumkin. Muhimi, bolaning o'zini erkin his qilishi va tabiiy nutq namunasini taqdim eta olishi hisoblanadi.

Nutq namunalarini yig'ishda turli topshiriq turlaridan foydalaniladi: erkin suhbat, rasmga qarab hikoya qilish, rolli o'yinlar, savol-javob shaklidagi dialoglar, she'r yoki kichik matn o'qish va boshqalar. Yosh guruhlariga qarab metodlar farq qiladi. Masalan, 2–4 yoshli bolalar bilan ko'proq o'yinli yondashuv, 5–7 yoshdagilar bilan mavzuli suhbatlashuv, 7 yoshdan kattalar bilan esa matn asosida nutq materialini yig'ish samarali bo'ladi. Yosh bo'yicha balansni saqlash korpusning ilmiy qiymatini oshiradi.

Video bilan ishlash transkript qanday shaklda bo'lishidan qat'i nazar, u hech qachon o'zaro ta'sirda nima bo'lganligi haqida to'liq aniq yozuvni o'z ichiga olmaydi. O'zaro ta'sirning transkripti hech qachon audiotasmaning o'rnini to'liq bosa olmaydi, chunki o'zaro ta'sirning audio yozuvi sodir bo'lgan voqealarning haqiqiy tafsilotlarini saqlash nuqtayi nazaridan har doim aniqroq bo'ladi. Xuddi shu tarzda, audio yozuv hech qachon yuqori sifatli audio trekka ega video yozuv kabi ko'p tafsilotlarni saqlay olmaydi. Audio yozuvlar ko'pincha suhbat o'zaro ta'sirining asosini tashkil etuvchi og'zaki bo'lmagan o'zaro ta'sirlarning hech birini yozib olmaydi. Shuning uchun, ular o'zaro ta'sirni to'liq talqin qilish uchun muhim bo'lgan ma'lumot manbaini muntazam ravishda chiqarib tashlaydi. Hatto, video yozuvda ham tarfkashliklar mavjud bo'lsada, bu bizda mavjud bo'lgan o'zaro ta'sirning eng aniq yozuvidir. O'zaro ta'sirning to'liq, batafsil xususiyatini aks ettirish uchun transkripsiyadan foydalanishga harakat qilayotganlar uchun transkripsiya tahlilning barcha bosqichlarida takroriy ko'rib chiqilishi kerak bo'lgan video yozuvda amalga oshirilishi juda muhimdir [5].

Korpusni teglash matnni lingvistik tahlil qilishda foydalaniladi. Matnni razmetka qilish tilning turli jihatlarni qamrab olishi mumkin. Masalan, yozma matnlar razmetkasida matnning struktur birliklari, sarlavhasi hisobga olinsa, ogʻzaki matnlar uchun tekshirish obyekti sifatida belgilangan soʻzlovchilarning yoshi, jinsi, kasbi kabi tomonlari ham ifodalanishi mumkin. Kiritilgan maʼlumotlarni mashina oʻqiy oladigan formatda berish uchun matnni kodlash yoʻriqnomasi (Text Encoding Initiative (T.E.I.)) orqali standart tilda beriladi ” [1].

Audio maʼlumotni qayta ishlash jarayonlari. Yigʻilgan audio materiallar korpusga kiritilishidan oldin texnik qayta ishlash bosqichidan oʻtadi. Sifat parametrlari (44.1 kHz yoki 48 kHz, 16-bit yoki 24-bit) standartlarga moslashtiriladi. Shovqinlarni kamaytirish, keraksiz artefaktlarni bartaraf etish, ovozni normallashtirish kabi jarayonlar yozuv sifati uchun zarur.

Nutq segmentatsiyasi – korpus yaratishning muhim bosqichi boʻlib, yozuvlar soʻz, gap yoki fonema darajasida boʻlinadi. Pauza, takrorlash, talaffuzdagi xatoliklar va nutq uzilishlari alohida qayd etiladi. Segmentatsiya natijasida maʼlumotlar strukturaviy jihatdan puxta boʻladi va ularni keyingi annotatsiya jarayoniga tayyor holga keltiradi.

Annotatsiya va belgilash tizimining muammolari. Bolalar nutqi koʻplab oʻziga xos xususiyatlarga ega boʻlgani sababli annotatsiya jarayoni murakkab sanaladi. Fonetik annotatsiya fonemalardagi oʻzgarishlar, talaffuzdagi xatolar va fonologik jarayonlarni aniq qayd etishni talab qiladi. Leksik annotatsiya esa, bolalar tilidagi noverbal birliklar – tovushlar, undovlar, oʻyinga xos soʻzlar va shu kabilarni ham qamrab olishi lozim.

Prosodik belgilash intonatsiya, urgʻu, ritm va pauzalarni oʻz ichiga oladi. Bu koʻrsatkichlar ayniqsa, bolalarda har xil tarzda namoyon boʻlgani uchun ularga alohida eʼtibor qaratiladi. Qoʻlda annotatsiya aniqroq boʻlsada, koʻp vaqt talab qiladi. Avtomatlashtirilgan annotatsiya esa tezkor, biroq bolalar nutqidagi farqlarni toʻliq aniqlashda qiynaladi. Shu bois, odatda, kombinatsiyalashgan yondashuv qoʻllanadi.

Etik va huquqiy masalalar. Bolalar bilan ishlashda etik tamoyillarga qat'iy rioya qilish zarur. Tadqiqot boshlanishidan avval ota-onalardan yozma rozilik olinishi, ular yozuvlar qayerda va qanday maqsadda qo'llanishi haqida batafsil xabardor qilinishi kerak. Shaxsiy ma'lumotlar maxfiyligini ta'minlash maqsadida ismlar, manzillar kabi identifikatsiya qiluvchi ma'lumotlar olib tashlanadi yoki kodlanadi.

Audio materiallarning xavfsiz saqlanishi, uchinchi shaxslarga berilmasligi va faqat ilmiy maqsaddagina foydalanilishi bo'yicha qat'iy tartiblar ishlab chiqilishi lozim. Bu nafaqat huquqiy talab, balki axloqiy mas'uliyat hamdir.

Texnik muammolar va ularni bartaraf etish yo'llari. Bolalar nutqi korpusi yaratishda turli texnik qiyinchiliklar uchrashi tabiiy. Masalan, yozuv sifati past bo'lib qolishi, tashqi shovqinning kuchayishi, turli qurilmalarda yozilgan audio materiallarning bir-biridan farq qilishi kabi holatlar. Bu muammolarni bartaraf etish uchun yozuvlarni standartlashtirish, texnik sozlamalarni bir xil formatda saqlash, shovqin filtrlash vositalaridan foydalanish tavsiya etiladi.

Korpusni saqlash va boshqarish jarayonida maxsus dasturlar – masalan, ELAN, Praat, CHILDES vositalari, CLAN platformalari yoki o'ziga xos korpus boshqaruv tizimlari qo'llanadi. Bunday tizimlar materiallarni izlash, filtrlash va tahlil qilish jarayonini yengillashtiradi.

Transkriptlar va media fayllarni yuklab olish haqida CLAN qo'llanmasida shunday deyiladi: “Agar siz transkriptlarni batafsilroq o'rganmoqchi bo'lsangiz, ularni ko'rib chiqiladigan ma'lumotlar bazasi orqali yuklab olishni xohlaysiz. Forrester transkriptlarini misol sifatida ishlatib, buni qanday qilish kerak:

1. <https://childes.talkbank.org> saytida Index to Corpora, keyin Eng-UK va Forrester-ni bosing.

2. Yuklab olish transkriptlari deb nomlangan havolani bosing va .zip fayli kompyuteringizga yuklab olinadi.

3. Agar u avtomatik ravishda arxivdan chiqarilmasa, uni arxivdan chiqarishingiz kerak.

4. Media fayllarni yuklab olish uchun Media papkasiga havolani bosing va keyin alohida videolarni birma-bir yuklab olishingiz mumkin. Media fayllarni yuklab olish transkriptni yuklab olishdan ko‘ra ancha ko‘proq vaqt talab etadi [5].

Faqat audio materiallar bilan cheklanish ko‘p hollarda bolalar nutqining to‘liq manzarasini bera olmaydi. Shu sababli video materiallar jalb qilinadi. Video orqali artikulyatsiya, mimika, imo-ishora, emotsional reaksiyalar va nutqga bog‘liq boshqa noverbal belgilar qayd etiladi. Bu multimodal tahlillarni yanada kengaytiradi.

Rasm, animatsiya yoki o‘yinli media kontent bolalar nutqini rag‘batlantirishda samarali vosita hisoblanadi. Shuningdek, audio va videoni matn anotatsiyasi bilan birlashtirgan holda multimodal korpus yaratish ko‘plab tadqiqot yo‘nalishlari uchun katta imkoniyatlar ochadi.

Jahon tajribasida CHILDES, CU Boulder Child Speech Corpus, LENA tizimi asosidagi korpuslar keng qo‘llanadi. Ushbu loyihalar bolalar nutqini turli jihatlarida o‘rganish, avtomatik tanish tizimlarini yaratish, logopedik standartlarni ishlab chiqishda katta yordam bergan. Ushbu dasturlardan foydalangan holda o‘zbek tili korpusi tarkibida bolalar nutqi korpusini yaratishimiz mumkin. Masalan, CLAN programmasi ham korpus yaratish uchun asos bo‘la oladi. Dastur qo‘llanmasida shunday deyiladi:

“Agar siz suhbatlashish, til o‘rganish yoki til buzilishlarini o‘rganuvchi tadqiqotchi bo‘lsangiz, CLANDan foydalanishni o‘rganishingiz kerak bo‘ladi, chunki bu sizga asosiy tadqiqot savollariga javob berishga va ko‘plab turdagi tillarni o‘rganishga yordam beradi. Shuningdek, transkripsiyani tezlashtirish, transkriptlarni mediaga bog‘lash, ma’lumotlarni avtomatik akustik tahlilga yuborish va keng ko‘lamli morfosintaktik xususiyatlarni avtomatik hisoblash uchun kuchli usullarni taqdim etadi” [5].

O‘zbek tilida esa bolalar nutqi bo‘yicha keng ko‘lamli standart korpuslar hali shakllanish jarayonida. Shuning uchun, mazkur maqolada taqdim etilayotgan metodik yondashuvlar kelajakda o‘zbek bolalar nutqi korpusini yaratish uchun nazariy va amaliy asos bo‘lib xizmat qilishi mumkin.

Bolalar nutqi korpusi uchun multimedia va audio kontent yaratish jarayoni murakkab, ko'p bosqichli va mas'uliyatli ilmiy-amaliy faoliyat hisoblanadi. Tadqiqot davomida ko'rib chiqilgan jihatlar shuni ko'rsatadiki, bunday korpusni yaratishda texnik, metodik, lingvistik va etik talablarning uyg'un ravishda bajarilishi alohida ahamiyatga ega. Audio va video materiallarni sifatli yig'ish, ularni standartlarga mos ravishda qayta ishlash, aniq va tizimli annotatsiya qilish korpusning ilmiy qiymatini belgilovchi asosiy omillardir.

Bolalar nutqi o'ziga xos fonetik, prosodik va grammatik xususiyatlarga ega bo'lgani sababli, mazkur jarayon boshqa turdagi korpuslarga nisbatan ko'proq murakkabliklarni yuzaga keltiradi. Shu bois yozuv muhiti, topshiriq turlari va yosh bo'yicha segmentatsiya ilmiy talablarga muvofiq rejalashtirilishi lozim. Shuningdek, bolalar bilan ishlashda maxfiylik, xavfsizlik va ota-onalarning roziligini ta'minlash kabi etik masalalarga qat'iy rioya etilishi zarur.

Multimedia (video, rasm, animatsiya) va audio ma'lumotlarning uyg'un qo'llanilishi korpusning funksional imkoniyatlarini kengaytiradi, multimodal tahlil uchun keng sharoit yaratadi. Bu esa lingvistika, psixolingvistika, logopediya, nutq texnologiyalari va ta'lim sohalarida qo'llanish doirasini yanada boyitadi.

Umuman olganda, bolalar nutqi korpusini yaratish bo'yicha ishlab chiqilgan metodik va texnik yondashuvlar kelajakda o'zbek tilida zamonaviy, keng qamrovli va xalqaro standartlarga mos korpus shakllanishiga zamin yaratadi. Bunday korpus esa nafaqat ilmiy tadqiqotlar, balki raqamli ta'lim, nutqni avtomatik qayta ishlash tizimlari va logopedik amaliyotlar uchun mustahkam manba bo'lib xizmat qiladi.

FOYDALANILGAN ADABIYOTLAR :

1. Abdurahmonova N.Z. Korpus lingvistikasi. – GlobeEdit, 2023
2. Abdurakhmonova N., Urdishev K. (2019). Corpus based teaching Uzbek as a foreign language. Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL), 6(1-2019)
3. Abdurakhmonova N., Tuliyeu U., Ismailov A., Abduvahobo, G. (2022). Uzbek electronic corpus as a tool for linguistic analysis. In компьютерная обработка тюркских языков. Turklang 2022

4. MacWhinney B. (2000). "The CHILDES Project: Tools for Analyzing Talk".
5. MacWhinney B. (2024). "Tools for Analyzing Talk , Part 2: The CLAN Program"
6. MacWhinney B. (2024). "Tools for Analyzing Talk, Part 1: The CHAT Transcription Format"
7. <https://youtu.be/KtLlR4K28oQ?feature=shared>
8. <https://talkbank.org/childes/access/>