O‘ZBEK TILI BOLALAR NUTQI AUDIO KORPUSINI YARATISH

Masharipova Baxtigul Rajabboy qizi
baxtigulmasharipova94@gmail.com
UrDU magistranti
Qutlimuratova Barno Xursandbekovna
kutlimuratovab0712@urdu.uz
UrDU o‘qituvchisi

Annotatsiya. O‘zbek tili uchun bolalar nutqi korpusini yaratish O‘zbekistonda tilni qayta ishlash texnologiyalari va ta’lim resurslarini takomillashtirish yo‘lidagi muhim qadamdir. Ushbu loyiha keng qamrovli va vakillik korpusini yaratish uchun turli yosh guruhlari, dialektlari va mintaqalaridagi bolalarning nutq ma’lumotlarini to‘plash, tahlil qilish va tartibga solishga qaratilgan. Bolalar so‘zlashadigan o‘zbek tilining o‘ziga xos fonetik, leksik va sintaktik xususiyatlariga moslashtirilgan nutqni aniqlash va sintez qilish tizimlari, ta’lim vositalari va lingvistik tadqiqotlarni rivojlantirishni qo‘llab-quvvatlash asosiy vazifalardan iborat.

Abstract. Creating a children's speech corpus for the Uzbek language is a pivotal step toward enhancing language processing technologies and educational resources in Uzbekistan. This project aims to collect, analyze, and organize speech data from children across various age groups, dialects, and regions to create a comprehensive and representative corpus. The primary objectives are to support the development of speech recognition and synthesis systems, educational tools, and linguistic research, specifically tailored to the unique phonetic, lexical, and syntactic characteristics of the Uzbek language as spoken by children.

Аннотация. Создание детского речевого корпуса на узбекском языке является важнейшим шагом на пути совершенствования технологий обработки языка и образовательных ресурсов в Узбекистане. Целью этого проекта является сбор, анализ и систематизация речевых данных детей разных возрастных групп, диалектов и регионов для создания комплексного и репрезентативного корпуса. Основными задачами являются поддержка разработки систем распознавания и синтеза речи, образовательных инструментов и лингвистических исследований, специально адаптированных к уникальным фонетическим, лексическим и синтаксическим характеристикам узбекского языка, на котором говорят дети.

Kalit so‘zlar: *NLP, Nutq korpusi, bolalar nutqi, O‘zbek tili, korpus tahlili.*

Kirish.

Til texnologiyalarini rivojlantirish, ayniqsa kam vakil tillar uchun, til xilma-xilligini saqlash va inklyuziv ta’lim va muloqotni rivojlantirish uchun muhim ahamiyatga ega. Markaziy Osiyoda millionlab odamlar so‘zlashadigan o‘zbek tili

turli yosh guruhlari va mintaqalarda farq qiluvchi o‘ziga xos xususiyatlarga ega. Bolalar nutqi, ayniqsa, kattalar nutqidan sezilarli darajada farq qiladigan aniq fonetik, leksik va sintaktik xususiyatlarni namoyon qiladi. Buni anglagan holda, o‘zbek tili uchun bolalar nutqi korpusini yaratish hal qiluvchi ahamiyatga ega bo‘ladi.

O‘zbek tili uchun maxsus tayyorlangan bolalar nutqi korpusining joriy etilishi lingvistik resurslar va texnologiyalarni rivojlantirishdagi jiddiy bo‘shliqni bartaraf etadi. Hozirgi til texnologiyalari asosan yaxshi hujjatlashtirilgan tillarga xizmat qiladi va o‘zbek tili kabi tillar uchun resurslarda bo‘sh joy qoldiradi. Ushbu korpus nafaqat o‘zbek tilida so‘zlashuvchilarga mo‘ljallangan ilg‘or ta’lim vositalari va ilovalarni ishlab chiqishga yordam beradi, balki nutqni aniqlash va sintez qilish tizimlarini ham yaxshilaydi va ularni bolalar uchun yanada qulayroq va samaraliroq qiladi.

Bundan tashqari, bu tashabbus faqat texnologiyani rivojlantirish bilan bog‘liq emas; madaniy o‘ziga xoslik va merosni tarbiyalash haqida. Bolalarning o‘zbek tilida so‘zlashuv usullarini, jumladan, mintaqaviy shevalar va so‘zlashuv iboralarini hujjatlashtirish orqali korpus tilshunoslik tadqiqotlari va saqlash ishlari uchun qimmatli manba bo‘lib xizmat qiladi. Bu tilshunoslar va pedagoglarga yosh so‘zlovchilar orasida til o‘zlashtirish va evolyutsiyani tushunishga yordam beradi, o‘zbek tilining lingvistik kelajagi haqida tushuncha beradi.

Shu nuqtayi nazardan, o‘zbek tili uchun bolalar nutqi korpusini joriy etish texnologik tafovutni bartaraf etish, lingvistik tadqiqotlarni boyitish, o‘zbek tilida so‘zlashuvchi yoshlarning ta’lim va kommunikativ ehtiyojlarini qo‘llab-quvvatlashni va’da qiladigan istiqbolli ishdir. Ushbu loyiha orqali biz til texnologiyalari inklyuziv, qulay va O‘zbekiston aholisining til xilma-xilligini ifodalovchi kelajak uchun zamin yaratishni maqsad qilganmiz.

Nutq korpusini yig‘ish

Bolalar nutq korpusini yig‘ish jarayoni ma’lumotlarning har tomonlama, vakolatli va sifatli bo‘lishini ta’minlash uchun bir necha tizimli bosqichlarni o‘z ichiga oladi. Umuman olganda, bu bosqichlar rejalashtirish, ma’lumotlarni yig‘ish, qayta ishlash va tekshirishni o‘z ichiga oladi. Bizning yondashuvimizda biz yig‘ish jarayonini soddalashtirish va yaxshilash uchun texnologiyani integratsiyalash orqali, xususan, bolalar nutqi ma’lumotlarini to‘plashni osonlashtirish uchun mo‘ljallangan Telegram botini yaratish orqali innovatsiyalar qildik.

Rejalashtirish bosqichi

Dastlabki bosqich bolalarning yosh doirasi, qamrab olinadigan lahjalar va nutqiy o‘zaro ta’sir turlari (masalan, suhbat, hikoya yoki ko‘rsatma nutqi) kabi korpusning ko‘lami va maqsadlarini aniqlashni o‘z ichiga oladi. Shuningdek, u axloqiy me’yorlarni belgilash va vasiylardan zarur roziliklarni olish, ishtirokchilarning shaxsiy hayoti va himoyasini ta’minlashni o‘z ichiga oladi.

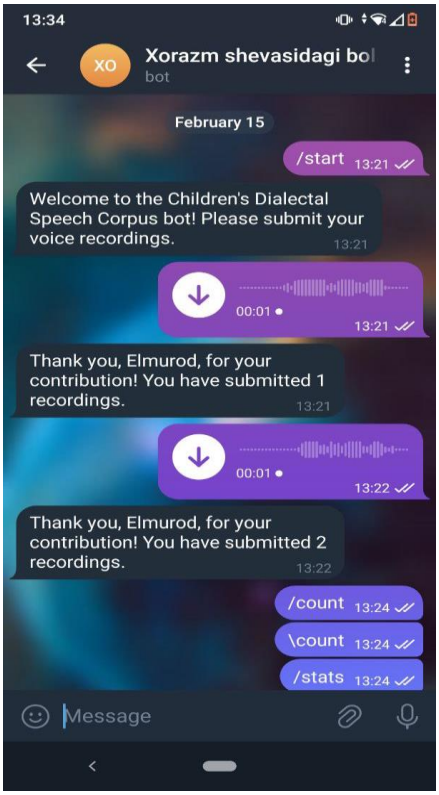
Ma’lumotlar yig‘ish

An’anaga ko‘ra, ma’lumotlarni yig‘ish maktablar yoki tadqiqot markazlari kabi boshqariladigan muhitda sessiyalar tashkil qilishni o‘z ichiga oladi. Biroq, kengroq demografiyaga erishish va jarayonni soddalashtirish uchun biz Telegram botidan foydalangan holda yangi yondashuvni qo‘lladik. Bu bot ota-onalar va vasiylar bilan muloqot qilish uchun mo‘ljallangan bo‘lib, ular o‘z farzandlarining nutqini bevosita platforma orqali yozib olishlari va yuklashlari mumkin. Bu usul tanish muhitda tabiiy nutqni to‘plash, bolalarning haqiqiyligi va qulayligini oshirish imkonini beradi.

1-rasm. Nutq korpusini yig‘ish platformasi QR kodi

Telegram botdan foydalanish

Telegram boti barcha ishtirokchilarning to‘liq xabardor bo‘lishini va loyiha shartlariga rozi bo‘lishini ta’minlab, rozilik olish jarayonida foydalanuvchilarni yo‘naltirish uchun dasturlashtirilgan. Rozilik olingandan so‘ng, bot nutq namunalarini yozib olish va yuklash bo‘yicha ko‘rsatmalar beradi. U to‘plangan ma’lumotlarning xilma-xilligini ta’minlash uchun muayyan turdagi nutq yoki mavzularni taklif qilishi mumkin. Masalan, u hikoyalar, kundalik faoliyatning tavsifi yoki turli til elementlarini keltirib chiqaradigan savollar va takliflarga javoblarni so‘rashi mumkin. Quyida Telegram botining ko‘rinishi



2-rasm. Nutq korpusini yig‘ish uchun yaratilgan Telegram bot interfeysi

Birlashtirish va saqlash

Barcha yig‘ilgan nutq ma’lumotlari avtomatik ravishda markazlashtirilgan serverga yuboriladi, u yerda saqlanadi va tizimli ravishda tashkil etiladi. Bu server boshqa omillar qatori yosh, dialekt va nutq turiga qarab ma’lumotlarni toifalarga ajratuvchi ombor vazifasini bajaradi. Ushbu tizimli yig‘ish va saqlash keyingi bosqichlarda ma’lumotlarni samarali qayta ishlash va tahlil qilishni osonlashtiradi.

Sifat nazorati va axloqiy mulohazalar

To‘plash jarayonida yuqori ma’lumotlar sifati va axloqiy me’yorlarni saqlash muhim ahamiyatga ega. Tizim yozuvlarning ravshanligi va foydalanish qulayligini tekshirish mexanizmlarini o‘z ichiga oladi va ishtirokchilarning maxfiyligini himoya qilish uchun barcha ma’lumotlarning anonim va xavfsiz saqlanishini ta’minlaydi.

Bolalar nutqi korpusini to‘plash uchun Telegram botidan foydalanish orqali biz O‘zbekiston bo‘ylab turli demografik guruhlardan keng ko‘lamli nutq namunalarini samarali to‘plashimiz mumkin. Bunday yondashuv nafaqat to‘plash jarayonini soddalashtiribgina qolmay, balki turli sheva va nutq turlarining ifodalanishini ta’minlab, pirovardida o‘zbek tili uchun boy va qimmatli lingvistik manba bo‘lishiga hissa qo‘shadi.

Xulosa

O‘zbek tili bo‘yicha bolalar nutqi korpusini yaratish tashabbusi lingvistik resurslarni ko‘paytirish va tilni qayta ishlashda texnologik yutuqlarni qo‘llab-quvvatlash yo‘lidagi muhim qadamdir. Bu sa’y-harakatlar nafaqat yosh

so‘zlovchilar uchun moslashtirilgan o‘quv va texnologik vositalarni ishlab chiqishda, balki o‘zbek tilining lingvistik rang-barangligini saqlashda ham muhim rol o‘ynaydi.

Kelajakdagi ish sifatida loyiha O‘zbekistonning turli mintaqalari va shevalari bo‘yicha yanada xilma-xil nutq namunalarini to‘plash orqali o‘z ma’lumotlar bazasini kengaytirishni maqsad qilgan. Ushbu kengayish yanada mustahkam mashina o‘rganish va sun’iy intellekt modellarini yaratish imkonini beradi, nutqni aniqlash va bolalar nutqi uchun sintez texnologiyalarini yaxshilaydi. Yakuniy maqsad bu yutuqlarni ta’lim platformalari va interfaol ilovalarga integratsiya qilish, shu tariqa O‘zbekistondagi bolalar uchun ta’lim tajribasini va texnologik foydalanish imkoniyatlarini boyitishdir.

Foydalanilgan adabiyotlar:

1. Salaev, U. I., Kuriyozov, E. R., & Matlatipov, G. R. (2023, November). Design and Implementation of a Tool for Extracting Uzbek Syllables. In *2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE)* (pp. 1750-1755). IEEE.

2. Madatov, K. (2019). A prolog format of uzbek WordNet’s entries. Human Language Technology as a Challenge for Computer Science and Linguistics, 316-320.

3. Allaberdiev, B., Matlatipov, G., Kuriyozov, E., & Rakhmonov, Z. (2024). Parallel texts dataset for Uzbek-Kazakh machine translation. Data in Brief, 110194.

4. Kutlimuratova, B., Kuriyozov, E., & Tillaeva, M. (2022). Teaching english as a foreign language for primary school children: literature reveiw. Foreign language teaching and applied linguistics, 161-171.

5. Sharipov, M., Salaev, U., & Matlatipov, G. (2021). O‘zbek tili fe’l so‘z turkumi uchun chekli avtomatlar asosida stemming algoritmini yaratish. Computer linguistics: problems, solutions, prospects, 1(1).

6. Kutlimuratova, B. (2021). Uzbek students learning English as a foreign language: Error analysis using corpora.

7. Sharipov, M., & Salaev, U. (2022). Uzbek affix finite state machine for stemming. arXiv preprint arXiv:2205.10078.

8. Madatov, K., Bekchanov, S., & Vičič, J. (2023). Uzbek text summarization based on TF-IDF. arXiv preprint arXiv:2303.00461.