ISSN 2181-922X

UZBEKISTAN LANGUAGE & CULTURE

O‘ZBEKISTON TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2024 Vol. 3 (6)

www.compling.tsuull.uz

ISSN 2181-922X

O‘ZBEKISTON

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2024 Vol. 3 (6)

compling.tsuull.uz

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti

Jurnal haqida ma’lumot

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi – Oliy attestatsiya komissiyasi ilmiy nashrlar ro‘yxatidagi “O‘zbekiston: til va madaniyat” akademik jurnalining ilovasi hisoblanib, unda professor-o‘qituvchilar, doktorantlar, stajor-tadqiqotchilar, mustaqil izlanuvchilar, magistrantlarning kompyuter lingvistikasi, jumladan, tabiiy tilga ishlov berish (NLP), o‘zbek tilining formal grammatikasi, korpus lingvistikasi, mashina tarjimasi, nutqni qayta ishlash tizimlari, intellektual tizimlar, kompyuter leksikografiyasi hamda lingvistik ontologiyalar kabi sohalarga oid tadqiqotlari nashr qilinadi.

Jurnal ilovasi bir yilda to‘rt marta chop etiladi.

O‘zbek, turk, rus va ingliz tillarida yozilgan maqolalar qabul qilinadi.

Jurnalda kitoblarga yozilgan taqrizlar, adabiyotlar sharhi, konferensiyalar hisobotlari va tadqiqot loyihalari natijalari ham e‘lon qilinadi.

Mualliflar fikri tahririyat nuqtayi nazaridan farq qilishi mumkin.

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi 2023-yildan chiqa boshlagan.

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti. O‘zbekiston, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko‘chasi, 103-uy.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

Information about the magazine

"Uzbekistan: language and culture. "Computer Linguistics" series is an appendix of the academic journal "Uzbekistan: Language and Culture" in the list of scientific publications of the Higher Attestation Commission, in which computer linguistics, including natural language processing (NLP) of professors-teachers, doctoral students, intern-researchers, independent researchers, master's students, researches related to formal grammar of the Uzbek language, corpus linguistics, machine translation, speech processing systems, intelligent systems, computer lexicography and linguistic ontologies are published.

The magazine supplement is published four times a year.

Articles written in Uzbek, Turkish, Russian and English languages are accepted.

The journal also publishes book reviews, literature reviews, conference reports, and research project results.

The opinion of the authors may differ from the editorial point of view.

"Uzbekistan: language and culture. "Computer Linguistics" series has been published since 2023.

Tashkent State University of Uzbek Language and Literature named after Alisher Navoi. Yusuf Khos Hajib street, 103, Yakkasaray district, Tashkent, Uzbekistan.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

MUNDARIJA

CONTENT

ISSN 2181-922X 115-125

AUDIOMATNLARNI KORPUSGA KIRITISH MUAMMOLARI TAHLILI

Mavluda Urazaliyeva[1]

Annotatsiya. Mazkur maqolada audiomatnlarning dialektlarni saqlash va lingvistik jihatdan tahlil qilishdagi o‘rni tadqiq qilingan. Dialektal xususiyatlarni saqlash texnologik jarayonlar orqali madaniy merosni asrash va lingvistik tahlillarni boyitishga xizmat qilishi ta’kidlangan. Zamonaviy texnologiyalar, xususan, Wav2Vec 2.0 va multimodal korpuslar yordamida dialektal talaffuz va grammatik o‘zgarishlarni tahlil qilish mumkinligi ko‘rsatilgan. Audiomatnlarni korpusga kiritishda duch kelinadigan texnologik va metodologik qiyinchiliklar ushbu jarayonni murakkablashtiradi. Bu maqolada audiomatnlarni lingvistik korpusga kiritishda duch keladigan texnik va lingvistik muammolarni yoritish maqsad qilingan.

Kalit so‘zlar: *audiomatnlar, dialektlar, lingvistik tahlil, madaniy meros, texnologiyalar, Wav2Vec 2.0.*

Kirish

Audiomatnlarni lingvistik tadqiqotlar uchun ishlatishning muhim afzalliklaridan biri ularning yuqori aniqlikdagi lingvistik ma’lumotlarni taqdim etishidir. Bunday ma’lumotlar, xususan, tilda mavjud talaffuz, urg‘u va intonatsiya farqlarini chuqur tahlil qilish imkonini beradi. Bundan tashqari, audiomatnlar tabiiy nutq namunalarini qayd etishga imkon berib, tilshunoslikda ilg‘or yondashuvlar uchun asos yaratadi.

Tilshunoslik va tabiiy tilni qayta ishlash sohalarida audiomatnlar tilning dinamikasi, dialektal tafovutlar va talaffuz o‘zgarishlarini tahlil qilishda yangi imkoniyatlar yaratmoqda. Biroq, audiomatnlarni korpusga kiritish murakkab jarayon bo‘lib, texnologik va lingvistik muammolarni o‘z ichiga oladi [Rastorgueva, 2018; Norqobilov, 2021]. ASR tizimlari nutqni matnga aylantirishda

[1]*Urazaliyeva Mavluda Yangiboyevna – O’zbekiston Milliy universiteti. Til nazariyasi. Amaliy va kompyuter lingvistikasi mustaqil izlanuvchisi.*
E-pochta: urazaliyeva_m@nuu.uz
RG: https://www.researchgate.net/profile/Mavluda-Urazaliyeva

samarali bo‘lsa-da, aksent va dialektal tafovutlar aniqlikni pasaytiradi. Shu sababli audiomatnlarni lingvistik xususiyatlarga moslashtirish muhim ahamiyatga ega.

Asosiy qism

Audiomatnlarni qayta ishlash jarayoni nafaqat texnologik masalalar, balki ijtimoiy va madaniy omillarni ham hisobga olishi kerak. Masalan, o‘zbek tilida mavjud bo‘lgan dialektal farqlar turli hududlarda ma’nolar va talaffuzlarda sezilarli o‘zgarishlarga olib keladi. Toshkent shevasidagi *“qizil”* so‘zi rangni bildirsa, Farg‘ona vodiysida bu so‘z siyosiy ma’noda ishlatilishi mumkin. Bunday tafovutlar audiomatnlar korpusiga kiritilganda tahlil qilish jarayonini murakkablashtiradi.

Shuningdek, audiomatnlarni lingvistik korpusga kiritish jarayonida kontekstual tafovutlarni aniqlash muhimdir. Har bir matnning ijtimoiy va madaniy konteksti uning tahliliga ta’sir qiladi. Misol uchun, *“salom”* so‘zining intonatsiya va talaffuzi Samarqand va Toshkent hududlarida sezilarli farq qiladi. Shu sababli, maqolada ushbu jarayon bilan bog‘liq lingvistik va texnologik muammolar batafsil yoritiladi.

1. Nutqni avtomatik tanib olishning [ASR] imkoniyatlari va cheklovlari

ASR tizimlari audiomatnlarni tahlil qilishda asosiy vosita hisoblanadi. Ushbu tizimlar ovozni matnga aylantirish orqali tabiiy tilni qayta ishlash (NLP) sohasida keng qo‘llaniladi. Google Speech-to-Text va Mozilla DeepSpeech kabi mashhur platformalar turli tildagi nutqni tahlil qilish va uni matnga aylantirishda yetakchi o‘rin tutadi [Kudryashov, 2020]. Ammo bu tizimlar universal bo‘lmaganligi sababli, ularning ishlash samaradorligi turli tillar va dialektlar bo‘yicha sezilarli darajada farqlanadi. ASR tizimlari duch keladigan muammolarni ko‘rib chiqish va ularga yechim topish tilshunoslar va dasturchilar uchun ustuvor yo‘nalishlardan biridir.

1-jadval. ASR tizimlaridagi asosiy muammolar quyidagilardan iborat:

Muammo turi	**Misol**	Taklif qilingan yechimlar
Dialektal o‘zgarishlar	O‘zbek tilidagi *“qizil”* so‘zi Toshkentda *“meva rangi”*ni bildirsa, Farg‘onada *“siyosiy qarash”* ni anglatadi.	Dialektal tahlil uchun Wav2Vec 2.0 kabi chuqur o‘rganish modellaridan foydalanish.

Aksentlar va talaffuz	Ingliz tilida *“car”* so‘zi amerika aksentida [kɑːr], Britaniya aksentida esa [kɒː] talaffuz qilinadi.	Aksentlar bo‘yicha maxsus moslashtirilgan akustik modellar yaratish.
Shovqinli muhit	Bozor kabi shovqinli joylarda yozilgan nutqni tanib olishda xatoliklarning yuqori darajasi.	Shovqin filtrlash algoritmlari va mustahkam akustik modellarni rivojlantirish.

2. Dialektal o‘zgarishlar

ASR tizimlarining dialektlarni to‘g‘ri farqlash imkoniyati cheklangan bo‘lishi mumkin. Masalan, O‘zbek tilida *“yig‘lamoq”* so‘zi Toshkentda aniq talaffuz qilinsa, Andijon va Qashqadaryo kabi hududlarda bu so‘z qisqaroq va o‘zgacha talaffuz qilinadi. Dialektal tafovutlar ko‘pincha mintaqaviy xususiyatlarga bog‘liq bo‘lib, ular tizimning aniqligiga ta’sir qiladi. Ushbu muammoni hal qilish uchun Wav2Vec 2.0 kabi chuqur o‘rganish texnologiyalari samarali hisoblanadi. Bu texnologiyalar katta hajmdagi nutq ma’lumotlarini o‘rganib, dialektal va fonetik xususiyatlarni aniqlashda yordam beradi [Baevski et al., 2020].

Dialektal tahlil uchun quyidagi usullar qo‘llaniladi:

– *Korpus yaratish*: Turli dialektal hududlardan yig‘ilgan ovozli ma’lumotlar asosida maxsus korpuslarni shakllantirish.

– *Dialekt farqlovchi modellarni ishlab chiqish*: Transformer va neyron tarmoqlari yordamida nutqdagi dialektlarni aniqlash texnologiyalarini yaratish.

– *Intonatsion va fonetik xususiyatlarni o‘rganish*: Hududiy talaffuzdagi farqlarni tahlil qilish orqali tizimlarni optimallashtirish.

Misol sifatida, o‘zbek tilidagi *“kelayapti”* so‘zi Toshkentda aniq talaffuz qilinadi, Qashqadaryo hududida esa bu *“kelyapti”* shakliga qisqaradi. Shunday xususiyatlarni inobatga oluvchi tizimlar yaratish audiomatnlarni aniqlik bilan tahlil qilish imkonini beradi.

ASR tizimlari tabiiy tilni qayta ishlash sohasida muhim o‘rin tutsa-da, ularning imkoniyatlari turli cheklovlarga duch kelmoqda. Dialektal farqlar, aksent va talaffuzdagi o‘zgarishlar, shuningdek, shovqinli muhit kabi omillar tizimlarning aniqligiga salbiy ta’sir ko‘rsatadi. Ushbu muammolarni hal qilish uchun chuqur o‘rganish texnologiyalari, shovqin filtrlash algoritmlari va kontekstual tahlil usullaridan foydalanish zarur. O‘zbek tilining dialektal va fonetik xususiyatlarini hisobga olgan holda maxsus ASR tizimlarini yaratish

ushbu tilni raqamlashtirish va global texnologiyalarga integratsiya qilishda muhim qadamdir.

3. Lingvistik va kontekstual murakkabliklar

Audiomatnlarda so‘zlarning semantik yuklamasi ko‘pincha hududiy, ijtimoiy va madaniy omillar bilan bog‘liq bo‘ladi. Har bir hududda so‘zlarning ma’nosi yoki ishlatilish konteksti o‘zgarishi mumkin, bu esa audiomatnlarni lingvistik jihatdan tahlil qilishni qiyinlashtiradi. Masalan, *“qora”* so‘zi Toshkentda *“rangi qora”*ni anglatadi, ammo Qashqadaryo viloyatida bu so‘z *“noqonuniy”* yoki *“qorong‘i”* ma’nosini bildirishi mumkin [Norqobilov, 2022; Vasilyeva, 2019]. Bu kabi tafovutlar mintaqalararo muloqotda tushunmovchiliklar keltirib chiqarishi yoki noto‘g‘ri transkribatsiyaga olib kelishi mumkin.

2-jadval. Quyidagi jadvalda semantik tafovutlarning ayrim misollari keltirilgan:

Hudud	*So‘z*	*Toshkentdagi ma’nosi*	*Boshqa hududlardagi ma’nosi*
Toshkent	*qora*	Rangi qora	Qorong‘i yoki noqonuniy
Farg‘ona	*aylan*	Doira shaklida aylanish	Ko‘chib yurmoq
Andijon	*uzum*	Meva turi	Ko‘p ma’noda ishlatiladi

Semantik tafovutlar nafaqat hududiy farqlarga, balki ijtimoiy qatlamlar va yosh guruhlariga ham bog‘liq bo‘lishi mumkin. Misol uchun, yosh avlod nutqida *“like qilish”* yoki *“post tashlash”* kabi iboralar tez-tez ishlatilsa, keksa avlod bu iboralarni kam tushunadi [Kudryashov, 2020]. Bunday vaziyatlarda semantik tahlil modellaridan foydalanish, xususan, BERT yoki RoBERTa kabi transformer asosidagi texnologiyalar, so‘zlarning ma’nosini kontekstga mos ravishda tahlil qilishga yordam beradi.

Semantik tafovutlarni kamaytirishda taklif qilinishi mumkin bo‘lgan texnologik yondashuvlar:

Korpus yaratish: Har bir hududdagi semantik tafovutlarni aniqlash uchun maxsus korpuslar yaratish [Baevski et al., 2020].

Modellarni moslashtirish: Transformer texnologiyalarini semantik tahlilga moslashtirish [Williams, Brown, Smith, 2019].

Hududiy va ijtimoiy omillarni tahlil qilish: Ma’lumotlarni hudud va ijtimoiy qatlamlarga bo‘lib o‘rganish.

Semantik tafovutlar tilning mintaqaviy boyligini aks ettiradi, lekin ushbu tafovutlar transkribatsiya va lingvistik tahlil jarayonida

qiyinchilik tug‘diradi. Shuning uchun, har bir hudud uchun alohida moslashtirilgan modellar ishlab chiqilishi zarur.

4. Pragmatik tahlil

Pragmatik tahlil audiomatnlardagi kontekstual ma’lumotlarni aniqlash uchun muhimdir. Pragmatika nutq intonatsiyasi, urg‘u va pauzalar orqali ma’noni tushunishga yordam beradi. Masalan, “salom” so‘zi Toshkentda tez va qisqa aytilib, rasman yoki norasmiy holatni bildirsa, Samarqandda bu so‘z uzoq intonatsiya bilan aytilib, hurmat yoki samimiylikni anglatadi [Ivanov, 2021]. Pragmatik tahlil intonatsion xususiyatlarni hisobga olish orqali nutqning hissiy ohangini va ma’nosini to‘g‘ri aniqlash imkonini beradi.

3-jadval. Quyidagi jadval pragmatik tafovutlarning ayrim misollarini keltiradi:

Hudud	So‘z	Intonatsiya farqlari	Ma’no ta’siri
Toshkent	*Salom*	Tez va qisqa aytiladi	Rasman yoki norasmiy holatni bildiradi.
Samarqand	*Salom*	Uzoq intonatsiya bilan aytiladi	Hurmat yoki samimiylikni anglatadi.

Pragmatik tahlilni avtomatlashtirish uchun quyidagi yondashuvlardan foydalanish mumkin:

Intonatsion tahlil modellarini ishlab chiqish: Transformer modellarini intonatsion xususiyatlarni aniqlashga moslashtirish [Park, 2018].

Ma’lumotlarni annotatsiyalash: Nutqdagi urg‘u va pauzalarni belgilash uchun maxsus annotatsiyalar kiritish [Schmidt, 2020].

Hissiy tahlilni integratsiya qilish: Nutqning hissiy tarkibini o‘rganish orqali pragmatik xususiyatlarni aniqlash.

Pragmatik tahlil nafaqat lingvistik, balki ijtimoiy va madaniy kontekstlarni tushunishda ham muhimdir. Misol uchun, o‘zbek tilida “aka” so‘zi birodarni bildiradi, lekin intonatsiya va kontekstga qarab bu so‘z hurmat ifodasi sifatida ham ishlatilishi mumkin. Ushbu ma’nolarni to‘g‘ri tahlil qilish uchun yuqori aniqlikdagi modellar va keng qamrovli korpuslar zarur [Kibrik, 2017].

Lingvistik va kontekstual murakkabliklar audiomatnlarni tahlil qilish jarayonida asosiy muammolardan biridir. Semantik va pragmatik tafovutlarni o‘rganish uchun zamonaviy texnologiyalar va modellarni rivojlantirish muhim ahamiyat kasb etadi. Har bir hudud uchun maxsus moslashtirilgan tahlil vositalarini ishlab chiqish, lingvistik tahlilning aniqligini oshiradi va madaniy boylikni saqlashga

yordam beradi. Transformer texnologiyalari va intonatsion tahlil usullari bu yo‘nalishda samarali natijalar berishi mumkin.

5. Dialektlarning saqlanishi va tahlili

Audiomatnlar dialektlarni saqlash va ularni lingvistik jihatdan tahlil qilish uchun asosiy vositadir. Dialektal farqlarni saqlab qolish orqali nafaqat madaniy merosni asrash, balki lingvistik izlanishlar uchun muhim manbalarni yaratish mumkin. Dialektlarning saqlanishi tilning rivojlanish jarayonlarini kuzatishda muhim ahamiyat kasb etadi. Misol uchun, Toshkent shevasidagi “kelayapti” va Qashqadaryo shevasidagi “kelyapti” farqlari nafaqat fonetik, balki grammatik o‘zgarishlarni ham aks ettiradi [Norqobilov, 2022; Williams, Brown, Smith, 2019].

Dialektlarni saqlash bo‘yicha amalga oshirilayotgan tadbirlar quyidagi asosiy yo‘nalishlarni o‘z ichiga oladi:

4-jadval.

Asosiy yo‘nalish	Amalga oshirilayotgan jarayonlar	Misollar
Arxivlash	Dialektlarning fonetik va morfologik xususiyatlarini saqlash.	Toshkent shevasidagi “kelayapti” va Qashqadaryo shevasidagi “kelyapti” farqi.
Madaniy merosni saqlash	Audiomatnlar xalq og‘zaki ijodining dostonlar, maqollar va ertaklar kabi namunalarini tahlil qilish.	Alisher Navoiy asarlarining dialekt bo‘yicha tahlili [Kudryashov, 2020].
Tilni rivojlantirish	Dialektlarning zamonaviy o‘zgarishlarini tahlil qilish va ularni texnologik jarayonlarga integratsiya qilish.	Hududiy dialektlarda yangi paydo bo‘lgan iboralarni qayd etish.

Dialektal xususiyatlarni saqlash va tahlil qilish tilshunoslik uchun yangi imkoniyatlar yaratadi. Hozirgi vaqtda dialektlarni saqlash uchun raqamlashtirish jarayonlari keng qo‘llanilmoqda. Audiomatnlar bu jarayonning asosiy komponenti hisoblanadi, chunki ular orqali nutqning tabiiy namunalari qayd etiladi va saqlanadi [Baevski et al., 2020; Schmidt, 2020].

Dialektal xususiyatlarni saqlash va tahlil qilish uchun texnologik yondashuvlar:

Audiokorpus yaratish: Hududiy shevalarning ovozli namunalarini yig‘ish va ularni fonetik xususiyatlariga qarab

tasniflash. Masalan, Toshkent va Andijon shevalaridagi talaffuz farqlarini qayd etuvchi maxsus korpus yaratish;

Neyron tarmoqlarni moslashtirish: Dialektlarni tahlil qilish uchun Wav2Vec 2.0 kabi chuqur o‘rganish modellaridan foydalanish [Baevski et al., 2020];

Multimodal ma’lumotlarni integratsiya qilish: Nutq intonatsiyasi, urg‘u va semantik yuklarni lingvistik tahlilga kiritish [Schmidt, 2020].

5-jadval. “kel” fe’lining turli dialektal talaffuzlari:

Hudud	Talaffuz	Ma’no
Toshkent	kelayapti	oddiy holatda jarayon davom etayotganini bildiradi.
Qashqadaryo	kelyapti	talaffuz qisqaroq, fonetik ixchamlikka ega.
Andijon	kelyaptu	dialektga xos intonatsiya va semantik urg‘u bilan boyitilgan.

Audiomatnlar xalq og‘zaki ijodi namunalarini, jumladan, maqollar, dostonlar va ertaklarni saqlashda muhim rol o‘ynaydi. Ushbu jarayon orqali hududiy dialektlarda saqlanib qolgan madaniy boyliklar tahlil qilinadi va kelgusi avlodlarga yetkaziladi. Masalan, Alisher Navoiy asarlarida ishlatilgan o‘zbek shevalari turli hududlarda farqli tarzda o‘qiladi va tushuniladi [Norqobilov, 2022]. Masalan, Samarqand shevasidagi “boraylikmi?” va Toshkent shevasidagi “boramizmi?” iboralari nafaqat fonetik, balki sintaktik o‘zgarishlarni ham aks ettiradi. Ushbu o‘zgarishlarni tahlil qilish uchun korpus texnologiyalaridan foydalanish talab etiladi [Schmidt, 2020].

Dialektlarni texnologik jarayonlarga integratsiya qilish nafaqat lingvistik tadqiqotlar uchun, balki zamonaviy texnologiyalarni rivojlantirish uchun ham muhimdir. Misol uchun, ovozli yordamchilar [virtual assistantlar] hududiy dialektlarni aniqlash va ularni qayta ishlash imkoniyatiga ega bo‘lishi kerak. Bu esa foydalanuvchilarga qulay interfeys va yuqori aniqlikdagi xizmatlarni taqdim etadi [Williams, Brown, Smith, 2019].

6. Kelajakdagi tadqiqot yo‘nalishlari

Avtomatik dialekt aniqlash: O‘zbekiston hududlaridagi 10 dan ortiq dialektni aniqlash uchun maxsus algoritmlar ishlab chiqish. Ushbu texnologiyalar dialektal farqlarni avtomatik qayd qilish va ularni tahlil qilish imkonini beradi [Baevski et al., 2020].

Lingvistik dinamikani kuzatish: Internetda trendga aylangan so‘z va iboralarni tahlil qilish orqali til evolyutsiyasini kuzatish.

Masalan, “like qilish” yoki “post tashlash” kabi iboralarning paydo bo‘lishi va yoshlar nutqida qanday rivojlanayotganini aniqlash [Williams, Brown, Smith, 2019].

Tabiiy talaffuzni sintez qilish: Tacotron 2 texnologiyasidan foydalanib, ovozli chiqishlarni yanada tabiiy qilish. Ushbu texnologiya hududiy talaffuzni o‘z ichiga olgan ovozli yordamchilarni yaratish uchun ishlatiladi [Kudryashov, 2020].

Korpusni boyitish: Turli hududiy sheva namunalarini korpuslarga kiritish va ularni annotatsiyalash jarayonlarini avtomatlashtirish. Bu esa lingvistik tahlillarni kengaytirish va tilning rivojlanish dinamikasini kuzatishga yordam beradi [Schmidt, 2020].

Xulosa

Dialektlarni saqlash va ularni lingvistik tahlil qilish madaniy merosni asrash va tilni zamonaviy texnologiyalarga integratsiya qilish uchun zarurdir. Audiomatnlar dialektal xususiyatlarni qayd etish va ularni tahlil qilish uchun samarali vositadir. Zamonaviy texnologiyalar, xususan, chuqur o‘rganish modellaridan foydalanish dialektlarni tahlil qilish va rivojlantirishda muhim ahamiyat kasb etadi. Raqamlashtirish va korpus yaratish jarayonlari esa tilshunoslik tadqiqotlarini yangi bosqichga olib chiqadi.

Audiomatnlarni korpusga kiritish bo‘yicha olib borilayotgan tadqiqotlar lingvistik tahlil va tabiiy tilni qayta ishlash texnologiyalarining rivojlanishiga xizmat qiladi. ASR tizimlarining dialektal farqlarni aniqlash qobiliyatini oshirish, audiomatnlardan madaniy merosni saqlashda foydalanish va past resursli tillar uchun moslashtirilgan modellar yaratish muhim yo‘nalishlardan biridir. Shu bilan birga, audiomatnlardagi talaffuz farqlarining tahlili orqali tilning tarixiy va madaniy rivojlanishi haqida kengroq tasavvurga ega bo‘lish mumkin.

Foydalanilgan adabiyotlar

Rastorgueva, T. A. 2018. Корпусная лингвистика и её прикладные исследования. Москва: Наука.

Norqobilov, Sh. 2021. O‘zbek tili dialektlarining semantik tahlili. O‘zbek tilshunosligi jurnali, 3(1), 45-56.

Kudryashov, B. И. 2020. Нейронные сети в системах автоматического распознавания речи ASR. Computational linguistics review, 12(4), 89-102.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. [2020]. Wav2Vec 2.0: Self-supervised learning for speech recognition. Advances in neural information processing systems, 33, 12449-12460.

Yusupov, A. 2022. O‘zbek tili dialektlarini o‘rganishda audiomatnlardan foydalanish. Toshkent davlat universiteti nashriyoti.

Williams, J., Brown, S., & Smith, A. 2019. Cross-linguistic challenges in asr systems. Speech communication, 114, 34-47.

Vasilyeva, N. 2019. Linguistic analysis in the context of digital humanities. Russian linguistics journal, 46(2), 78-93.

Schmidt, T. 2020. Multimodal corpora and their applications. International journal of corpus linguistics, 25(3), 195-210.

Ivanov, P. 2021. ASR technology adaptation for low-resource languages. Proceedings of inter speech, 45-50.

Abduraxmonova, N. Z. 2018. Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) il dis. aftoref.

Abdurakhmonova, N. 2016. The bases of automatic morphological analysis for machine translation. Izvestiya Kyrgyzskogo gosudarstvennogo tekhnicheskogo universiteta, 2(38), 12-7.

Suleymanov, D., Nevzorova, O., Gatiatullin, A., Gilmullin, R., & Khakimov, B. 2013. National corpus of the Tatar language “Tugan Tel”: grammatical annotation and implementation. Procedia-Social and Behavioral Sciences, 95, 68-74.

Abdurakhmonova, N., Tuliyev, U., & Gatiatullin, A. 2021, November. Linguistic functionality of Uzbek Electron Corpus: uzbekcorpus. uz. In 2021 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-4). IEEE.

Park, D. 2018. Contextual variations in speech recognition. Ieee transactions on audio, speech, and language processing, 27(4), 331-342.

Kibrik, A. A. 2017. Pragmatics in linguistic research. Moscow: linguistic studies.

Kim, Y., & Kang, H. 2021. Deep learning models for dialect recognition. Journal of computational linguistics, 35(7), 102-115.

Tanaka, H. (2020). Semantic differences in multilingual corpora. Multilingual studies review, 15(5), 67-82.

Johnson, R., & Blake, S. 2022. Advancements in asr for regional dialects. Language technology journal, 19(1), 1-20.

Müller, K. 2021. Comparative linguistics and corpus studies. Berlin: springer verlag.

ANALYSIS OF PROBLEMS IN INCORPORATING AUDIO TEXTS INTO A CORPUS

Mavluda Urazaliyeva[1]

Abstract. This article examines the role of audiotexts in preserving and linguistically analyzing dialects. It emphasizes that preserving dialectal features through technological processes contributes to the conservation of cultural heritage and enriches linguistic analyses. Modern technologies, such as Wav2Vec 2.0 and multimodal corpora, are highlighted for their ability to analyze dialectal pronunciations and grammatical changes. However, incorporating audiotexts into linguistic corpora is complicated by technological and methodological challenges. This article aims to address the technical and linguistic issues encountered during the process of integrating audiotexts into linguistic corpora.

Keywords: *audiotexts, dialects, linguistic analysis, cultural heritage, technologies, Wav2Vec 2.0.*

References

Rastorgueva, T. A. 2018. Korpusnaya lingvistika i yeyo prikladnie issledovaniya. Moskva: Nauka.

Norqobilov, Sh. 2021. O‘zbek tili dialektlarining semantik tahlili. O‘zbek tilshunosligi jurnali, 3(1), 45-56.

Kudryashov, V. I. 2020. Neyronnie seti v sistemax avtomaticheskogo raspoznavaniya rechi ASR. Computational linguistics review, 12(4), 89-102.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. [2020]. Wav2Vec 2.0: Self-supervised learning for speech recognition. Advances in neural information processing systems, 33, 12449-12460.

Yusupov, A. 2022. O‘zbek tili dialektlarini o‘rganishda audiomatnlardan foydalanish. Toshkent davlat universiteti nashriyoti.

Williams, J., Brown, S., & Smith, A. 2019. Cross-linguistic challenges in asr systems. Speech communication, 114, 34-47.

Vasilyeva, N. 2019. Linguistic analysis in the context of digital humanities. Russian linguistics journal, 46(2), 78-93.

[1]*Urazaliyeva Mavluda Yangiboyevna – National University of Uzbekistan. The theory of language. Independent researcher in applied and computational linguistics.*
E-pochta: urazaliyeva_m@nuu.uz
RG: https://www.researchgate.net/profile/Mavluda-Urazaliyeva

Schmidt, T. 2020. Multimodal corpora and their applications. International journal of corpus linguistics, 25(3), 195-210.

Ivanov, P. 2021. ASR technology adaptation for low-resource languages. Proceedings of inter speech, 45-50.

Abduraxmonova, N. Z. 2018. Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) il dis. aftoref.

Abdurakhmonova, N. 2016. The bases of automatic morphological analysis for machine translation. Izvestiya Kyrgyzskogo gosudarstvennogo tekhnicheskogo universiteta, 2(38), 12-7.

Suleymanov, D., Nevzorova, O., Gatiatullin, A., Gilmullin, R., & Khakimov, B. 2013. National corpus of the Tatar language "Tugan Tel": grammatical annotation and implementation. Procedia-Social and Behavioral Sciences, 95, 68-74.

Abdurakhmonova, N., Tuliyev, U., & Gatiatullin, A. 2021, November. Linguistic functionality of Uzbek Electron Corpus: uzbekcorpus. uz. In 2021 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-4). IEEE.

Park, D. 2018. Contextual variations in speech recognition. Ieee transactions on audio, speech, and language processing, 27(4), 331-342.

Kibrik, A. A. 2017. Pragmatics in linguistic research. Moscow: linguistic studies.

Kim, Y., & Kang, H. 2021. Deep learning models for dialect recognition. Journal of computational linguistics, 35(7), 102-115.

Tanaka, H. (2020). Semantic differences in multilingual corpora. Multilingual studies review, 15(5), 67-82.

Johnson, R., & Blake, S. 2022. Advancements in asr for regional dialects. Language technology journal, 19(1), 1-20.

Müller, K. 2021. Comparative linguistics and corpus studies. Berlin: springer verlag.

Jurnal 2017-yil 26-oktyabrda O‘zbekiston Respublikasi Matbuot
va axborot agentligi tomonidan 0936-raqam bilan ro‘yxatdan o‘tgan.

Jurnal O'zbekiston Respublikasi Oliy Attestatsiya Komissiyasi tomonidan
filologiya fanlari bo'yicha falsafa doktori (PhD) va fan doktori (DSc)
dissertatsiyalari asosiy ilmiy natijalari chop etilishi lozim bo'lgan ro'yxatga
kiritilgan (30.10.2021. № 308/6).

Tahririyatga kelgan maqolalar mualliflarga qaytarilmaydi.

Manzil: Toshkent shahri, Yakkasaroy tumani, Yusuf Xos
Hojib ko‘chasi 103-uy.
Telefonlar: +99871 281-45-11, +99871 281-41-93.
Website: compling.tsuull.uz
E-mail: kompling@navoiy-uni.uz

Bosishga **.**.****-yilda ruxsat etildi.
Bichimi 70x100 1/16, Ofset bosma. “Cambria” garniturasi.
Shartli b.t. 7,51. Nashr b.t. 7,62.

“O‘zbekiston: til va madaniyat” jurnali tahririyatida
tayyorlandi va sahifalandi.
“YASHNOBOD NASHR” bosmaxonasida chop etildi.
Adadi 300 nusxa. Buyurtma №2.
Bosmaxona manzili: Toshkent shahar Yashnobod tumani,
58-a harbiy shaharcha.