ТАТУ ХАБАРЛАРИ
ВЕСТНИК ТУИТ
TUIT BULLETIN

MUHAMMAD AL-XORAZMIY NOMIDAGI TOSHKENT AXBOROT TEXNOLOGIYALARI UNIVERSITETINING
ILMIY-TEXNIKA VA AXBOROT-TAHLILIY JURNALI

JURNAL 2007 YILDA TASHKIL ETILGAN.
BIR YILDA TO‘RT MARTA NASHR QILINADI

1 (73)/2025

https://uzjurnal.uz

«TATU xabarlari» jurnali («Вестник ТУИТ», “TUIT Bulletin”) O‘zbekiston matbuot va axborot agentligida 2007 yil 22 yanvarda 0204 - son bilan ro‘yxatdan o‘tgan.
O‘zR OAK tomonidan doktorlik dissertatsiyalari yuzasidan ilmiy maqolalar chop etilishi lozim bo‘lgan ilmiy jurnallar ro‘yxatiga kiritilgan
(2008 yil 2 yanvardagi 001-I-sonli buyruq).
Tahririyat manzili:
100202, Toshkent sh., Amir Temur ko‘chasi, Ф408-xona. Tel.: (+99871) 207-59-41
E-mail: tuit_xabar@tuit.uz
Qo‘lyozmalar taqrizlanmaydi va qaytarilmaydi.

TOSHKENT - 2025

UO‘K 81.322

OG‘ZAKI NUTQNI TANIB OLISH TIZIMIDA MATNDAGI TINISH BELGILARINI TIKLASH USULLARI

Elov B.B.

Tinish belgilarini tiklash masalasi og‘zaki nutqni avtomatik aniqlash (ASR) tizimlarida keng uchraydigan muammolardan biri bo‘lib, avtomatik transkriptlarda tinish belgilarining yo‘qligi matn tushunarliligini pasaytiradi. Mazkur tadqiqotda ilk bor o‘zbek tili uchun ASR tizimi chiqishidagi matnlarda tinish belgilarini avtomatik tiklash vazifasi ko‘rib chiqilgan bo‘lib, ushbu yo‘nalishda o‘zbek tilida hozirga qadar deyarli tadqiqotlar olib borilmagani qayd etiladi. Ushbu maqsadda chuqur o‘rganishga asoslangan PR (punctuation restoration) modeli ishlab chiqilib, UzbPunct korpusi misolida sinovdan o‘tkazildi. UzbPunct ma’lumotlar to‘plami 32 000 dan ortiq matnlar va taxminan 36 soatlik ovozli ma’lumotlarni o‘z ichiga oladi, jami 105 nafar spiker (51 erkak, 54 ayol) ishtirok etgan. Korpus o‘quv (1000 ta yozuv) va test (274 ta yozuv) bo‘limlariga bo‘lingan bo‘lib, norasmiy suhbat matnlari (UzbTalks, ~20%) va rasmiy axborot matnlari (UzbNews, ~80%) kabi ikki turli bo‘limni o‘z ichiga oladi. Taklif etilayotgan yechim doirasida tinish belgilarini bashorat qilish uchun zamonaviy chuqur neyron tarmoqli yondashuvlar qo‘llanildi: LSTM, ikki yo‘nalishli BLSTM, transformer arxitekturasi hamda oldindan o‘qitilgan BERT modeli yordamida turli usullar sinovdan o‘tkazildi. Model matndagi tinish belgi joylashuvlarini aniqroq aniqlash uchun faqat lingvistik (matn) ma’lumotlargagina emas, balki nutqdan olingan pauza kabi prozodik xususiyatlarga ham tayanadi. Punktuatsiyani bashorat qiluvchi modellarning ishlash samaradorligi aniqlik (precision), eslab qolish (recall) va F1 ko‘rsatkichlari orqali baholandi, natijalar har bir tinish belgisi uchun alohida tahlil qilindi. Eksperimentlar natijasi ishlab chiqilgan model o‘zbek tilidagi ASR transkriptlarida tinish belgilarini yuqori aniqlikda tiklashga erishishini ko‘rsatdi – modelning umumiy F1 bahosi taxminan 80 % ga teng. Avtomatik tinish belgilarini tiklash ASR yordamida olingan matnlarning o‘qilishini sezilarli darajada yaxshilab, ularni tabiiy yozma ko‘rinishga keltiradi.

Kalit so‘zlar: Tinish belgilarini tiklash, punctuation restoration, Automatic Speech Recognition, ASR, og‘zaki nutqni tanib olish, nutq transkriptlari, UzbPunct ma’lumotlar to‘plami.

Задача восстановления знаков препинания в расшифровках, получаемых системами автоматического распознавания речи (ASR), является распространенной проблемой, поскольку отсутствие пунктуации в автоматически распознанном тексте затрудняет его понимание и снижает читаемость. В настоящей работе впервые рассматривается автоматическое

восстановление пунктуации для узбекского языка, поскольку ранее подобные исследования для узбекской речи практически не проводились. Для решения этой задачи разработана модель восстановления пунктуации (PR) на основе глубинного обучения, которая обучена и протестирована на корпусе UzbPunct. Корпус UzbPunct содержит более 32 000 текстовых предложений и около 36 часов аудиозаписей речи общим количеством 105 дикторов (51 мужчина и 54 женщины). Данные разделены на обучающую выборку (1000 аудиозаписей) и тестовую выборку (274 записи), а также на две тематические части: UzbTalks (разговорный жанр, ~20%) и UzbNews (новостной стиль, ~80%), что охватывает как неформальный диалоговый, так и информационный контекст. В рамках исследования были рассмотрены современные методы восстановления пунктуации с помощью нейронных сетей, включая рекуррентные модели LSTM (в том числе двунаправленные BLSTM), архитектуры на базе Transformer, а также предобученную языковую модель BERT. Предлагаемый подход учитывает не только текстовые, но и просодические признаки устной речи (например, длительности пауз) для более точного прогнозирования расположения знаков препинания. Эффективность модели оценивается по метрикам точности (precision), полноты (recall) и F1-меры для каждого вида знаков препинания. Экспериментальные результаты показывают, что предложенный метод обеспечивает высокое качество восстановления пунктуации в узбекских ASR-транскриптах – суммарное значение F1 достигает порядка 80 %. Автоматическое добавление знаков препинания значительно повышает читаемость распознанного текста, приближая его по форме к нормативно написанному.

Ключевые слова: восстановление знаков препинания, punctuation restoration, автоматическое распознавание речи, ASR, распознавание устной речи, транскрипты речи, набор данных UzbPunct.

Punctuation restoration in automatic speech recognition (ASR) transcripts is a common natural language processing challenge, as the absence of punctuation in recognized text greatly reduces readability and hinders understanding. This paper presents, for the first time, a study on automatic punctuation restoration for the Uzbek language, addressing a gap since such research on Uzbek speech has been virtually nonexistent to date. We develop a deep learning–based punctuation restoration (PR) model and evaluate it on the UzbPunct dataset of Uzbek speech transcripts. The UzbPunct corpus contains over 32,000 sentences of text and approximately 36 hours of speech audio from 105 speakers (51 male and 54 female). The data are split into a training set of 1000 recordings and a test set of 274 recordings, and divided into two sections: UzbTalks (conversational content, ~20%) and UzbNews (news content, ~80%), covering both informal dialogue and formal informational contexts. We explore state-of-the-art neural network approaches for punctuation restoration, including Long Short-Term Memory (LSTM) networks

(with bidirectional BLSTM), transformer-based architectures, and a pre-trained BERT model. The proposed method leverages both textual features and prosodic cues from speech (such as pause durations) to more accurately predict punctuation marks. Model performance is evaluated using precision, recall, and F1-score metrics for each punctuation category. Experimental results demonstrate that the model achieves a high overall F1-score on the order of 80% in restoring punctuation for Uzbek ASR outputs. This automatic punctuation restoration notably improves the readability of Uzbek speech transcripts, making the transcribed text much closer to well-formed written language.

Keywords: punctuation restoration, Automatic Speech Recognition, ASR, spoken language recognition, speech transcripts, UzbPunct dataset.

I. KIRISH

Bugungi kunda raqamli yozma muloqotning aniq va sifatliligi juda muhim. Norasmiy chatlarda qatnashish yoki professional hujjatlarni tayyorlash, fikrimizni samarali yetkazish tilimizning aniqligiga tayanadi. An’anaviy imlo va grammatik tekshiruvlar xatolarni aniqlash uchun qimmatli vosita bo‘lgan bo‘lsa-da, ular ko‘pincha kontekstni tushunishga nisbatan ko‘proq narsani talab qiladi. Ushbu cheklov tabiiy tilni qayta ishlash (Natural Language Processing, NLP) vositalaridan foydalanadigan yanada ilg‘or yechimlarga yo‘l ochdi. Ushbu maqolada NLP vositalaridan foydalangan holda zamonaviy imlo va grammatika tekshirgichni ishlab chiqishni ko‘rib chiqamiz. Taklif qilinayotgan yechimning an’anaviy qoidalarga asoslangan tizimlardan ustunligini va raqamli aloqa asrida foydalanuvchi tajribasini yanada qulayroq taqdim etishilishi qayt etish lozim.

Raqamli aloqaning tobora ortib borayotgan ahamiyati yozma matnning ravshanligi va aniqligiga katta ahamiyat berdi. Onlayn suhbatlardan tortib professional yozishmalargacha o‘z fikrimizni samarali ifoda etishimiz tabiiy tilimizning aniqligi bilan chambarchas bog‘liq. An’anaviy imlo va grammatik tekshiruv usullari uzoq vaqtdan beri xatolarni aniqlash va tuzatishda qimmatli vosita bo‘lib kelgan, ammo ularni kontekstual tushunish va moslashish cheklovlari hali hal qilingan emas. Bu esa tabiiy til bilan bog‘liq vazifalarga yanada kengroq yondashuvni taklif qiluvchi NLPning yanada ilg‘or yechimlarni ishlab chiqishga turtki bo‘ldi.

Tabiiy tilni qayta ishlash – bu kompyuterlarga inson tilini qayta ishlash va tushunish imkonini berish uchun tilshunoslik, informatika va sun'iy intellekt tajribasini birlashtirgan fanlararo soha. NLP vositalaridan foydalangan holda, bizning imlo va grammatika tekshirgichimiz foydalanuvchilarga aniqroq va kontekstga bog‘liq xatolarni aniqlash va tuzatish imkoniyatini taqdim etadi. NLPga asoslangan tekshiruv ilovalari matndagi imlo va grammatik xatolarni aniqlaydi va aniqroq tuzatishlar va takliflar berish uchun **kontekst, sintaksis** va **semantikani tahlil qiladi**.

So‘nggi bir necha yil ichida Google Home, Amazon Echo, Siri va Cortana kabi **ovozli yordamchilar (Voice Assistants, VA)**dan keng miqyosida foydalanilmoqda. Ushbu ilovalar asosini og‘zaki nutqni avtomatik aniqlash (Automatic Speech Recognition, ASR) NLP vazifasi tashkil etadi. Yuqorida keltirilgan NLP ilovalarida audio formatidagi ma’lumotlardan matnli fotmatdagi ma’lumotlar shakllantiriladi. Shu sababli bu turdagi algoritmlar **Speech-to-Text algorithms** deb ham ataladi. Google Home va Siri kabi NLP ilovalari nafaqat audiodan matnni aniqlaydi, balki aytilgan so‘zning semantik ma’nosini ham izohlaydi va tushunadi, tahlil qiladi hamda javob qaytaradi. Shuning uchun ular savollarga javob berishlari yoki foydalanuvchi buyruqlari asosida harakatlarni amalga oshirishlari mumkin.

Inson nutqi bizning kundalik shaxsiy va ish hayotimiz uchun asos bo‘lib, og‘zaki nutqni matnga o‘tkazish funksiyasi juda ko‘p ilovalarga ega. Ushbu turdagi NLP ilovalaridan *mijozlarni qo‘llab-quvvatlash* yoki *savdo qo‘ng‘iroqlari mazmunini transkripsiya qilish*, *ovozli chatbotlar* uchun yoki *uchrashuvlar* va *boshqa muhokamalar mazmunini yozib olish* uchun foydalanish mumkin.

Audio ma’lumotlar *tovushlar* va *shovqinlardan* iborat. Inson nutqi o‘ziga xos holatdir. Audio ma’lumotlarni qayta ishlash uchun ular *raqamlashtiriladi* va *spektrogrammalarga aylantiriladi*. Biroq, og‘zaki nutq murakkabroq, chunki u tabiiy tilni kodlaydi. Audio ma’lumotlarni tasniflash kabi muammolar audio klipdan boshlanadi va berilgan sinflar to‘plamidan bu audio qaysi sinfga tegishli ekanligini bashorat qiladi.

Og‘zaki nutqni tanib olish (Automatic Speech Recognition, ASR) tizimlari tomonidan shakllantirilgan *nutq transkriptlarida (Speech transcripts)* odatda *tinish belgilari* yoki *bosh harflar* mavjud emas. Avtomatik tanib olingan og‘zaki nutqning fragmentlarida tinish belgilarining yo‘qligi sababli matnining tushunish jarayoniga ta'sir qiladi [1]. Tabiiy tilni qayta ishlash (Natural Language Processing, NLP) vazifalari hisoblangan *tinish belgilari (punctuation restoration, PR)* va bosh harflarni tiklash *(capitalization restoration, CR)*ning asosiy maqsadi ASR tomonidan yaratilgan tinish belgilarisiz matnlarning o‘qilishini yaxshilashdir.

PR va CR hal qilish masalasidan *nomlangan obyektlarni aniqlash (Named Entity Recognition, NER)* [2]*, POS teglash (Part-of-Speech, POS)*, *semantik tahlil qilish* va og‘zaki dialogni segmentlash [3] kabi NLP vazifalari samaradorligini oshirishi mumkin. Tabiiy tildagi og‘zaki matn transkriptlarini tahlil qilishga asoslangan PR tizimini baholash murakkab bo‘lib, asosan, tinish belgilarining qoidalari dastlab yozilgan matnlar uchun ham noaniq bo‘lishi mumkinligi va tabiiy ravishda yuzaga keladigan og‘zaki nutqning so‘z birikmalari va gaplar chegaralarini aniq belgilashni qiyinlashtiradi [4,5].

II. ASOSIY QISM

ASR tizimi audio ma’lumotlardan tinish belgilarisiz uzluksiz so‘zlar ketma-ketligini natija sifatida taqdim etadi. Bu esa insonning o‘qish qobiliyatini va ASR

matnida tabiiy tilni qayta ishlash vazifalarining bajarilishini pasaytiradi. Tinish belgilarini bashorat qilish vazifasini oldindan o‘qitilgan BERT metodidan foydalanadigan arxitekturani K.Makhija va T.N.Chng (2019) [6] taklif qilganlar. Taklif etilgtan BERT modeli IWSLT ma’lumotlar to‘plamidagi holatini sezilarli darajada yaxshilagan, nutqa, vergul va so‘roq belgisini birgalikda bashorat qilish bo‘yicha umumiy F1-score 81.4% ko‘rsatkichga teng bo‘lgan.

Long Short-Term Memory (LSTM) neyron tarmoqqa asoslangan metod asosida matndagi tinish belgilarni bashorat qilish usulini K.Xu, L.Xie, va K.Yao (2016) [7] taqdim etishgan. Taklif etilayotgan LSTM modeli bir nechta qatlamlar orqali kirish xususiyatlari va chiqish yorliqlari o‘rtasidagi bog‘liqlikni modellashtiradi. Shuningdek, ular chiqish belgilari o‘rtasidagi bog‘liqlikni modellashtirishning ta'sirini empirik tarzda o‘rganganlar. Olingan natijalar shuni ko‘rsatadiki, chuqur ikki yo‘nalishli LSTM metodidan foydalanish tinish belgilarini bashorat qilishda eng zamonaviy ko‘rsatkichlarga erishishi qayd etilgan.

X.Wu, S.Zhu, Y.Wu va K.Yu (2016) [8] multi-view LSTM modeli asosida katta hajmdagi til korpusi matnlaridagi tinish belgilarni bashoratlashni amalga oshirdilar. Tajribalar shuni ko‘rsatdiki, LSTM an’anaviy CRF metodiga asoslangan modeldan sezilarli darajada ustundir.

Xitoy tili harflari boshqa tillaridan ancha farq qiladi. X.Liu, Y.Liu va X.Song (2018) [9] xitoy tili korpusidagi tinish belgilarini bashorat qilish uchun uzoq qisqa muddatli xotiraga (LSTM) asoslangan uch bosqichli tizimni joriy qilishgan. Ular leksik xususiyatlar va pauza vaqti o‘rtasidagi kombinatsiyaga asoslanib, tinish belgilarni bashorat qilish uchun leksik va pauza vaqtiga ohang ma’lumotlarini birlashtirishgan. Bu uchta leksik, pauza vaqti va tovush balandligi kabi multimodal xususiyatdan foydalanadigan usul bo‘lib, ma’lumotlar to‘plami bo‘yicha eksperimental natijalar tavsiya etilgan modelning samaradorligi va ustunligini uch bosqichli takroriy neyron tarmog‘idan foydalangan holda ko‘rsatilgan.

ASR tizimlarini turli sohalardagi amaliy masalarni hal qilishda qo‘llash mumkin. ASR mavjud texnologiyalarning foydaliligi va samaradorligini sezilarli darajada yaxshilaydi, lekin odatda natija sifatida tinish belgilarisiz faqat so‘zlar ketma-ketligi chiqadi. Bu esa foydalanuvchi maqsadini aniqlashda noaniqlikka olib kelishi mumkin. A.Silva, B.J.Theobald va N.Apostoloff (2021) [10] ASR tizimlarini maishiy elektronika sohasida qo‘llaganlar. Ular birinchi navbatda tinish belgilarini bashorat qilish uchun transformerga asoslangan yondashuvdan foydalanib IWSLT 2012 TED topshirig‘ini 8% ga yaxshilashga erisganlar.

Hozirda ko‘plab tinish belgilarini bashorat qilish usullari asosan standart toza ma’lumotlarga o‘qitiladi va ASR tizimida transkripsiya xatolari bilan umumlashtirishni yo‘qotadi. Toza o‘quv ma’lumotlari va shovqinli test ma’lumotlari o‘rtasidagi bo‘shliqni bartaraf etish uchun A.Zheng, N.Ye, X.Wang va X.Song (2020) [11] uchta tasodifiy (three random, 3R) ma’lumotlarni ko‘paytirish strategiyasini taklif qilishgan: tasodifiy so‘zlarni o‘chirish (random word deletion, RWD), tasodifiy so‘zlarni almashtirish (random word substitution, RWS) va

tasodifiy fonemalarni tahrirlash (random phoneme edition, RPE). Ular akustik jihatdan o‘xshash so‘zlarni almashtirish uchun fonema darajasidagi o‘xshash so‘zlar lug‘atni taqdim etishgan. Shuningdek, RoBERTa – katta modelini tinish belgilarini bashorat qilish metodi vositasida tildagi semantika va uzoq masofaga bog‘liqliklar qo‘lga kiritilgan.

M.Fang, H.Zhao, X.Song va X.Wang (2019) [12] BLSTM va BERT metodlarini birlashtirib, xitoycha matndagi tinish belgilarini bashorat qilishga asoslangan usulni taklif qilishgan. Bu esa BERT metodidan kontekstli so‘zlarni o‘rganish uchun matn kodlash qatlamlari sifatida foydalanadi. BLSTM tarmog‘i samaradorligini yaxshilash uchun Recurrent Neural Network (RNN) asosidagi oldingi tinish belgilarini bashorat qilish usullari bilan solishtirganda, taklif qilingan usulning tinish belgilarini bashorat qilish samaradorligini xitoycha segmentlanmagan matnda semantika va uzoq masofaga bog‘liqliklarni olishning samaradorligini oshirgan. Xitoy yangiliklari ma’lumotlar to‘plami bo‘yicha eksperimental natijalar shuni ko‘rsatadiki, BERT-BLSTM asosidagi usul umumiy mikro-F1 bo‘yicha mutlaq 31,07% ga asosiy ko‘rsatkichdan ustunligi qayd etilgan.

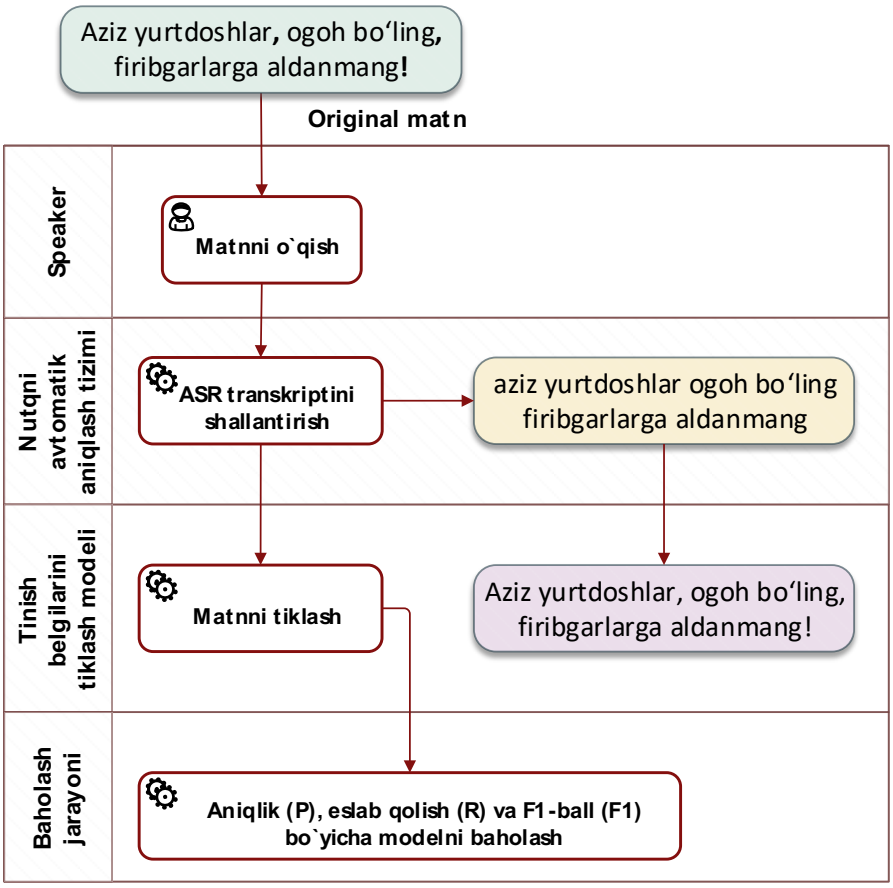
Bugungi kunda o‘zbek tilidagi matnlarda tinish belgilarni to‘g‘ri joylashtirilganligini aniqlash bo‘yicha ilmiy va amaliy tadqiqotlar deyarli olib borilmagan.

Og‘zaki nutqni tanib olish tizimi. Ushbu maqolada yoqorida keltirilgan talablar va cheklovlarni hisobga olgan holda, o‘zbek tili korpusiga asoslangan PR vazifasini hal qilish masallasini ko‘rib chiqiladi. Shu maqsadda o‘zbek tili korpusidan 450.000 so‘z shakldan iborat 1000 dan ortiq matnlar tanlab olindi. Ushbu manbalardagi mavjud tinish belgilari boshqa matnlarni baholash uchun foydalaniladi. Biroq, bu matnlarda ba'zi imloviy xatolar bo‘lishi mumkin. Ajratib olingan matnlar yuzdan ortiq turli ma'ruzachilar tomonidan o‘qildi. Tinish belgilariga ega asl matnlar ovozli yozuvlar bilan majburiy ravishda moslashtirildi va ideal ASR chiqishi sifatida ishlatildi.

ASR vazifasining maqsadi - bu topshiriq uchun jamlangan test to‘plamida tinish belgilarini tiklash uchun yechimni taqdim etishdan iborat. **Test to‘plami (test set)** korpus matnlarining moslashtirilgan ASR transkripsiyalaridan iborat. Taklif qilinayotgan yechimda tinish belgilarini aniqlash uchun *matnga asoslangan* va *og‘zaki nutqdan olingan xususiyatlardan* foydalanish tavsiya etiladi [13]. Shuningdek, **o‘quv to‘plami (train set)** sifatida o‘zbek tili korpusidan olingan matnlar ishlatiladi. Ushbu matnlarni o‘qitishda va tinish belgilarini sozlashda foydalanish mumkin.

Masalaning qo‘yilishi

Matnda tinish belgilarini joylashtirish NLP vazifasining maqsadi og‘zaki nutqni tanib olish (ASR)da tinish belgilarini tiklashdan iborat. Ushbu jarayon quyidagi 1-rasmda o‘z aksini topgan:



1. 1-rasm. Matnda tinish belgilarini joylashtirish NLP vazifasining arxitekturasi

UzbPunct ma’lumotlar to‘plami

UzbPunct – ToshDO‘TAU kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi xodimlari tomonidan o‘zbek tili korpusi matnlarining *kraudsorsingli matn va audio ma’lumotlar to‘plami (crowdsourced text and audio data set)*. Ma’lumotlar to‘plami ikki qismga bo‘lingan: **suhbat (UzbTalks)** va **axborot (UzbNews)**. Audio komponentni ishlab chiqarishda yuzdan ortiq xodimlar, ilmiy tadqiqotchilar, magistrantlar va talabalar ishtirok etdi. Audio ma’lumotlarning umumiy uzunligi (hajmi) test to‘plamini o‘z ichiga olgan holda deyarli **o‘ttiz olti soatga** yetadi. Audio ma’lumotlarni shakllantirishda erkak va ayol nisbatini muvozanatlash uchun choralar ko‘rildi. UzbPunct 32000 dan ortiq matn va 1200 ta audio faylga ega bo‘lib, ulardan 1000 tasi o‘quv to‘plamida va 200 tasi test to‘plamida joylashtirilgan. Har bir matn uchun *avtomatik tanib olingan og‘zaki nutq* va *qo‘lda tekislangan matnning* **transkripti** mavjud. Ma’lumotlar formati va baholash ko‘rsatkichlari haqidagi ma’lumotlar maqolaning keingi qismlarida keltirilgan.

UzbPunct ma’lumotlar to‘plamining statistikasi quyida keltirilgan:

– **Matn:** o‘zbek tili korpusidan olingan 32000 ta matn;

– **Audio:**

▪ *Tanlash tartibi:*

▪tasodifiy tanlangan UzbNews (80%, ya'ni o‘quv to‘plami uchun 800 ta yozuv) so‘zlar soni [150..300] oralig‘ida;

▪tasodifiy tanlangan UzbTalks (20%), so‘zlar soni [150..300] oralig‘ida va kamida bitta so‘roq belgisidan iborat gap;

▪*Ma’lumotlar to‘plamining bo‘linishi:*

▪O‘quv ma’lumotlari (Training data): 1000 ta yozuv;

▪Test ma’lumotlari (Test data): 274 ta yozuv;

▪*Spikerlar:*

▪Erkak: 51 ma’ruzachi, 16,7 soat nutq;

▪Ayol: 54 ma’ruzachi, 19 soat nutq;

Strukturlanmagan (qayta ishlanmagan) matn uchun tinish belgilari statistikasi quyidagi 1-jadvalda keltirilgan.

1-jadval

Strukturlanmagan matn uchun tinish belgilari statistikasi

Tinish belgisi	yozilishi	mean	median	max	sum	included
fullstop	**.**	12.44	7.0	1129.0	404378	yes
comma	**,**	10.97	5.0	1283.0	356678	yes
question mark	**?**	0.83	0.0	130.0	26879	yes
exclamation mark	**!**	0.22	0.0	55.0	7164	yes
hyphen	**-**	2.64	1.0	363.0	81190	yes
colon	**:**	1.49	0.0	202.0	44995	yes
ellipsis	**...**	0.27	0.0	60.0	8882	yes
semicolon	**;**	0.13	0.0	51.0	4270	no
quote	**"**	3.64	0.0	346.0	116874	no
Jami so‘zlar soni		**169.50**	**89.0**	**17252.0**	**5452032**	-

Yuqorida keltirilganidek, ma’lumotlar to‘plami ikki qismdan iborat: **suhbat (UzbTalks)** va **axborot (UzbNews)**.

1-qism. UzbTalks

Ma’lumotlar o‘zbek tili korpusidagi turtli soha va uslublarga mansub matnlardan olingan. Bu turdagi ma’lumotlarda O‘zbekiston va jahondagi itqisodiyot, jamiyat, sport, fan-texnika va ob havo haqidagi kontenlar o‘z aksini topgan. Suhbat sahifalari suhbat ma’lumotlari sifatida xizmat qiladi. Bunda, foydalanuvchilar bir-birlari bilan izoh yozish orqali muloqot qilishadi. UzbTalks ma’lumotlar to‘plamida imloviy va tinish belgi xatolari mavjud. Ushbu ma’lumotlar to‘plami *og‘zaki shakldagi ma’lumotlarning (spoken data)* 20% ni qamrab oladi.

2-qism. UzbNews

Uzbnews — bepul kontentli yangiliklar bo‘lib, ma’lumotlar kun.uz, daryo.uz, … kabi elektron nashrlardan olingan. Sayt hamkorlikdagi jurnalistika orqali ishlaydi. Uzbnews ma’lumotlar to‘plamining sifati yuqori bo‘lsa-da, ba’zi matnlarda

imloviy va tinish belgilarda xatoliklar mavjud bo‘lishi mumkin. Ushbu ma’lumotlar to‘plami *og‘zaki shakldagi ma’lumotlarning* 80% ni qamrab oladi.

Ma’lumotlar formati

Kirish fayli ikki ustundan iborat bo‘lgan fayldab iborat:

1. Matn identifikatori;

2. Tinish belgilarisiz kichik harflardagi matn.

Chiqish matnidafi satrlar soni kirish faylidagi satrlar soniga teng bo‘lishi kerak bo‘lib, har bir satrda tinish belgilari bilan matn joylashadi.

Majburiy tekislangan transkripsiyalar (Forced-aligned transcriptions)

Og‘zaki nutqni tanib olish tizimining chiqish qismidagi natijani hosil qilishda asl matnlarning moslashtirilgan transkripsiyalaridan foydalaniladi. **“*.timestamp”** formatidagi fayllar asl matnni bir guruh ko‘ngillilar tomonidan o‘qilgan audio fayl bilan moslashtirilgan juftliklarini o‘z ichiga oladi. Ba’zi fayllarda matnni noto‘g‘ri o‘qish (bo‘laklarni o‘tkazib yuborish, asl matnda yetishmayotgan so‘zlarni qo‘shish) va matn hamda audio fayllar uchun moslashtirish vositasini sozlash natijasida yuzaga keladigan tekislash xatolarini o‘z ichiga olishi mumkin. Konfiguratsiya o‘zbek tiliga mo‘ljallangan bo‘lib, boshqa tillardagi neologizmlar va NERlar noto‘g‘ri aniqlanishi mumkin. Ushbu turdagi so‘zlarning davomiyligi nolga teng (boshlanish va tugash vaqt belgilari teng). Ma’lumotlar quyidagi formatda taqdim etiladi:

(timestamp_start,timestamp_end) word

…

</s>

“*.timestamp” formatidagi faylga namuna quyidagi 2-jadvalda keltirilgan:

2-jadval

“*.timestamp” formatidagi faylga namuna

transcript	transcript	transcript
(780,810) Bu	(2790,2960) ilg‘or	(7260,7510) viloyati,
(830,890) haqda	(3010,3430) hisoblanadi.	(7570,7620) Olot
(910,910) O‘zA	(3520,3670) Bu	(7680,7920) tumanida
(950,1020) xabar	(3760,4030) haqda	(8030,8420) yashovchi
(1070,1300) berdi.	(4120,4600) Shavkat	(8480,8510) 28
(1330,1360) U	(4630,5230) Mirziyoyev	(8580,8620) yoshli
(1400,1560) odamga	(5350,5530) videoselektorda	(8650,8650) K.
(1600,1820) taqlid	(5590,5890) aytib	(8680,8680) L.
(1890,2100) qilish	(6010,6010) o‘tdi.	(8700,8890) sodir
(2160,2290) nuqtai	(6880,6300) Ushbu	(8950,9220) etganligi
(2350,2570) nazaridan	(6360,6810) jinoyatni	(9280,9420) aniqlandi.
(2640,2710) eng	(6930,7200) Buxoro	</s>

Modelni baholash tartibi. Modelni baholash bir nechta bosqichda amalga oshiriladi.

Tinish belgilari

Og‘zaki nutqni tanib olish vazifasini bajarish davomida quyidagi 3-jadvalda keltirilgan tinish belgilari baholanadi:

3-jadval

Tinish belgilari va ularning belgilanishi

№	Tinish belgisi	Punctuation mark	symbol
1.	nuqta	fullstop	.
2.	vergul	comma	,
3.	so‘roq belgisi	question mark	?
4.	undov belgisi	exclamation mark	!
5.	tire	hyphen	-
6.	ikki nuqta	colon	:
7.	uch nuqta	ellipsis	...
8.	nuqtai vergul	semi-colon	;
9.	bo‘sh joy	blank (no punctuation)	

Baholash kerak bo‘lgan natija faqat tinish belgilari qo‘shilgan matndir.

Baholash ko‘rsatkichlari

Yakuniy natijalar har bir tinish belgisini alohida bashorat qilish uchun **aniqlik (precision), eslab qolish (recall)** va **F1 ballari (F1 scores)** baholash ko‘rstakichlari vositasida baholanadi. Taqdim etilgan qiymatlar har bir tinish belgisi uchun F1 ballarining o‘rtacha og‘irligiga nisbatan taqqoslanadi.

Hujjat bo‘yicha ball:

$$\textbf{Aniqlik (precision)} = \frac{TP}{TP + FP};$$

$$\textbf{Eslab qolish (recall)} = \frac{TP}{TP + FN};$$

$$\textbf{F1} = 2 * \frac{\textbf{Precision} * \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}};$$

Har bir tinish belgisi uchun global (p) ball:

$$\textbf{Precision}_P = \textbf{avg}_{micro}\textbf{Precision}_P = \frac{\sum_{d \in Documents} TP}{\sum_{d \in Documents} TP + FP};$$

$$\textbf{Recall}_P = \textbf{avg}_{micro}\textbf{Recall}_P = \frac{\sum_{d \in Documents} TP}{\sum_{d \in Documents} TP + FN};$$

Modelning yakuniy ball ko‘rsatkichi global ballarning o‘rtacha og‘irligi sifatida hisoblanadi.

$$\frac{1}{N}\sum_{p\in Punctuation} support(p) * \mathbf{avg}_{micro}\mathbf{F_1}(p)$$

I. XULOSA

An’anaviy imlo va grammatik tekshiruv usullari uzoq vaqtdan beri xatolarni aniqlash va tuzatishda qimmatli vosita bo‘lib kelgan. Bugungi kunda AI usullari vositasida matndagi imlo va grammatik xatolarni aniqlash va tuzatish bo‘ycha juda ko‘p ilmit tadqiqotlar olib borilmoqda. Jumladan, NLPga asoslangan tekshiruv ilovalari matndagi imlo va grammatik xatolarni aniqlaydi va aniqroq tuzatishlar va takliflar berish uchun kontekst, sintaksis va semantikani tahlil qiladi. Hozirda Google Home, Amazon Echo, Siri va Cortana kabi ovozli yordamchilar (Voice Assistants, VA) asosini og‘zaki nutqni avtomatik aniqlash (Automatic Speech Recognition, ASR) NLP vazifasi tashkil etadi. Ushbu NLP ilovalarida audio formatidagi ma’lumotlardan matnli fotmatdagi ma’lumotlar shakllantiriladi. Og‘zaki nutqni tanib olish (Automatic Speech Recognition, ASR) tizimlari tomonidan shakllantirilgan *nutq transkriptlarida (Speech transcripts)* odatda *tinish belgilari* yoki *bosh harflar* mavjud emas. Avtomatik tanib olingan og‘zaki nutqning fragmentlarida tinish belgilarining yo‘qligi sababli matnining tushunish jarayoniga ta'sir qiladi. Tabiiy tilni qayta ishlash (Natural Language Processing, NLP) vazifalari hisoblangan *tinish belgilari (punctuation restoration, PR)* va bosh harflarni tiklash *(capitalization restoration, CR)*ning asosiy maqsadi ASR tomonidan yaratilgan tinish belgilarisiz matnlarning o‘qilishini yaxshilashdan iborat. Ushbu maqolada matnda tinish belgilarini joylashtirish NLP vazifasining arxitekturasi keltirildi. ToshDO‘TAU kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi xodimlari tomonidan ishlab chiqilgan UzbPunct ma’lumotlar to‘plami (32000 dan ortiq matn va 1200 ta audio fayl)da ASR modeli sinovdan o‘tkazildi va ishlab chiqilgan LM baholash usullari keltirildi. Xulosa sifatida, ASR modeli samaradorligini oshirish maqsadida quyidagi omillarni hisobga olgan holda baholash ko‘rsatkichlarida muayyan o‘zgartirishlarni amalga oshrish lozim:

- *ASR va majburiy hizalama xatolari;*
- *izohlovchilar o‘rtasidagi nomuvofiqliklar;*
- *tinish belgilarining ozgina siljishi ta'siri;*
- *har xil turdagi xatolarga turli og‘irliklarni belgilash.*

II. FOYDALANILGAN ADABIYOTLAR

[1] Yi, J., Tao, J., Bai, Y., Tian, Z., & Fan, C. (2020). Adversarial transfer learning for punctuation restoration. arXiv preprint arXiv:2004.00248.

[2] Nguyen, T. B., Nguyen, Q. M., Nguyen, T. T. H., Do, Q. T., & Luong, C. M. (2020). Improving vietnamese named entity recognition from speech using word

capitalization and punctuation recovery models. *arXiv preprint arXiv:2010.00198*.

[3] Hlubík, P., Španěl, M., Boháč, M., & Weingartová, L. (2020, September). Inserting punctuation to asr output in a real-time production environment. In *International Conference on Text, Speech, and Dialogue* (pp. 418-425). Cham: Springer International Publishing.

[4] Sirts, K., & Peekman, K. (2020). Evaluating sentence segmentation and word tokenization systems on Estonian web texts. In *Human Language Technologies–The Baltic Perspective* (pp. 174-181). IOS Press.

[5] Wang, X. (2020, February). Analysis of Sentence Boundary of the Host's Spoken Language Based on Semantic Orientation Pointwise Mutual Information Algorithm. In *2020 12th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)* (pp. 501-506). IEEE.

[6] Makhija, K., Ho, T. N., & Chng, E. S. (2019, November). Transfer learning for punctuation prediction. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 268-273). IEEE.

[7] Xu, K., Xie, L., & Yao, K. (2016, October). Investigating LSTM for punctuation prediction. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)* (pp. 1-5). IEEE.

[8] Wu, X., Zhu, S., Wu, Y., & Yu, K. (2016, October). Rich punctuations prediction using large-scale deep learning. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)* (pp. 1-5). IEEE.

[9] Liu, X., Liu, Y., & Song, X. (2018, November). Investigating for punctuation prediction in Chinese speech transcriptions. In *2018 International Conference on Asian Language Processing (IALP)* (pp. 74-78). IEEE.

[10] Silva, A., Theobald, B. J., & Apostoloff, N. (2021, June). Multimodal punctuation prediction with contextual dropout. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3980-3984). IEEE.

[11] Zheng, A., Ye, N., Wang, X., & Song, X. (2020, November). 3r: Word and phoneme edition based data augmentation for lexical punctuation prediction. In *2020 16th International Conference on Computational Intelligence and Security (CIS)* (pp. 1-5). IEEE.

[12] Fang, M., Zhao, H., Song, X., Wang, X., & Huang, S. (2019, December). Using bidirectional LSTM with BERT for Chinese punctuation prediction. In *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)* (pp. 1-5). IEEE.

[13] Sunkara, M., Ronanki, S., Bekal, D., Bodapati, S., & Kirchhoff, K. (2020). Multimodal semi-supervised learning framework for punctuation prediction in conversational speech. *arxiv preprint arXiv:2008.00702*.