

https://doi.org/10.51885/3134-8025_IJCS_2026_3_5

XFTAP 16.31.21

ҚАЗАҚ ТІЛІ ҮШІН АВТОМАТТЫ СӨЙЛЕУДІ ТАҢУ ЖҮЙЕСІНДЕ ПУНКТУАЦИЯНЫ ҚАЛПЫНА КЕЛТІРУ

ВОССТАНОВЛЕНИЕ ПУНКТУАЦИИ В СИСТЕМЕ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ ДЛЯ КАЗАХСКОГО ЯЗЫКА

RESTORING PUNCTUATION IN AUTOMATIC SPEECH RECOGNITION FOR KAZAKH LANGUAGE

А.С. Кәрібаева ^{1*}, О. Мыссов ¹, А.С. Карыбаева ²

¹Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан

²Абылай хан атындағы Қазақ халықаралық қатынастар және әлем тілдері университеті,
Алматы қ., Қазақстан

*Жауапты автор: А.С. Кәрібаева, a.s.karibayeva@gmail.com

Түйінді сөздер:

Қазақ тілі, сөйлеуді
автоматты тану, тыныс
белгілері, Трансформер,
mBert, код-ауысым.

ТҮЙІНДЕМЕ

Мақалада қазақ тілін автоматты сөйлеуді тану жүйелеріндегі (ASR) тыныс белгілерін қалпына келтіру мәселесін қарастырады. Ұсынылған тәсіл қазақ тілінің ASR жүйелеріндегі тыныс белгілерін қалпына келтірудің негізгі төрт тыныс белгісі классы (үтір – COMMA, нүкте – PERIOD, сұрақ – QUESTION және леп белгісі – EXCLAMATION) және тыныс белгісі жоқ (O) класс бойынша шешуді көрсетеді. Тәжірибелер бір тілді қазақ мәтіндерінде де, аралас қазақ-орыс деректерінде де сынақтан өткізіліп, қазіргі ASR қолданбаларындағы практикалық көптілді байланыс параметрлерін қамтыды. Леп белгісі мен сұрақ белгісіндегі класс теңгерімсіздігін азайту үшін толықтыру әдісі негізінде қазақ тіліндегі тыныс белгілерін қалпына келтіруге арналған арнайы таңбаланған корпус құрылды. Тәжірибелер тыныс белгілерін қалпына келтіру моделінің төрт тыныс белгілері классы (COMMA, PERIOD, QUESTION және EXCLAMATION) бойынша моно-тілді тест жиынында 93,23 % макро-F1 көрсеткішіне жеткенін көрсетті: нүкте үшін 98,99 %, сұрақ белгісі үшін 94,35 %, үтір белгісі үшін 87,77 % және леп белгіні тану үшін 91,82 % F1 көрсеткіштері алынды. Оқыту жиынына code-switching деректерін қосу арқылы бөлек code-switching тест жиынындағы Macro-F1 32,81 %-тен 63,88 %-ке дейін өсті, дегенмен EXCLAMATION классы бойынша сапа әлі де шектеулі (26,47 % F1) болып қалады.

Ключевые слова:

Казахский язык,
автоматическое
распознавание речи, знаки
препинания, Трансформер,
mBert, переключение кода.

АННОТАЦИЯ

В статье рассматривается проблема восстановления пунктуации в системах автоматического распознавания речи (ASR) казахского языка. Предложенный подход демонстрирует решение задачи восстановления пунктуации в казахских ASR для четырех основных классов пунктуации (запятая - COMMA, точка - PERIOD, вопрос -



© 2026 А.С. Кәрібаева, О. Мыссов, А.С. Карыбаева

Бұл жұмыс Creative Commons Attribution 4.0 халықаралық лицензиясы
(CC BY 4.0) бойынша таратылады.

<https://creativecommons.org/licenses/by/4.0/>

QUESTION и восклицательный знак - EXCLAMATION) и класса без пунктуации (O). Эксперименты проводились как на монологичных казахских текстах, так и на смешанных казахско-русских данных, охватывающих практические параметры многоязычной коммуникации в современных приложениях ASR. Для уменьшения дисбаланса классов восклицательных и вопросительных знаков был создан специально размеченный корпус на основе реальных предложений из открытых казахскоязычных источников и LLM-ассистированной аугментации. Эксперименты показали, что модель восстановления пунктуации достигла показателя макро-F1 93,23 % для четырех классов пунктуации (COMMA, PERIOD, QUESTION и EXCLAMATION) на монологичном тестовом наборе, показав F1-меры 98,99 % для точки, 94,35 % для вопросительного знака, 87,77 % для запятой и 91,82 % для восклицательного знака. На отдельном тестовом наборе с казахско-русским code-switching добавление таких данных в обучение подняло Macro-F1 с 32,81 % до 63,88 %, хотя класс EXCLAMATION остается наиболее сложным (26,47 % F1).

Keywords:

Kazakh language, Automatic speech recognition, punctuation, Transformer, mBert, code-switching.

ABSTRACT

This paper considers the problem of punctuation restoration in Kazakh automatic speech recognition (ASR) systems. The proposed approach demonstrates a solution to the punctuation restoration problem in Kazakh ASR for four main punctuation classes (COMMA, PERIOD, QUESTION, and EXCLAMATION) and the unpunctuated class (O). Experiments were conducted on both monolingual Kazakh texts and mixed Kazakh-Russian data, covering the practical parameters of multilingual communication in modern ASR applications. To reduce the imbalance between the exclamation and question mark classes, a specially labeled corpus was created using both real sentences collected from open Kazakh-language sources and LLM-assisted paraphrasing. Experiments showed that the punctuation restoration model achieved a macro-F1 score of 93.23% for four punctuation classes (COMMA, PERIOD, QUESTION, and EXCLAMATION) on the monolingual test set, with F1 scores of 98.99% for period, 94.35% for question mark, 87.77% for comma, and 91.82% for exclamation mark. On a held-out code-switching test set, adding code-switched sentences to training raised macro-F1 from 32.81% to 63.88%, though the exclamation class remains the weakest (26.47% F1), pointing to prosody-free intonation cues as the main remaining challenge.

КІРІСПЕ

Автоматты сөйлеуді тану (ASR) – жасанды интеллекттің (ЖИ) тез дамып келе жатқан саласының бірі болып табылады. Қазіргі заманғы жасанды интеллект жүйелерінің ауызекі сөйлеуге байланысты жылдам транскрипциялауға, автоматты түрде субтитрлерді құруға және дауыстық интерфейстерді қолданбаларға үздіксіз интеграциялауға мүмкіндік береді. Қазіргі таңда ASR модельдерінің сапа жағынан айтарлықтай жақсы сапа көрсетуіне қарамастан, бұл жүйелер шығысында алынған транскрипциялар көбінесе лингвистикалық тұрғыдан толық емес болып қалуда. Көп жағдайда автоматты сөйлеуді танудан шыққан нәтиже әдетте тыныс белгілері мен сөз және сөйлем шекараларының болмауымен үздіксіз сөз ағыны ретінде алынууда. Бұл өз кезегінде модельді, жүйені оқыту барысына тікелей әсер береді, сонымен қатар түсінуді қиындатады. Сөйлем ішінде дұрыс тыныс белгілерінің, шекараларын дұрыс емес анықтау ол автоматты суммаризациялау, мәтіннің эмоционалды

талдауы, семантикалық талдау және машиналық аударма сияқты табиғи тілді өңдеу есептері үшін шектеулерге алып келеді.

Ауызекі сөйлеу кезінде тыныс белгілері тікелей айтылмаса да, олар кідірістер, интонациялық үлгілер, сөйлеу темпі және дауыс қарқындылығы сияқты дыбыстық ерекшеліктерде жанама түрде көрініс табады. Сондықтан тыныс белгілерін дәл қалпына келтіру лингвистикалық контекстті де, акустикалық белгілерді де модельдеуді қажет етеді. Нәтижесінде, тыныс белгілерін қалпына келтіру NLP зерттеулеріндегі маңызды және өзекті есептердің бірі болып табылады.

Қазақ тілі синтаксистік және морфологиялық тұрғыдан бай тілдерге жатады, сондықтан тыныс белгілерінің дұрыс қойылуы семантикалық тұрғыдан маңызды рөлді атқарады. Тыныс белгілеріндегі шағын айырмашылықтар сөйлемнің мағынасы мен эмоционалды реңкін өзгерте алады. Мысал ретінде, «Шын.» және «Шын? (күнделікті өмірде шылауларсыз айтылуы мүмкін)» сөйлемдерін алсақ. Бірінші жағдайда хабарлы сөйлемді білдіреді, ал «Шын?» сұрақ белгісі тұрғандықтан сөйлемді сұраулы сөйлемге айналдырды. Нүктенің сұрақ белгісімен алмастырылуы тек грамматикалық функцияны ғана емес, сонымен қатар коммуникативтік ниетті де өзгертеді. Сол сияқты, дұрыс қойылмаған үтірлер немесе сөйлемнің соңғы тыныс белгілері мағынаны бұрмалап, мазмұн мен эмоцияны дұрыс түсіндірмеуге әкелуі әбден мүмкін.

Қазақ тілінде тыныс белгілері шығыс транскрипцияда көп жүйелерде дұрыс қоймайды, яғни қазақ тілін танитын ASR жүйелерінің тыныс белгілерін қалпына келтіру мүмкіндігі шектеулі. Кейбір модельдерде, Whisper және ElevenLabs сияқты қазақ сөйлеуді тану жүйелері транскрипцияларға тыныс белгілерін автоматты түрде қояды. Бұл осы жүйелердегі тыныс белгілерін қалпына келтіру архитектураға кіріктірілгендігіне байланысты. Сонымен қатар, MMS (Massive Multilingual Speech), TurkicASR және Soyle модельдері қазақ және басқа түркі тілдері үшін тыныс белгілерінсіз автоматты түрде енгізбей сөйлеу транскрипциясын қамтамасыз етеді, яғни сөйлем басы аяғы анықталмаған. Егер ASR жүйесіне бірнеше сөйлем кіріс ретінде енгізілсе, шығыста барлығы бір ағын ретінде тыныс белгілерінсіз шығарылады. Соңғы аталған жүйелерде қазақ тіліндегі тыныс белгілерін қалпына келтіру үшін қосымша өңдеуден кейінгі модульдер немесе арнайы модельдерді құру қажеттілігі туындайды. Бұл айырмашылық қазақ тіліндегі тыныс белгілерін қалпына келтіруді ASR зерттеулерінде бөлек қарастыру қажеттілігіне алып келеді.

Қазақ тіліндегі ASR жүйелерін дамыту бағытында отандық зерттеулердің де елеулі үлесі бар. Атап айтқанда, Mamyrbayev et al. (2019) қазақ тілінің үздіксіз сөйлеуін тануға арналған жүйені ұсынса, ал Mamyrbayev et al. (2020) аглютинативті тілдер үшін end-to-end сөйлеуді тану моделін зерттеді. Кейінгі жұмыстарда Mamyrbayev et al. (2022) гибриді end-to-end архитектураны, ал Orken et al. (2022) трансформер негізіндегі end-to-end сөйлеуді тану жүйесін қазақ тілі үшін бейімдеп, жоғары дәлдік көрсетті. Дегенмен, аталған зерттеулердің басым бөлігі тек сөйлеуді тану (WER) сапасына бағытталған және тыныс белгілерін қалпына келтіру мәселесін қарастырмайды, бұл дәл осы жұмыстың өзектілігін негіздейді.

Бүгінгі таңда тыныс белгілерін қалпына келтіру, ең алдымен, нейрондық желіге негізделген архитектураларды, әсіресе Трансформер (Vaswani et al., 2017) және BERT (Devlin et al., 2019) сияқты алдын ала оқытылған тіл модельдерін, сондай-ақ мәтіндік және акустикалық ақпаратты біріктіретін гибриді құрылымдарды пайдалану арқылы шешілуде. Бұл жүйелердің өнімділігі оқыту деректерінің сапасы мен репрезентативтілігіне, мақсатты тілдің құрылымдық сипаттамаларына және сөйлеу саласына (мысалы, әңгіме, академиялық дәрістер, сұхбаттар немесе ресми презентациялар) байланысты. Сонымен қатар, модельдің шу, аймақтық диалектілер және әртүрлі сөйлеу стильдері болған жағдайларда әлі де шешілмеген күйінде қалып отыр.

Трансформер архитектурасы негізіндегі тыныс белгілерін қалпына келтірудегі айтарлықтай жетістіктерге қарамастан, қазақ тіліндегі ASR шығыстарын кейінгі өңдеу әлі де толық зерттелмеген. Көптеген шешімдер көптілді алдын ала оқытылған модельдерге тәуелді немесе тану қателеріне шектеулі сапаны көрсетеді.

BERT тәрізді модельдерді қолдана отырып тыныс белгілерін қалпына келтіру бұрын басқа жоғары ресурсты (ағылшын, француз, неміс және т.б. аналитикалық тілдер) тілдерде қолданылғанымен, жұмыстың жаңалығы үш негізгі аспектіде: (i) оны қазақ тіліне қолдану, (ii) нақты әлемдегі көп тілді байланысты көрсететін код аралас қазақ-орыс мәтінін жүйелі бағалау және (iii) осы тілдік орта үшін арнайы жасалған арнайы аннотацияланған корпусты жасау.

Жұмыста сегменттелмеген ASR мәтініндегі тыныс белгілерін қалпына келтіру және токен деңгейіндегі контекстік тәуелділіктерді анықтау үшін дәл бапталған көптілді BERT негізіндегі архитектура моделі қолданылды. Токен деңгейіндегі тыныс белгілері схемасын жобалауға және қазақ тіліне бейімделген деректер жиынтығын құруға ерекше назар аударылады. Сонымен қатар, бұл тәсілді қазақ тіліне кириллица және латын графикасында қолдануға болады, бұл орфографиялық ауысулар кезінде және көптілді контексттерде икемділікті қамтамасыз етеді. Сондай-ақ, бұл жұмыстың тағы бір ерекшелігі - ол code-switching (адам бір әңгімеде, сөйлемде немесе тіпті сөз тіркесінде екі немесе одан да көп тілді, диалектілерді немесе лингвистикалық стильдерді кезектестіріп қолдануы) деректері бойынша модельдерді бағалау және тестілеу болып табылады.

ӘДЕБИЕТТІК ШОЛУ

Тыныс белгілерін автоматты түрде қалпына келтіру саласындағы зерттеулер жиырма жылдан астам уақыт бойы жүргізіліп келеді. Алғаш лингвистикалық зерттеулерде тыныс белгілерінің орналасуының негізгі болжаушылары ретінде просодикалық ерекшеліктерге – үзілістерге, дыбыс биіктігінің өзгеруіне және сөйлеу жылдамдығына сүйенуді ұсынған (Batliner et al., 1998).

ASR мәтіндеріндегі тыныс белгілерін қалпына келтірудің алғашқы тәсілдері максималды энтропия моделіне негізделген және тыныс белгілерін болжау үшін сөздер, сөйлеу бөліктері, кідірістер және сөйлеушінің сипаттамалары туралы ақпаратты пайдаланған (Batista et al., 2007; Batista et al., 2008). Бастапқы жұмыстар ағылшын тіліндегі жаңалықтар саласы үшін орташа нәтижелерді көрсеткен n-граммға негізделген статистикалық әдістер негізінде тыныс белгілерін анықтайтын тіл моделі негізінде шешу ұсынған (Gravano et al., 2009).

Тыныс белгілері мен бас әріптерді қалпына келтірудің тарихи дамуы мен заманауи әдістеріне жан-жақты жүйелі талдау қолданыстағы әдістердің артықшылығы мен шектеулері талданып, ресурстары аз тілдерде, дамыту үшін құнды ұсыныстар жасалынды (Păiș & Tufiş, 2022). Алғашқы зерттеулер мәселені негізінен ауызекі тіл транскрипцияларын кейінгі өңдеу аясында қарастырды. Бұл тәсілдердің негізгі шектеуі – олардың алыс синтаксистік тәуелділіктерді ескере алмауы, бұл әсіресе агглютинативті тілдер үшін өте маңызды.

Тыныс белгілерін қалыпқа келтірудегі келесі буынында қайталанатын нейрондық желілерді пайдалану арқылы шешу қарастырылды. Екі бағытты LSTM желілері ағылшын жаңалықтарының корпусында (IWSLT) 68 %-дан астам F1 мәндеріне қол жеткізіп, n-грамм модельдерінен айтарлықтай асып түсетінін көрсетті (Tilk & Alumäe, 2015; Tilk & Alumäe, 2016).

Одан кейін тыныс белгілерін болжауға арналған екі терең нейрондық желіге негізделген тізбекті белгілеу архитектурасын тыныс белгілерінің аннотациясын қамтитын корпуста оқыту қарастырылды. Екі бағытты ұзақ қысқа мерзімді жады желісі (BLSTM) және конволюциялық нейрондық желі (CNN) негізінде тыныс белгілерін болжау ұсынылған (Żelasko et al., 2018). Мұнда BLSTM моделі жалпы қателіктердің төмен деңгейін көрсеткенімен, CNN тыныс белгілерін қоюда, әсіресе «сұрақ» белгілерін анықтауда жоғары дәлдікке қол жеткізілген.

2017 жылы ұсынылған Трансформер архитектурасының жарыққа шығуымен байланысты көптеген NLP есептерін шешуде айтарлықтай жақсы нәтижелер көрсете бастады (Vaswani et al., 2017). Тыныс белгілерін қалпына келтіру тапсырмасы үшін Трансформерлердің рекуррентті нейронды желілерге қарағанда екі айқын артықшылығы бар: өзіне назар аудару механизмі қайталанатын жинақтаусыз кез келген ұзындықтағы тәуелділіктерді тікелей қарастыра алады, ал үлкен корпустарда бай контекстік көріністерді қамтамасыз етеді.

BERT моделін ағылшын тіліндегі IWSLT деректер жиынтығында оқытып, «үтір» тыныс белгісі үшін 77,3 % F1_score мәнін алған (Courtland et al., 2020). Тыныс белгілерін қалыпқа келтіруді T5 моделі көмегімен тізбекті генерациялау тапсырмасы ретінде қарастырып, ағылшын және венгр тілдері үшін нәтижелер алған (Nagy & Bial, 2021).

Трансформер архитектурасы түрік тілінде де қолданыс тапты. Авторлар нәтижесінде BERTurk-пен жалпы макро-F1 ұпайы 78,5% тыныс белгілері үшін дұрыс қойылуын алған (Alam et al., 2020). Түрік тілінде тыныс белгілерін қалпына келтіру бойынша зерттеулерде жаңа, арнайы дайындалған деректер жиынтығы пайдаланылды (Kurt & Çayır, 2023; Görmez et al., 2025). Трансформерлерді қоса алғанда жұмыста әртүрлі модельдермен тәжірибе жасаған, авторлар «түрік тілі үшін «нүкте» мен «бос орындар» үшін F-score шамамен 90%-тей көрсеткішін алған.

Сонымен қатар, тыныс белгілерін қалпына келтіру үшін токен деңгейіндегі оқыту жүйесі ұсынылды (Huang et al., 2021). Бұл әдіс модельдің контекстік жағынан ұқсас және әртүрлі токендерді ажырату қабілетін күшейтеді, осылайша тыныс белгілерін болжау дәлдігін жақсартады. Эксперименттік бағалаулар бұл тәсілдің ASR арқылы жасалған және басқа да шулы мәтіндердегі тыныс белгілерін тиімді түрде қалпына келтіретінін, дәстүрлі бақыланатын оқыту әдістерінен асып түсетінін растады.

Қазіргі таңда тыныс белгілерін қалыпқа келтіруде контекстік ендірімелер енгізу тізбектерін қосымша қолданылуда, бұл синтаксистік құрылым мен семантикалық қатынастарды дәлірек модельдеуге мүмкіндік береді. Модельдер ASR қателіктерінің әдеттегі үлгілерін көрсететін ауызекі тілдік корпустарда оқытылады, осылайша олардың транскрипция шуына төзімділігін арттырады. Эмпирикалық нәтижелер трансформерлерге негізделген архитектуралардың дәлдік пен F1 ұпайы тұрғысынан дәстүрлі әдістерден асып түсетінін көрсетеді. Бұл сөйлеуді тану жүйелерінде кейінгі өңдеу үшін терең оқыту тәсілдерінің тиімділігін көрсетеді.

Қазақ тіліндегі сөйлеуді тану саласындағы отандық зерттеулердің нәтижелерін тыныс белгілерін қалпына келтіру контекстінде салыстырмалы түрде қарастыру маңызды. Mamyrbayev et al. (2019, 2020, 2023) және Mamyrbayev et al. (2022) жұмыстарында ұсынылған end-to-end және трансформер негізіндегі архитектуралар негізінен сөз деңгейіндегі тану дәлдігін (WER, CER) жақсартуға бағытталған: соңғы аталған зерттеуде Трансформер мен CTC тілдік моделін біріктіру арқылы 3,7 % символдық қателік деңгейіне (CER) қол жеткізілген. Сол сияқты, Mussakhodzhaeva et al. (2021) қазақ, орыс және ағылшын тілдері үшін көптілді end-to-end ASR жүйесін зерттеп, деректерді толықтыру (data augmentation) әдістерінің тану сапасына әсерін бағалады. Дегенмен, аталған зерттеулердің барлығында тыныс белгілерін қалпына келтіру бөлек тапсырма ретінде қарастырылмаған, ал ASR жүйесінің шығысы тыныс белгілерсіз үздіксіз сөз ағыны түрінде қалады.

Бұған қарағанда, Umbet & Kozhimbayev (2025) жұмысы қазақ тіліндегі пунктуация мен бас әріптерді қалпына келтіруге тікелей арналған, бірақ леп белгісі сияқты сирек кластар бойынша төмен көрсеткіш ($F1 = 32,85\%$) көрсеткен. Осы жұмыста ұсынылған тәсіл дәл осы олқылықты жоюға бағытталған: деректерді мақсатты толықтыру арқылы леп белгісі бойынша F1 көрсеткішін 89,89 %-ке дейін жеткізу отандық зерттеулермен салыстырғанда айтарлықтай жетістік болып табылады.

Ал, қазақ тілі үшін буын тыныс белгілерін қалпына келтіру және бас әріптермен жазуды қалпына келтіруге арналған арнайы модель енгізілді (Umbet & Kozhirbayev, 2025). Жүйе BERT негізіндегі көптілді-кассацияланбаған архитектура негізінде жасалған және NVIDIA NeMo құрылымын пайдаланып жүзеге асырылған. Ол бас әріптермен жазу үшін 98,04 микро-F1 ұпайларымен және тыныс белгілерін болжау үшін 96,75 шамасынан жоғары өнімділікке қол жеткізілген. Дегенмен, модель сирек кездесетін таңбалар мен сирек кездесетін тілдік үлгілерді өңдеуде шектеулерін көрсеткен.

Трансформер негізіндегі тыныс белгілерін қалпына келтірудегі айтарлықтай жетістіктерге қарамастан, қазақ тіліндегі ASR шығыстарын кейінгі өңдеу әлі де толық зерттелмеген. Көптеген қолданыстағы шешімдер көптілді алдын ала дайындалған модельдерге тәуелді немесе тану қателеріне шектеулі төзімділікті көрсетеді, бұл одан әрі зерттеулер мен тілге тән оңтайландыруға мүмкіндік береді.

ЗЕРТТЕУ МАТЕРИАЛДАРЫ МЕН ӨДІСТЕРІ

Тыныс белгілерін қалпына келтіру токен деңгейіндегі тізбекті белгілеу мәселесі ретінде тұжырымдалды. Әрбір токен $x = (x_1, \dots, x_n)$ тізбектілігі үшін ASR жүйесінің тыныс белгісі жоқ шығысы үшін, модель әрбір x_i токени үшін $y_i \in L$ тыныс белгісін болжайды. Осындағы жарамды белгілер жиынтығы келесідей түрде болады (1):

$$L = \{O, COMMA, PERIOD, QUESTION, EXCLAMATION\} \quad (1)$$

y_i белгісі қалпына келтірілген мәтіндегі x_i белгісінен кейінгі тыныс белгіні білдіреді. Ал O тыныс белгісінің жоқтығын, $COMMA$ – үтір, $PERIOD$ – нүкте, $QUESTION$ – сұрақ, $EXCLAMATION$ – леп белгісін білдіреді. Тыныс белгілері олардың алдындағы сөзге тағайындалады. Бұл орнату осы тапсырма үшін стандартты болып табылады және акустикалық ерекшеліктер туралы ақпаратты қажет етпейді, модель тек ASR шығысымен, яғни мәтінмен жұмыс жасайды.

Тыныс белгілерінің модельмен дұрыс тағайындалуы үшін барлық тыныс белгілерін қамтитындай корпус құрылды. Корпус қазақ тіліндегі мәтіндердің тәуелсіз жиналған деректер жиынтығынан тұрады. Оның жалпы көлемі – 109 772 сөйлем. Деректер жиынтығы ашық қазақ тіліндегі веб-ресурстардан (жаңалықтар порталдары мен энциклопедиялық мәтіндер, соның ішінде zakon.kz, tengrinews.kz, nur.kz, ult.kz) автоматтандырылған краулинг арқылы жиналды. Жинау кезеңі 2026 жылдың ақпан-шілде айларын қамтиды. Тазалау кезеңінде HTML-белгілер мен навигациялық элементтер алынып тасталды, сөйлемдер ұзындығы (5–60 токен) бойынша сүзгіленді, қайталанатын жазбалар дедупликацияланды, ал кодтау артефактілері бар сөйлемдер (мысалы, бұзылған UTF-8 таңбалары) бас тартылды.

Сирек кездесетін тыныс белгілерін қоса алғанда, барлық тыныс белгілері кластарын жеткілікті түрде қамту үшін деректер жиынтығы екі кезеңде жасалды:

1. Негізгі корпус ашық қазақ тіліндегі ресурстардан автоматтандырылған веб-краулинг арқылы 109 772 сөйлем жиналды. Бұл жиналған мәліметтер бойынша белгіленген класстар бойынша класстардың саны бойынша тең еместігін көрсетті. Жиналған мәліметтер бойынша O класы токендердің 88,4 %-ын, $PERIOD$ 5,89 %-ын және $COMMA$ 5,24%-ын құрады, ал $QUESTION$ мен $EXCLAMATION$ бірге 0,5 %-дан құрады.

2. Сирек кездесетін кластарды толықтыру екі көзден жүргізілді. Біріншіден, $QUESTION$ және $EXCLAMATION$ белгілерімен аяқталатын 11 429 нақты сөйлем (5 393 сұраулы және 6 036 лепті) осы ашық веб-корпустан сөйлем шекарасы бойынша сүзгіленіп бөлек алынды. Екіншіден, осы нақты сөйлемдердің 6 000-ы (класс бойынша 3 000-нан) үлгі ретінде пайдаланылып, жергілікті орналастырылған Qwen3-32B тілдік моделі арқылы тақырыптық парафраз жасалды; генерацияланған сөйлемдердің сапа критерийлерінен (дұрыс тыныс белгісімен аяқталуы, 4–40 сөз аралығы, қайталанбауы) өткен 5 584-і (93%) түпкі корпусқа қосылды. Барлық генерацияланған сөйлемдер өз seed-сөйлемімен және

генерациялаған модельмен бірге бөлек тіркелді. Соңғы деректер жиынтығы 126 785 сөйлемнен тұрады (барлығы 1 976 506 токен).

Толықтырудан кейінгі корпуста келесі тыныс белгілері үлестері анықталды: О класы токендердің 88,89 %-ын, PERIOD 4,92 %-ын, COMMA 5,31 %-ын, QUESTION 0,43 %-ын және EXCLAMATION 0,46 %-ын құрады (барлығы 1 976 506 токен).

Сөйлемнің орташа ұзындығы 17 токен. Корпус әрбір деректер жиынтығындағы сценарийлердің үлесі бойынша оқытуға (80 %), валидация (10 %) және тест (10 %) жиынтықтарына бөлінді.

Модель стандартты transformer encoder мен token classification head архитектурасын қолданады. Модельдегі энкодер әрбір кіріс токенді көпқабатты трансформер арқылы өңдеп, оның контекстік векторын (hidden state) шығарады (2):

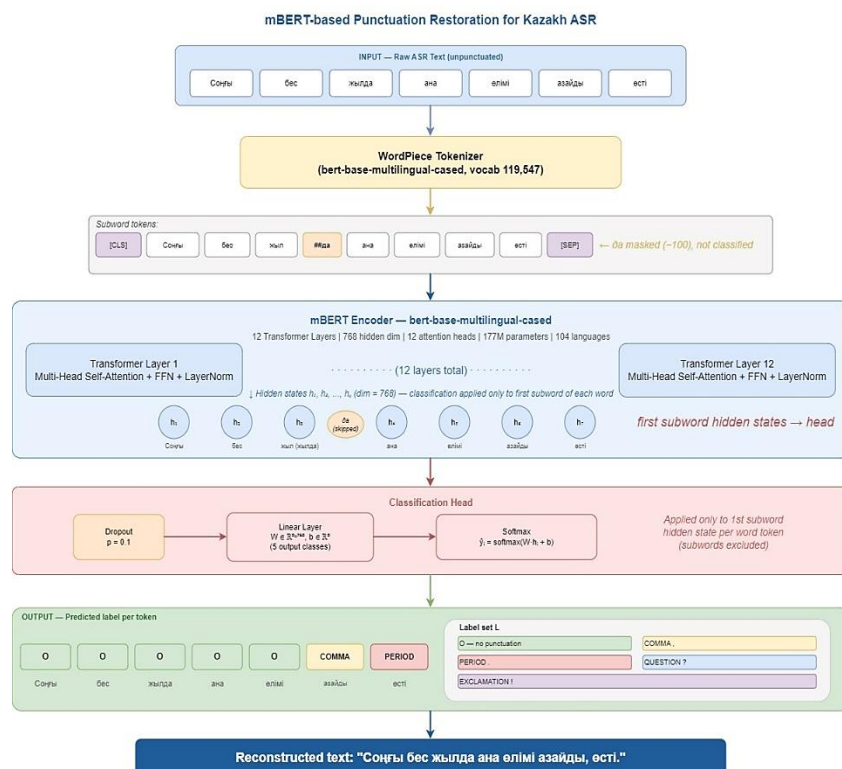
$$h_i \in \mathbb{R}^d \quad (2)$$

мұндағы h_i – i -ші токеннің жасырын көрінісі, d – энкодер өлшемі (әдетте 768). Dropout қабаты артық оқытылуды (overfitting) азайту үшін қолданылады. Сызықтық классификация басы (Linear Head) әрбір токеннің векторын (h_i) 5 класқа келесідей түрлендіреді (3):

$$y_i = \text{softmax}(W \cdot h_i + b) \quad (3)$$

мұндағы: $W \in \mathbb{R}^{5 \times d}$ – салмақ матрицасы – $b \in \mathbb{R}^5$ – ығысу (bias) – $y_i \in \mathbb{R}^5$ – i -ші токеннің ықтималдық векторы (5 класс бойынша: O, COMMA, PERIOD, QUESTION және EXCLAMATION).

BPE (Byte pair encoding) алгоритмі бойынша бірнеше ішкі сөздерге (subwords) бөлінген токендер үшін классификация тек бірінші ішкі сөзге ғана қолданылды. Ал, қалған позициялар -100 мәнімен маскеланып, шығын функциясын (loss) есептеу кезінде ескерілмейді. Қазақ тіліндегі ASR жүйесіне арналған mBERT негізіндегі пунктуацияны қалпына келтіру архитектурасы төмендегі 1-суретте көрсетілген.



1-сурет. Қазақ ASR жүйесі үшін тыныс белгілерін қалпына келтірудің архитектурасы
Ескерту – авторлармен құрастырылған

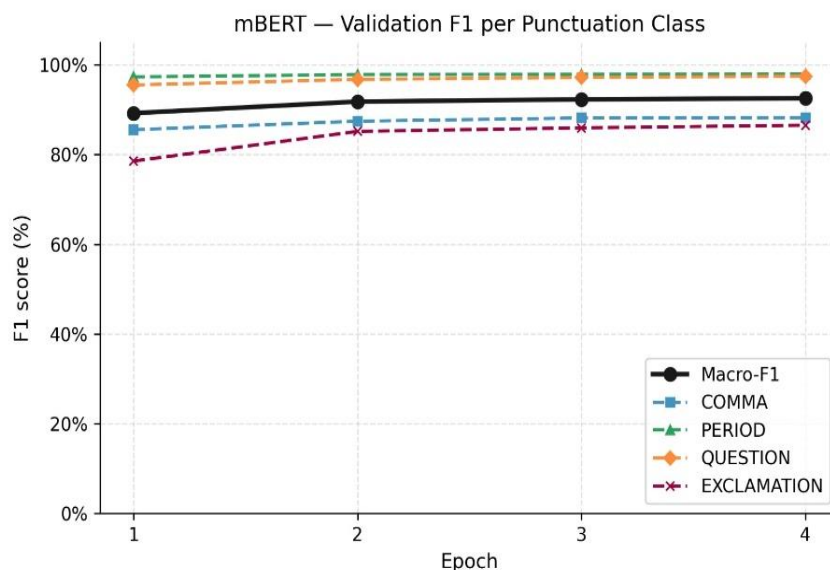
Оқыту шығыны (training loss) барлық бес эпоха барысында тұрақты түрде төмендеді (1-эпоха 0,1418, 2-эпоха 0,0512, 3-эпоха 0,0389, 4-эпоха 0,0310, 5-эпоха 0,0254), бұл өз кезегінде модельдің ауытқусыз бірқалыпты оқытылғанын көрсетеді. Валидациялық Macro-F1 көрсеткіші әр эпохада жақсарып отырды; ең жақсы бақылау нүктесі (checkpoint) 5-эпохадан кейін сақталды, мұнда валидациялық Macro-F1 мәні 93,32 % болды (1-кесте).

Кесте 1. Эпохалар бойынша оқытылудың сапа көрсеткіштері

Эпоха	Train Loss	Val Loss	Macro F1
1	0.1418	0.0589	91.18%
2	0.0512	0.0495	92.53%
3	0.0389	0.0484	92.80%
4	0.0310	0.0521	93.13%
5 ✓	0.0254	0.0540	93.32%

Ескерту – авторлармен құрастырылған

Төмендегі 2-суретте әрбір пунктуация класы (5 класс) үшін эпоха ішіндегі F1 валидациясы көрсетілген.

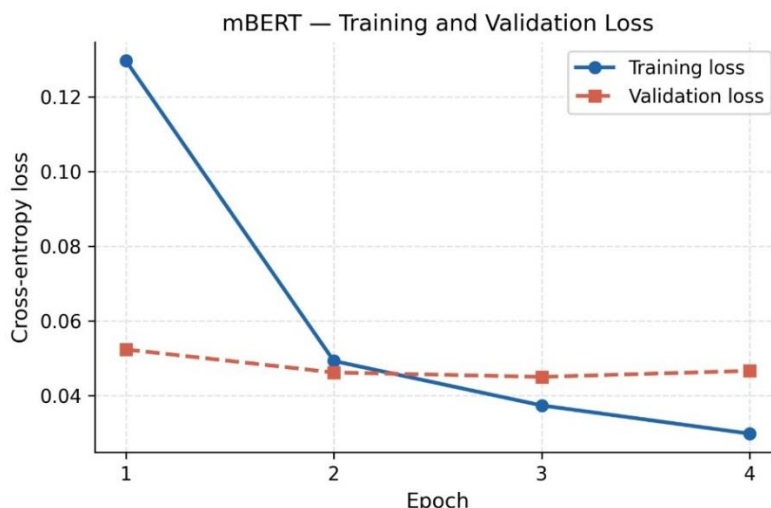


2-сурет. Әр эпоха бойынша тыныс белгілері класының F1 көрсеткіші)

Ескерту – авторлармен құрастырылған

Төмендегі 3-суретте модельдің оқу деректеріндегі қателігі мен модельдің бұрын көрмеген валидациялық деректерде қателіктер жасауы көрсетілген. Суреттен оқыту және валидациялық шығындардың эпохалар бойынша төмендеуін көруге болады, бұл модельдің дұрыс оқытылғанын көрсетеді.

Оқытудың гиперпараметрлері келесідей болды: шығыс қабаты ретінде Viterbi декодтауы бар CRF, dropout = 0.1, AdamW optimizer, learning rate = 2×10^{-5} , batch size = 16, maximum sequence length = 128 және ерте тоқтату шыдамдылығымен 10 эпох дейінгі оқыту = 3. Класс теңгерімсіздігін азайту үшін EXCLAIM және QUESTION үлгілері сәйкесінше 5 және 3 коэффициенттерімен артық іріктелді.



3-сурет. Эпохалар бойынша оқыту және валидациялық шығындары
Ескерту – авторлармен құрастырылған

Барлық өлшемдер NVIDIA RTX 5090 графикалық процессорында FP32 дәлдігімен (precision) жүзеге асырылды. Learning rate scheduler ретінде linear decay with warmup қолданылды. Нәтижелердің қайта өндірілуін қамтамасыз ету үшін барлық эксперименттерде random seed = 42 тұрақты етіп бекітілді. Қолданылған mBERT-base моделінің жалпы параметрлер саны шамамен 178 миллион құрайды (bert-base-multilingual-cased архитектурасына сәйкес).

НӘТИЖЕЛЕР ЖӘНЕ ОЛАРДЫ ТАЛҚЫЛАУ

Корпустағы тыныс белгілерінің үлесі тең емес: мысалы, О (тыныс белгілері жоқ) 88,89%, ал леп белгісі тек 0,46% құрайды. Мұндай теңсіздік модельдің болжамдарын ең көп таралған класқа қарай бұрмалауы мүмкін. Бұл жұмыста салмақ қосу нәтижелерді қайталауды қиындатуына байланысты және нақты мәндерге тым тәуелді ететіндіктен, салмақ коэффициенттерін әдейі енгізілмеді. Мұнда Macro-F1 индикаторы ерте тоқтату критерийі ретінде пайдаланылды. Себебі, ол сирек кездесетін таңбалардың сапасын сақтауға көмектеседі, мұнда модель жалпы дәлдікті ғана емес, орташа көрсеткішті жақсартумен тоқтайды.

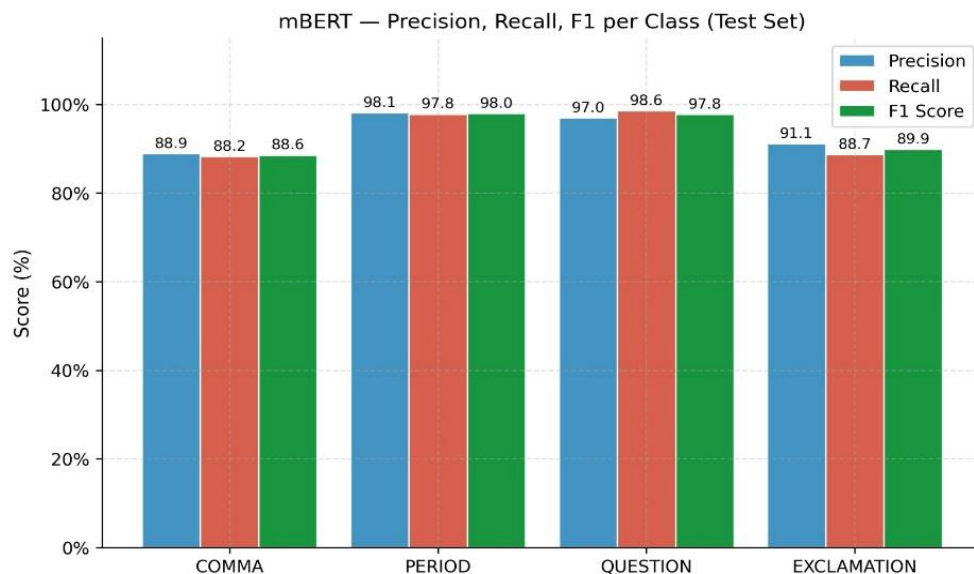
Ең жақсы модель ұзақ уақытқа созылған тест жиынтығында бағаланды (12 826 сөйлем / 200 092 токен). 2-кестеде тыныс белгісінің әр класы бойынша соңғы тест көрсеткіштері көрсетілген.

2-кесте. mBERT тест жиынтығының нәтижелері: әр класс бойынша F1, Precision және Recall

Class	F1	Precision	Recall	Support
O (no punct.)	99.23	99.20	99.27	177,406
COMMA	87.77	88.45	87.10	10,955
PERIOD	98.99	98.91	99.06	9,935
QUESTION	94.35	94.81	94.18	877
EXCLAMATION	91.82	92.68	90.97	919
Macro F1 (punct. only)	93.23	93.71	92.83	—
Weighted F1 (all)	98.54	—	—	200,092

Ескерту – авторлармен құрастырылған

mBERT моделі төрт тыныс белгілері класы бойынша 93,23% Macro-F1 тестілеу деректерінде қол жеткізді. Ал, PERIOD (98,99%) және QUESTION (94,35%) ең жоғары жеке F1 шамаларын көрсетті, бұл олардың контекстік болжамдылығын көрсетеді. COMMA бойынша F1 шамасы 87,77% көрсетті. Атап айтқанда, EXCLAMATION корпус құру кезінде деректерді мақсатты түрде көбейтуге байланысты алдыңғы жұмыспен (Umbet et al., 2025) салыстырғанда айтарлықтай жақсарумен 91,82% F1 ұпайына қол жеткізді. 4-суретте әр класс үшін F1, Precision және Recall көрнекі түрде көрсетілген, ал 5-суретте тест жиынтығындағы тыныс белгілері кластары үшін шатастыру матрицасы көрсетілген.

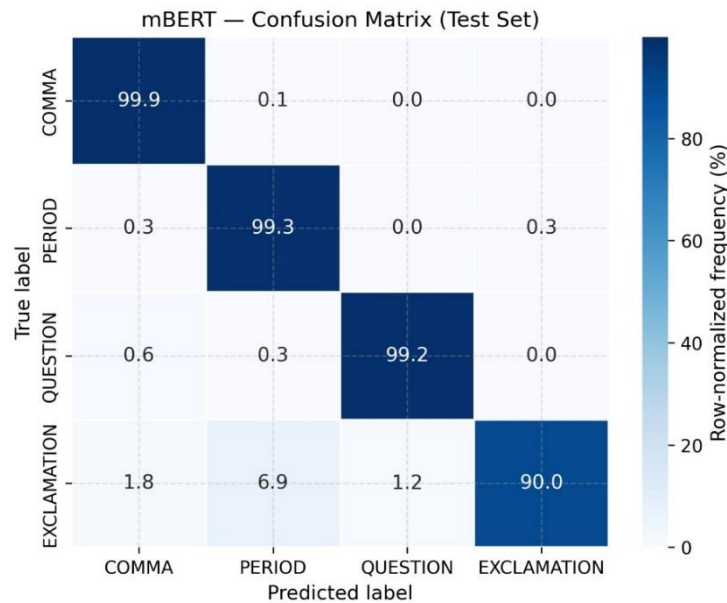


4-сурет. Тест жинағындағы тыныс белгілері класы үшін дәлдік, еске түсіру және F1 (mBERT)

Шатастыру матрицасы барлық кластар бойынша диагональды басымдықтың күшті екенін растайды. Мұнда О (99,25%) және PERIOD (99,06%) толық дерлік жіктелінген. Ал, EXCLAMATION ең жоғары қатені көрсетеді: ол 4,24% жағдайда QUESTION ретінде және 2,94% жағдайда О ретінде қате жіктелінген. QUESTION да, өз кезегінде, 3,42% жағдайда EXCLAMATION ретінде қате танылады. Бұл екі класс арасындағы шатасу лингвистикалық тұрғыдан күтілетін жағдай, өйткені сұраулы және лепті сөйлемдер жиі ұқсас лексикалық белгілермен (сұрақ есімдіктері, күшейткіш бөлшектер) басталады және негізінен ASR мәтінінде жоқ интонациялық белгілермен ажыратылады (5-сурет).

Тәжірибенің нәтижесінде mBERT жеткілікті үлкен және класс бойынша теңдестірілген деректер жиынында оқытылған кезде қазақ тіліндегі ASR шығысындағы тыныс белгілерін тиімді түрде қалпына келтіре алатынын көрсетті. Оқыту динамикасы 5 эпоха бойы тұрақты конвергенцияны көрсетті, шамадан тыс сәйкестік белгілері болмады, бұл көрінбейтін мәтінге жақсы жалпылауды көрсетеді.

Оқыту мақсатынан класс салмағын әдейі алып тастау қайталануды қамтамасыз ету үшін әдістемелік таңдау болды. Оның орнына, класс теңгерімсіздігін шешу үшін деректер жиыны деңгейіндегі күшейту қолданылды, ал Macro-F1 ерте тоқтату критерийі (О қоспағанда) оқыту кезінде сирек кездесетін класстардың сапасының сақталуын қамтамасыз етті.



5-сурет. Шатастыру матрицасы

Confusion matrix нәтижелері бойынша негізгі қалған қате EXCLAMATION белгісінің QUESTION (4,24%) және O (2,94%) кластарымен шатасуы болып табылады. Бұл лингвистикалық тұрғыдан негізделген: сұраулы және лепті сөйлемдер жиі ұқсас лексикалық белгілермен басталады, ал лепті сөйлемнің интонациясы «таза» ASR мәтінінде жоқ болғандықтан, EXCLAMATION белгісін тек контекст арқылы ажырату ең қиын класс болып қалады. Болашақта осы қатені азайту үшін просодикалық белгілерді (интонация, екпін, ырғақ) немесе акустикалық эмбеддингтерді қолдануды қарастыруға жоспарлануда.

Қазақ-орыс код-ауысымды (code-switching) мәтін үшін бөлек бағалау жиынтығы құрылды: орыс тіліндегі жабық класты функционалды сөздер (конечно, вообще, кстати, нельзя, потому, поэтому және т.б., қарыз сөздермен шатаспайтын) негізінде 8 107 сөйлем анықталды. Осы жиынтықтың 80 %-ы модельдің оқыту жиынына қосылды, ал 20 %-ы (909 сөйлем) толығымен бөлек ұсталып, тек соңғы бағалау үшін пайдаланылды. Кесте 3-те көрсетілгендей, code-switching деректерін оқыту жиынына қосу жүйенің сапасын айтарлықтай жақсартты: Macro-F1 32,81 %-тен 63,88 %-ке дейін өсті. PERIOD (91,37 %) және COMMA (71,42 %) кластары бойынша нәтиже қанағаттанарлық деңгейге жетті, алайда EXCLAMATION класы бойынша сапа айтарлықтай төмен болып қалды (26,47 % F1). Бұл лепті сөйлемдердің code-switching контекстінде интонациялық белгілерсіз анықталуының қиындығын көрсетеді, сондықтан келешек зерттеулердің басым бағыты осы болуы тиіс.

3-кесте. Code-switching тест жиынтығындағы нәтижелер
(моно-тілді тестпен салыстырғанда)

Class	Моно-тест F1 (%)	CS-тест F1 (%)	CS Precision (%)	CS Recall (%)
COMMA	87,77	71,42	71,96	70,88
PERIOD	98,99	91,37	92,55	90,22
QUESTION	94,35	66,27	64,37	68,29
EXCLAMATION	91,82	26,47	26,09	26,87
Macro-F1	93,23	63,88	—	—

Ескерту – авторлармен құрастырылған. CS-тест 909 сөйлемнен тұрады, модель оқыту кезінде бұл сөйлемдерді мүлдем көрмеген

ҚОРЫТЫНДЫ

Бұл зерттеудің нәтижелері қазақ тіліндегі ASR жүйелерінде тыныс белгілерін қалпына келтіру мүмкін екенін көрсетіп, екі негізгі үлес ұсынады: біріншіден, бағалау қазақ-орыс код-ауысымды мәтіндерінде жүргізіліп, оқыту жиынына тиісті деректерді қосу арқылы жүйенің екітілді жағдайлардағы сапасы екі есеге жуық жақсарды (Macro-F1 32,81 %-тен 63,88 %-ке дейін), дегенмен EXCLAMATION класы бойынша сапа әлі де шектеулі; екіншіден, осы сценарий үшін арнайы тыныс белгілерімен (O, COMMA, PERIOD, QUESTION, EXCLAMATION) таңбаланған корпус құрылып, тыныс белгілерін қалпына келтіру міндеттері үшін сенімді дереккөз ұсынылды.

Бұл мақалада қазақ тіліндегі ASR шығысы үшін тыныс белгілерін қалпына келтірудің Трансформер негізді тәсілі ұсынылды. Сирек кездесетін класстарды жақсырақ қамту үшін нақты жиналған және LLM-ассистенттелген парафраз сөйлемдермен толықтырылған 126 785 сөйлемнен тұратын корпусты fine-tune жасалған mBERT моделін қолдана отырып, төрт тыныс белгісі сыныбы бойынша моно-тілді тест жиынында Macro-F1 = 93,23 % нәтижесіне қол жеткіздік, ал code-switching тест жиынында Macro-F1 = 63,88 % алынды.

Жұмыста алынған негізгі нәтижелер келесідей:

- Леп белгісі мен сұрақ белгісіндегі класс теңгерімсіздігін азайту үшін толықтыру қолданылған, қазақ тіліндегі тыныс белгілерін қалпына келтіруге арналған арнайы таңбаланған корпус жасалды.

- Қазақ-орыс код-ауысымды мәтінге бағалау жүргізіліп, оқыту жиынына тиісті деректерді қосу арқылы жүйенің екітілді ортадағы сапасы екі есеге жуық жақсарғаны көрсетілді (Macro-F1 32,81 %-тен 63,88 %-ке дейін), дегенмен EXCLAMATION класы бойынша сапа әлі де шектеулі.

- Umbet & Kozhirbayev (2025) еңбегімен салыстырмалы талдау мақсатты толықтырудың леп белгісі бойынша F1 көрсеткішін 32,85 %-тен 91,82 %-ке дейін, ал сұрақ белгісі бойынша F1 көрсеткішін 69,54 %-тен 94,35 %-ке дейін едәуір арттырғанын көрсетті.

Болашақ жұмыста code-switching сценарийіндегі EXCLAMATION класының сапасын жақсарту үшін просодикалық белгілерді (интонация, кідіріс) пайдалану, сондай-ақ жалпы өнімділікті әрі қарай жақсарту, әртүрлі домендерге төзімділікті күшейту және морфологиялық ерекшеліктері ұқсас басқа түркі тілдеріне кеңейту мақсатында XLM-RoBERTa-large моделін зерттеу жоспарлануда. Ұсынылған әдістеме қазақ тіліндегі ASR транскрипцияларын оқуға ыңғайлы етіп, ресурсы шектеулі тілдерде тыныс белгілерін қалпына келтіру сапасын арттыруға мүмкіндік береді.

МҮДДЕЛЕР ҚАЙШЫЛЫҒЫ: Авторлар мүдделер қайшылығы жоқ екенін мәлімдейді

ҚАРЖЫЛАНДЫРУ: Бұл мақала Ғылым және жоғары білім министрлігінің ғылыми зерттеулерге арналған гранттық қаржыландыру жобасы аясында AP23488624 «Түркі тілдерінің параллель сөйлеу корпусының автоматты генерациясын және олардың нейрондық модельдер үшін қолданылуын зерттеу» тақырыбы бойынша дайындалды.

ДЕРЕКТЕРДІҢ ҚОЛЖЕТІМДІЛІГІ ТУРАЛЫ МӘЛІМДЕМЕ: Осы зерттеуде пайдаланылған деректер жиынтығы көпшілікке қолжетімді емес. Осы техникалық зерттеудің нәтижелерін растайтын деректерді тиісті автордан сұраныс жіберу арқылы алуға болады.

АЛҒЫС: Осы мақаланың ғылыми негізділігі мен сапасын арттыруға үлес қосқан құнды пікірлері үшін жасырын пікір жазғандарға ерекше алғыс білдіреміз. Әл-Фараби атындағы Қазақ ұлттық университетінің табиғи тілді компьютерлік өңдеу зерттеушілер тобының қазақ тілін дамытуға, оның мемлекеттік тіл ретіндегі мәртебесін цифрландыру арқылы нығайтуға және білім беруде, ғылымда және мемлекеттік басқаруда қолданылуын арттыруға бағытталған жұмыстары бұл жобаны жүзеге асыруда маңызды рөл атқарды.

ЖАСАНДЫ ИНТЕЛЛЕКТ ТЕХНОЛОГИЯЛАРЫН ПАЙДАЛАНУ ТУРАЛЫ ХАБАРЛАМА: Бұл ғылыми мақаланы дайындау кезінде авторлар генеративті жасанды интеллектті (жергілікті орналастырылған Qwen3-32B моделі) сирек кездесетін тыныс белгілері кластары (QUESTION, EXCLAMATION) үшін қосымша оқыту сөйлемдерін парафраздау мақсатында, сондай-ақ сөздерді жақсарту және мақаланы рәсімдеу үлгісіне келтіру үшін пайдаланды. Барлық генерацияланған сөйлемдер нақты дереккөзден алынған үлгі (seed) сөйлемдер негізінде жасалды және сапа критерийлері бойынша сүзгіленді; олардың seed-сөйлемі мен генерациялаған моделі бөлек тіркеліп сақталды. Барлық ғылыми идеялар, талдау, нәтижелерді түсіндіру және қорытындылар тек авторлардікі.

ӘДЕБИЕТТЕР ТІЗІМІ

- Alam, T., Khan, A., & Alam, F. (2020). Punctuation restoration using Transformer models for high- and low-resource languages. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)* (pp. 132–142). Association for Computational Linguistics. <https://aclanthology.org/2020.wnut-1.18/>
- Batista, F., Caseiro, D., Mamede, N. J., & Trancoso, I. (2007). Recovering punctuation marks for automatic speech recognition. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007)* (pp. 2153–2156). ISCA. <https://api.semanticscholar.org/CorpusID:7805805>
- Batista, F., Caseiro, D., Mamede, N. J., & Trancoso, I. (2008). Recovering capitalization and punctuation marks for automatic speech recognition: Case study for Portuguese broadcast news. *Speech Communication*, 50(10), 847–862. <https://doi.org/10.1016/j.specom.2008.05.001>
- Batliner, A., Kompe, R., Kießling, A., Mast, M., Niemann, H., & Nöth, E. (1998). M = Syntax + Prosody: A syntactic-prosodic labelling scheme for large spontaneous speech databases. *Speech Communication*, 25(4), 193–222. [https://doi.org/10.1016/S0167-6393\(98\)00037-5](https://doi.org/10.1016/S0167-6393(98)00037-5)
- Courtland, M., Faulkner, A., & McElvain, G. (2020). Efficient automatic punctuation restoration using bidirectional transformers with robust inference. In *Proceedings of the 17th International Conference on Spoken Language Translation* (pp. 272–279). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.iwslt-1.33>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Gravano, A., Jansche, M., & Bacchiani, M. (2009). Restoring punctuation and capitalization in transcribed speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009)* (pp. 4741–4744). IEEE. <https://doi.org/10.1109/ICASSP.2009.4960700>
- Görmez, Y., Arslan, H., & Elyakan, M. (2025). Developing deep learning models for Turkish automatic punctuation restoration using a novel dataset. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(11), 1–15. <https://doi.org/10.1145/3772087>
- Huang, Q., Ko, T., Tang, H. L., Liu, X., & Wu, B. (2021). Token-level supervised contrastive learning for punctuation restoration. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2021)* (pp. 3241–3245). ISCA. <https://doi.org/10.21437/Interspeech.2021-661>
- Kurt, U., & Çayır, A. (2023). Transformer-based punctuation restoration for Turkish. In *Proceedings of the 8th International Conference on Computer Science and Engineering (UBMK 2023)* (pp. 169–174). IEEE. <https://doi.org/10.1109/UBMK59864.2023.10286690>
- Nagy, A., & Bial, B. (2021). Automatic punctuation restoration with BERT models. *arXiv*. <https://doi.org/10.48550/arXiv.2101.07343>

- Mamyrbayev, O., Turdalyuly, M., Mekebayev, N., Mukhsina, K., Keylan, A., BabaAli, B., Nabieva, G., Duisenbayeva, A., & Akhmetov, B. (2019). Continuous speech recognition of Kazakh language. ITM Web of Conferences, 24, 01012. <https://doi.org/10.1051/itmconf/20192401012>
- Mamyrbayev, O., Alimhan, K., Zhumazhanov, B., Turdalykyzy, T., & Gusmanova, F. (2020). End-to-end speech recognition in agglutinative languages. In Intelligent Information and Database Systems: 12th Asian Conference (ACIIDS 2020), Phuket, Thailand, 23–26 March 2020 (Part II, pp. 391–401). Springer. https://doi.org/10.1007/978-3-030-42058-1_33
- Mamyrbayev, O. Z., Oralbekova, D. O., Alimhan, K., & Nuranbayeva, B. M. (2023). Hybrid end-to-end model for Kazakh speech recognition. International Journal of Speech Technology, 26(6), 261–270. <https://doi.org/10.1007/s10772-022-10015-5>
- Mamyrbayev, O., Oralbekova, D., Alimhan, K., Turdalykyzy, T., & Othman, M. (2022). A study of transformer-based end-to-end speech recognition system for Kazakh language. Scientific Reports, 12, 8337. <https://doi.org/10.1038/s41598-022-12260-y>
- Mussakhojayeveva, S., Khassanov, Y., & Varol, H. A. (2021). A study of multilingual end-to-end speech recognition for Kazakh, Russian, and English. In Speech and Computer (SPECOM 2021), Lecture Notes in Computer Science, vol. 12997 (pp. 448–459). Springer. https://doi.org/10.1007/978-3-030-87802-3_41
- Păiș, V., & Tufiș, D. (2022). Capitalization and punctuation restoration: A survey. Artificial Intelligence Review, 55, 1681–1722. <https://doi.org/10.48550/arXiv.2111.10746>
- Tilk, O., & Alumäe, T. (2015). LSTM for punctuation restoration in speech transcripts. In Proceedings of INTERSPEECH 2015 (pp. 683–687). <https://doi.org/10.21437/Interspeech.2015-240>
- Tilk, O., & Alumäe, T. (2016). Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In Proceedings of INTERSPEECH 2016 (pp. 3047–3051). <https://doi.org/10.21437/Interspeech.2016-1517>
- Umbet, S., & Kozhirbayev, Z. (2025). Restoring punctuation and capitalization in Kazakh: A BERT-based approach for text normalization. In Proceedings of the 10th International Conference on Computer Science and Engineering (UBMK 2025) (pp. 326–330). IEEE. <https://doi.org/10.1109/UBMK67458.2025.11207023>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS 2017) (Vol. 30). <https://doi.org/10.48550/arXiv.1706.03762>
- Želasko, P., Szymański, P., Mizgajski, J., Szymczak, A., Carmiel, Y., & Dehak, N. (2018). Punctuation prediction model for conversational speech. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2018) (pp. 1–6). ISCA. <https://doi.org/10.48550/arXiv.1807.00543>

Авторлар туралы мәліметтер
Информация об авторах
Information about authors



Кәрібаева Айдана Сейілғазықызы – PhD, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан

Кәрібаева Айдана Сейілғазықызы – PhD, Казахский национальный университет имени аль-Фараби, г. Алматы, Казахстан

Karibayeva Aidana – PhD, Al-Farabi Kazakh National University, Almaty, Kazakhstan,

e-mail: Aidana.Karibayeva@kaznu.edu.kz, a.s.karibayeva@gmail.com

ORCID: <https://orcid.org/0000-0002-2023-1573>,



Мысов Олег – 3-курс студенті, әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан

Мысов Олег – студент 3 курса, Казахский национальный университет имени аль-Фараби, г. Алматы, Казахстан

Myssov Oleg – student of 3rd course, Al-Farabi Kazakh National University, Almaty, Kazakhstan,

e-mail: myssov_oleg@live.kaznu.kz ,

ORCID: <https://orcid.org/0009-0006-9609-4073>



Карыбаева Адина – педагогикалық ғалымдар магистрі, Абылай хан атындағы Қазақ халықаралық қатынастар және әлем тілдері университеті, Алматы қ., Қазақстан

Карыбаева Адина – магистр педагогических наук, Казахский университет международных отношений и мировых языков им. Аблай-хана, г. Алматы, Казахстан

Karybayeva Adina – Master of Pedagogical Sciences, Ablai Khan Kazakh University of International Relations and World Languages, Almaty, Kazakhstan,

e-mail: Adinakarybayeva@gmail.com ,

ORCID: <https://orcid.org/0009-0003-3297-9006>