

АНДАТПА
Абдуали Балжан Абдирасилқызы
философия докторы (PhD) дәрежесін алу үшін дайындалған
диссертациясына
8D07109 – «Автоматтандыру және Internet of Things» білім беру
бағдарламасы бойынша
«Қырғыз-қазақ тілдік жұбы үшін кездесу стенограммаларын
жасауға арналған автоматтандырылған жүйенің моделі мен құралдарын
әзірлеу және зерттеу» тақырыбында

Зерттеу тақырыбының өзектілігі. Қазіргі таңда цифрлық технологиялардың қарқынды дамуы табиғи тілдерді өңдеу, автоматты сөйлеуді тану және машиналық аударма салаларының маңыздылығын арттыруда. Соған қарамастан, қырғыз және қазақ тілдері сияқты түркі тілдері ресурсы шектеулі тілдер қатарына жатады. Бұл тілдер үшін таңбаланған аудио деректердің, параллель мәтіндердің және сапалы корпустардың жеткіліксіздігі заманауи интеллектуалды жүйелерді тиімді әзірлеуге кедергі келтіреді.

Қырғыз-қазақ тілдік жұбы лингвистикалық тұрғыдан жақын болғанымен, осы бағытта сөйлеуді өңдеу, автоматты аударма және мәтінді қорытындылау жүйелері жеткілікті деңгейде зерттелмеген. Түркі тілдерінің агглютинативті табиғаты, морфологиялық күрделілігі және сөз формаларының көптүрлілігі модельдерді құру процесін қиындатады.

Сонымен қатар, қазіргі қоғамда кездесулер мен жиналыстардың стенограммаларын автоматты түрде қалыптастыруға деген сұраныс артып отыр. Мұндай жүйелер сөйлеуді мәтінге түрлендіру, мәтінді өңдеу, негізгі ақпаратты бөліп алу және қысқаша мазмұндау сияқты бірнеше күрделі кезеңдерді қамтиды. Әсіресе көптілді ортада мұндай шешімдер ақпарат алмасуды жеделдетіп, адами ресурстарды үнемдеуге мүмкіндік береді.

Осыған байланысты қырғыз-қазақ тілдік жұбы үшін кездесулер стенограммаларын автоматты түрде құрастыруға арналған модельдер мен құралдарды әзірлеу және зерттеу өзекті ғылыми әрі практикалық міндет болып табылады.

Қазіргі таңда жасанды интеллект технологиялары адамның сөйлеуін өңдеу, мәтінді талдау және ақпаратты автоматты түрде құжаттау салаларында кеңінен қолданылып келеді. Соның ішінде дыбысты автоматты тану, машиналық аударма және жиналыс стенограммаларын автоматтандыру бағыттары ерекше маңызға ие болып отыр. Бұл бағыттар білім беру, мемлекеттік басқару, бизнес, медиа, халықаралық коммуникация және цифрлық сервистерді дамытуда маңызды рөл атқарады. Әсіресе көптілді ортада жұмыс істейтін жүйелерге сұраныс артып, қазақ тілі сияқты ресурсы салыстырмалы түрде аз тілдер үшін тиімді модельдерді зерттеу өзекті мәселеге айналды.

Автоматты сөйлеуді тану жүйелері адамның ауызша айтқан сөзін мәтінге айналдыруға мүмкіндік береді. Мұндай жүйелер аудиоархивтерді

цифрландыруда, субтитр жасауда, көмекші технологияларда, call-center жүйелерінде және жиналыстарды автоматты тіркеуде қолданылады. Бұл салада соңғы жылдары Whisper, MMS және түркі тілдеріне бағытталған Turkic ASR секілді модельдер үлкен қызығушылық тудырып отыр. Олардың басты ерекшелігі — көптілділік, төмен ресурсты тілдерге бейімделу және нақты өмір жағдайындағы сөйлеуді тану сапасын арттыруға бағытталуы.

Машиналық аударма саласы да қазіргі кезде өте қарқынды дамуда. Бұрын аударма жүйелері негізінен белгілі бір тіл жұптарымен ғана шектелсе, қазіргі нейрондық модельдер ондаған және жүздеген тілдер арасында аударма жасауға мүмкіндік береді. Осыған байланысты mT5, Gemma және NLLB секілді заманауи модельдер ерекше назар аудартады. Бұл модельдер тек аударма жасау ғана емес, көптілді мәтіндерді түсіну, өңдеу және генерациялау міндеттерін де атқара алады. Сондықтан олардың архитектурасын, жұмыс істеу принциптерін және қолданылу салаларын салыстырып қарастыру ғылыми әрі практикалық тұрғыдан маңызды.

Сонымен бірге, сөйлеуді тану мен машиналық аударма нәтижелерінің сапасын бағалау үшін арнайы метрикалар қолданылады. Олардың ішінде BLEU, chrF, WER және TER көрсеткіштері ең кең таралған. Бұл метрикалар жүйенің нәтижесін эталондық мәтінмен салыстырып, қателер санын, ұқсастық деңгейін және аударманың немесе танудың сапасын сандық түрде бағалауға мүмкіндік береді. Метрикаларды дұрыс түсіну зерттеудің объективтілігін қамтамасыз етеді және модельдердің артықшылықтары мен шектеулерін нақты көрсетуге көмектеседі.

Бұдан бөлек, соңғы жылдары жиналыс стенограммаларын автоматтандыру бағыты да өзекті зерттеу саласына айналды. Мұндай жүйелер жиналыс барысындағы сөйлеуді мәтінге айналдырып қана қоймай, негізгі ойларды, қабылданған шешімдерді, тапсырмаларды және қорытындыларды автоматты түрде бөліп көрсете алады. Бұл әсіресе ұйымдардағы уақытты үнемдеуге, құжат айналымын жеңілдетуге және ақпаратты жүйелеуге үлкен мүмкіндік береді. Алайда қазақ тілі үшін бұл бағыттағы зерттеулер әлі де шектеулі, сондықтан осы саладағы бар жұмыстарды жинақтап, салыстырмалы талдау жасау маңызды.

Осы жұмыстың мақсаты — дыбысты тануға арналған Whisper, Turkic ASR және MMS модельдерін, машиналық аудармаға арналған mT5, Gemma және NLLB модельдерін қарастыру, олардың математикалық модельдерін сипаттау, ғылыми еңбектерге шолу жасау, сондай-ақ BLEU, chrF, WER және TER метрикаларын түсіндіру. Сонымен қатар жиналыс стенограммаларын автоматтандыру бойынша заманауи зерттеулерді талдап, қазақ тілі үшін бұл бағыттың даму мүмкіндіктерін көрсету болып табылады.

Диссертациялық жұмыстың мақсаты. Қырғыз-қазақ тілдік жұбы үшін кездесулер стенограммаларын автоматты түрде қалыптастыруға арналған интеллектуалды жүйенің моделін әзірлеу және оның тиімділігін зерттеу.

Зерттеудің міндеттері. Зерттеу мақсатына жету үшін келесі міндеттер қойылады:

- Қырғыз тіліндегі кездесу стенограммаларын мәтінге түрлендіру үшін автоматты сөйлеуді тану (ASR) моделін зерттеу;
- CSE негізінде Morphology-Aware ASR Post-Processing әдісін әзірлеу
- Қырғыз тілінен қазақ тіліне мәтінді машиналық аударудың нейрондық моделін зерттеу;
- Автоматтандырылған кездесу стенограммаларына арналған интеграцияланған ASR → МТ жүйесін жасау;
- Әзірленген модельдердің сапасын эксперименттік зерттеу жүргізу.

Зерттеу нысаны. Қырғыз сөйлеуді қазақ мәтіндеріне өңдеу процесі.

Зерттеу пәні. Көптілді ортада сөйлеуді тану және машиналық аударма әдістері мен модельдері.

Алынған нәтижелердің ғылыми жаңалығы:

- Қырғыз–қазақ тілдік жұбы үшін 782 293 параллель сөйлемнен тұратын қомақты параллель корпус әзірленді;
- Қырғыз тілінде сөйлеуді тану нәтижелерін түзетуге арналған CSE технологиясына негізінделген Morphology Aware Post Processing әдісі ұсынылды;
- Қырғыз–қазақ тілдік жұбының кездесу стенограммаларын автоматты түрде жасау жүйесі әзірленді.

Қорғауға шығарылған тұжырымдар:

- Қырғыз–қазақ тілдік жұбы үшін 782 293 параллель сөйлемнен тұратын қомақты параллель корпус әзірлеу;
- Қырғыз тіліндегі сөйлеуді тану нәтижелерін түзетуге арналған CSE негізіндегі Morphology Aware Post Processing әдісі;
- Қырғыз–қазақ тіл жұбы үшін кездесу стенограммаларын автоматты түрде жасау жүйесін әзірлеу.

Жұмыстың теориялық және практикалық маңыздылығы:

- мемлекеттік органдарда, білім беру ұйымдарында және бизнес құрылымдарда кездесулер стенограммаларын автоматтандыруға мүмкіндік береді;
- қырғыз және қазақ тілдеріндегі сөйлеу технологияларын дамытуға үлес қосады;
- көптілді интеллектуалды жүйелерді құруда қолдануға болатын әмбебап шешім ұсынады;
- ресурсы шектеулі тілдер үшін тиімді бағдарламалық құралдар жасауға негіз болады.

Тақырыптың ғылыми-зерттеу бағдарламаларымен байланысы:

Диссертациялық жұмыс AP23487816 «Түркі тілдеріндегі сөйлеу және отырыс хаттамаларын қалыптастырудың нейрондық модельдерін зерттеу» - 2024-2026 жылдарға арналған гранттық қаржыландыру жобасы аясында орындалған.

Диссертациялық жұмыстың көлемі мен құрылымы.

Диссертация кіріспеден, үш тараудан, қорытындыдан және әдебиеттер тізімінен тұрады.

Бірінші бөлімде қырғыз–қазақ тіл жұбы үшін сөйлеуді автоматты тану (ASR) және машиналық аударма жүйелерінің теориялық негіздері қарастырылды. Қырғыз және қазақ тілдеріне арналған ASR технологиялары, нейрондық машиналық аударма тәсілдері, сондай-ақ автоматтандырылған стенограмма қалыптастыру жүйелері бойынша заманауи ғылыми зерттеулерге талдау жүргізілді.

Екінші бөлімде қырғыз–қазақ тіл жұбы үшін автоматтандырылған стенограмма қалыптастыру жүйесін әзірлеу әдістемесі ұсынылды. Жүйенің client-server архитектурасы, деректерді жинау және өңдеу әдістері, сөйлеуді тану модельдерін зерттеу кезеңдері, Morphology Aware Post Processing әдісі, машиналық аударма модельдерін әзірлеу және стенограмма жасауға арналған модельдер қарастырылды.

Үшінші бөлімде ұсынылған автоматтандырылған жүйенің эксперименттік нәтижелері мен талдауы келтірілді. ASR модельдерінің сапасы бағаланып, Morphology Aware Post Processing әдісінің тиімділігі талданды. Сонымен қатар қырғыз тілінен қазақ тіліне аудару нәтижелері ұсынылып, жүйе қолданыстағы әдістермен салыстырылды. Ұсынылған жүйенің практикалық қолдану ерекшеліктері, шектеулері және болашақта жетілдіру бағыттары қарастырылды. Ұсынылған тәсілдің ресурсы аз қырғыз–қазақ тілдік жұбы үшін тиімділігі негізделіп, автоматтандырылған стенограмма қалыптастыру жүйесін әрі қарай дамыту мүмкіндіктері сипатталды.

Жарияланымдар

Диссертация тақырыбы бойынша 6 баспа жұмысы жарияланды, оның ішінде 3 жұмыс халықаралық дәйексөздер базасына кіретін рецензияланған журналдарда («Web of Science» және/немесе «Scopus»), 1 жұмыс халықаралық конференциялар жинақтарында жарияланды. Сондай-ақ, зерттеу жұмысының нәтижесі негізінде «Қазақско–қырғызский параллельный корпус текстов «KazKyrParC»» атты Қазақстан Республикасының авторлық құқықпен қорғалатын объектілерге құқықтардың мемлекеттік тізілімге мәліметтерді енгізу туралы 2025 жылы 14 қазанда берілген №62995 куәлік тіркелді (Қосымша).

Жарияланымдар:

- U. Tukeyev, A. Shormakova, A. Karibayeva, D. Rakhimova, B. Abduali, D. Amirova, N. Rakhmanberdi, R. Aliyev. An Integrated Approach to Adapting Open-Source AI Models for Machine Translation of Low-Resource Turkic Languages 2026 Computers 2026, MDPI, 15(2), 73. (Процентиль 84, Q1) <https://doi.org/10.3390/computers15020073>
- A. Karibayeva, V. Karyukin, B. Abduali, D. Amirova, Speech Recognition and Synthesis Models and Platforms for the Kazakh Language

2025 Information 2025, MDPI, 16(10), 879; (Процентиль 74, Q2) <https://doi.org/10.3390/info16100879>

- A. Karibayeva, V. Karyukin, U. Tukeyev, B. Abduali, D. Amirova, D. Rakhimova, R. Aliyev, A. Shormakova. The Development and Experimental Evaluation of a Multilingual Speech Corpus for Low-Resource Turkic Languages 2025 Applied Science. MDPI, 2025, 15(24), 12880; (Процентиль 79, Q2) <https://doi.org/10.3390/app152412880>

- Abduali B., Milosz, M., Tukeyev, U., Karibayeva, A. Methodology for creating datasets of parallel sentences in low-resource languages by using AI. International Journal of Innovative Research and Scientific Studies, 8(9), 13–23. 2025 <https://doi.org/10.53894/ijirss.v8i9.10605>

- А.Кәрібаева, Б. Абдуали, Д.Амирова. Анализ технологии обратного перевода в англо-казахском нейронном машинном переводе, 31.10-1.11.2024, IX Международная научно-практическая конференция «Информатика и прикладная математика»

- Abduali B., Tukeyev U., Zhumanov Zh., Israilova N. Study of Kyrgyz-Kazakh Neural Machine Translation 2025 17th Asian Conference on Intelligent Information and Database Systems, proceedings conference paper (Scopus) Doi: 10.1007/978-981-96-5881-7_21.

- Авторлық куәлік «Қазақско–қырғызский параллельный корпус текстов «KazKyrParC»» № №62995. 14.10.2025