

әл-Фараби атындағы Қазақ Ұлттық университеті

ӘОЖ 004.912:81'322.2 (043)

Қолжазба құқығында

АБДУАЛИ БАЛЖАН АБДИРАСИЛҚЫЗЫ

**Қырғыз-қазақ тілдік жұбы үшін кездесу стенограммаларын жасауға
арналған автоматтандырылған жүйенің моделі мен құралдарын әзірлеу
және зерттеу**

8D07109 – Автоматтандыру және Internet of Things

Философия докторы (PhD) дәрежесін алу үшін дайындалған диссертация

Отандық ғылыми кеңесші:
техника ғылымдарының докторы, профессор Тукеев У.А.
әл-Фараби атындағы Қазақ ұлттық университеті,
Алматы, Қазақстан

Шетелдік ғылыми кеңесші:
Профессор, доктор, Марек Милош,
Lublin University of Technology, Люблин, Польша

Қазақстан Республикасы
Алматы, 2026

МАЗМҰНЫ

ҚЫСҚАРТУЛАР МЕН БЕЛГІЛЕУЛЕР	4
Кіріспе	5
1 ҚЫРҒЫЗ-ҚАЗАҚ ТІЛДЕРІ ҮШІН СӨЙЛЕУДІ ТАҢУ ЖӘНЕ МАШИНАЛЫҚ АУДАРМАНЫҢ ТЕОРИЯЛЫҚ НЕГІЗДЕРІ	10
1.1 Қырғыз-қазақ тілдік жұбы үшін сөйлеуді автоматты тану (ASR) технологиялары бойынша ғылыми зерттеулерді талдау	10
1.1.1 OpenAI Whisper	13
1.1.2 Meta MMS	14
1.1.3 Turkic ASR	16
1.2 Қырғыз-қазақ тілдік жұбы үшін заманауи нейрондық тәсілдер мен аударма жүйелері бойынша зерттеулерді талдау	18
1.3 Қырғыз-қазақ тілдік жұбы үшін автоматтандырылған стенограмма қалыптастыру жүйесіне теориялық шолу	22
2 ҚЫРҒЫЗ-ҚАЗАҚ ТІЛДІК ЖҰБЫ ҮШІН АВТОМАТТАНДЫРЫЛҒАН ЖҮЙЕНІ ӨЗІРЛЕУ ӘДІСТЕМЕСІ	28
2.1 Автоматтандырылған жүйенің архитектурасы	28
2.2 Қырғыз-қазақ тілдері үшін деректерді жинау, өңдеу және талдау әдістері	30
2.2.1. Бірінші кезең – Дайындық	31
2.2.2 Аударма сапасын бағалау метрикалары	32
2.2.3 Деректер жиынтықтарын анықтау және сынақ деректерін дайындау	33
2.2.4. Жасанды интеллект жүйелері және таңдау критерийлері	34
2.2.5. Синтетикалық параллель корпус жасау	34
2.2.5.1. Корпус жасаудың жалпы сипаттамасы	34
2.2.5.2. Аударма жүйесін таңдау	35
2.2.6. Қырғыз-қазақ корпустарын қателерден тазарту	35
2.2.6.1. Анықталған қате түрлері	35
2.2.6.2. Автоматты тазарту модулі	36
2.2.6.3. Қырғыз-қазақ корпусын тазарту нәтижелері	36
2.3 Сөйлеуді тану модельдерін зерттеу кезеңдері	37
2.3.1. Зерттелген ASR модельдері	37
2.3.2. Деректер жиынтығы және бағалау протоколы	38
2.3.3. Қырғыз тілі бойынша модельдерді салыстыру нәтижелері	38
2.3.4. MMS-1b-all моделін таңдаудың негіздемесі	39
2.3.5. Қырғыз сөйлеуін танудағы басты қиындықтар	40
2.3.6. ASR → аударма конвейерінің жалпы сұлбасы	40
2.4 Сөйлеуді танудан алынған деректерді түзету Morphology Aware Post Processing әдісі	41
2.4.1. Жүйенің жалпы сұлбасы	41
2.4.2. Есептің формалды қойылымы	42
2.4.3. BERT қате детекторы	43

2.4.4. Қате сөздерге баламаларды генерациялау	44
2.4.5. MaxEnt перанкері	44
2.4.6. Белгі (Feature) шаблондары.....	44
2.4.7. Септік белгілері.....	47
2.4.8. Көрші сөздермен келісім белгілері	48
2.4.9. MLM/BERT белгілері мен MAPR ASRтолық мысалды көрсету.....	49
2.5 Машиналық аударма модельдерін әзірлеу кезеңдері.....	54
2.5.1 Жалпы әдістемелік шеңбер	54
2.5.2 Модель таңдау критерийлері	55
2.5.3 Модельдерді бағалау және таңдау	55
2.5.4 Таңдалған модельдердің сипаттамасы.....	56
2.5.5 Fine-tuning әдістемесі.....	57
2.5.6 Бағалау метрикалары	58
2.5.7 Қырғыз–қазақ тіл жұбы үшін әдістеменің маңыздылығы	58
3 АВТОМАТТАНДЫРЫЛҒАН СТЕНОГРАММА ҚАЛЫПТАСТЫРУ	
ЖҮЙЕСІНІҢ НӘТИЖЕЛЕРІ МЕН ТАЛДАУЫ.....	60
3.1 ASR және машиналық аударма модельдерінің сапасын талдау	60
3.2 Morphology Aware Post Processing әдісінің нәтижесін талдау	62
3.3 Қырғыз-қазақ тіліне аудару нәтижелері	64
3.3.1. Модельді баптау (Fine-tuning) нәтижелері	64
3.3.2. Сыртқы бағалау: FLORES 200 деректер жиынтығы	65
3.3.3. Деректерді бөлу схемасы	66
3.3.4. Алдын ала тәжірибелер нәтижелері.....	66
3.4 Ұсынылған кездесу стенограммаларын жасауға арналған автоматтандырылған жүйенің нәтижелері.....	67
Қорытынды	71
Пайдаланылған әдебиеттер	73

ҚЫСҚАРТУЛАР МЕН БЕЛГІЛЕУЛЕР

NMT – neural machine translation

HMA – нейронды машиналық аударма

CSE – Complete set of endings (жалғаулардың толық жиынтығы)

WMT – world machine translation (әлемдік машиналық аударма симпозиумы)

BPE – byte-pair encoding (жұптық байттық кодтау)

РНЖ – рекурентті нейронды желі

seq2seq – sequence to sequence (тізбектіліктен тізбектілікке)

LSTM – long-short term memory

BLEU – Bilingual Evaluation Study (бір табиғи тілден екінші табиғи тілге автоматты түрде аударылған мәтіннің сапасын бағалауға арналған бағалау метрикасы)

MAPP - Morphology Aware Post Processing

КІРІСПЕ

Зерттеудің 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;

Morphology-Aware Post-Processing әзірлеу

Қырғыз тілінен қазақ тіліне мәтінді машиналық аударудың нейрондық моделін зерттеу;

Автоматтандырылған кездесу стенограммаларына арналған интеграцияланған ASR → MT жүйесінің прототипін 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. Диссертацияның жалпы көлемі 82 беттен, 16 суреттен, 36 кестеден, 80 қолданылған әдебиеттен 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 4 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 «Казахско–кыргызский 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:

Scopus:

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>

Халықаралық конференциялар:

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

1 ҚЫРҒЫЗ-ҚАЗАҚ ТІЛДЕРІ ҮШІН СӨЙЛЕУДІ ТАҢУ ЖӘНЕ МАШИНАЛЫҚ АУДАРМАНЫҢ ТЕОРИЯЛЫҚ НЕГІЗДЕРІ

1.1 Қырғыз–қазақ тілдік жұбы үшін сөйлеуді автоматты тану (ASR) технологиялары бойынша ғылыми зерттеулерді талдау

Автоматты сөйлеуді тану (Automatic Speech Recognition – ASR) – адамның сөйлеуін компьютер арқылы мәтінге айналдыру технологиясы. Бұл жүйелер дыбыстық сигналды өңдеп, оны тілдік бірліктерге (фонема, сөз, сөйлем) бөліп, сәйкес мәтіндік көрінісін құрастырады. ASR технологиясы қазіргі таңда көптеген салаларда кеңінен қолданылады: дауыстық көмекшілер (Siri, Google Assistant), субтитр жасау, call-center жүйелері, білім беру платформалары, аудио архивтерді цифрландыру және жиналыс стенограммаларын автоматты түрде жазу сияқты бағыттарда.

ASR жүйесінің жұмыс істеу принципі бірнеше негізгі кезеңдерден тұрады. Бірінші кезеңде дыбыстық сигнал алдын ала өңдеуден өтеді. Бұл кезеңде шу азайтылады, сигнал қалыпқа келтіріледі және аудио қысқа уақыттық фреймдерге бөлінеді. Кейін әрбір фреймнен акустикалық ерекшеліктер (features) алынады. Ең жиі қолданылатын әдістердің бірі – Mel-frequency cepstral coefficients (MFCC) немесе log-Mel spectrogram. Бұл ерекшеліктер дыбыстың адам құлағына ұқсас қабылдануын модельдеуге мүмкіндік береді.

Келесі кезеңде алынған ерекшеліктер нейрондық желіге беріледі. Қазіргі заманғы ASR жүйелері негізінен терең оқыту әдістеріне сүйенеді. Бұрынғы жүйелерде Hidden Markov Models (HMM) мен Gaussian Mixture Models (GMM) қолданылған болса, қазіргі уақытта оларды Transformer, Conformer және recurrent нейрондық желілер алмастырды. Бұл модельдер уақыт бойынша тәуелділікті жақсы ұстап, контекстті ескеріп, сөйлеуді дәлірек тануға мүмкіндік береді.

ASR жүйелерінің негізгі екі түрі бар: дәстүрлі (hybrid) және end-to-end жүйелер. Hybrid жүйелерде акустикалық модель, тілдік модель және декодер бөлек-бөлек құрылады. Ал end-to-end жүйелерде барлық процесс бір нейрондық желі арқылы жүзеге асады. Мұндай модельдерге Whisper және MMS жатады. Олар аудио сигналды бірден мәтінге айналдырады және қосымша компоненттерді қажет етпейді.

Математикалық тұрғыдан ASR міндеті – аудио сигнал (x) берілген кезде ең ықтимал мәтін (y)-ды табу:

$$y^* = \operatorname{argmax} P(y|x) \quad (1)$$

Көп жағдайда бұл ықтималдық Байес формуласымен қайта жазылады:

$$P(y|x) \propto P(x|y) \cdot P(y) \quad (2)$$

мұнда ($P(x|y)$) – акустикалық модель, ал ($P(y)$) – тілдік модель. Бірақ қазіргі end-to-end модельдер бұл екі компонентті біріктіріп, бір ғана модель арқылы есептейді.

Соңғы жылдары ASR саласында үлкен өзгерістер орын алды. Үлкен көлемдегі деректерде алдын ала үйретілген (pretrained) модельдер пайда болды. Мұндай модельдер әртүрлі тілдерде жұмыс істей алады және аз дерек болған жағдайда да жақсы нәтиже көрсетеді. Мысалы, Whisper моделі weakly supervised

оқыту арқылы көптеген тілдерде жоғары сапалы тану нәтижесін береді, ал MMS жобасы мыңнан астам тілге арналған speech технологияларын дамытуға бағытталған.

Түркі тілдері үшін ASR жүйелерін дамыту да маңызды бағыт болып табылады. Қазақ, қырғыз, өзбек сияқты тілдер үшін деректердің аз болуы модель сапасына әсер етеді. Сондықтан қазіргі зерттеулерде көптілді оқыту (multilingual learning) және бір тілден екінші тілге білім тасымалдау (transfer learning) әдістері қолданылады. Бұл тәсілдер ұқсас тілдердің ортақ ерекшеліктерін пайдаланып, төмен ресурсты тілдер үшін тиімді модельдер құруға мүмкіндік береді.

ASR – күрделі әрі қарқынды дамып жатқан сала. Ол дыбысты өңдеу, тілдік модельдеу және машиналық оқыту әдістерінің тоғысында орналасқан. Заманауи модельдер сөйлеуді тану дәлдігін едәуір арттырды және көптілді ортада тиімді жұмыс істеуге мүмкіндік берді. Бұл технологиялар алдағы уақытта да жиналыс стенограммаларын автоматтандыру, адам-компьютер өзара әрекеттесуін жақсарту және ақпаратқа қолжетімділікті арттыру бағыттарында маңызды рөл атқара береді.

Сөйлеуді автоматты тану (Automatic Speech Recognition – ASR) технологиялары қазіргі таңда жасанды интеллект, компьютерлік лингвистика және сөйлеу өңдеу салаларындағы маңызды ғылыми бағыттардың бірі болып табылады. ASR жүйелері адамның сөйлеу сигналын мәтіндік форматқа автоматты түрде түрлендіруге арналған. Бұл технологиялар виртуалды ассистенттерде, автоматтандырылған транскрипция жүйелерінде, мультимедиялық іздеу жүйелерінде, машиналық аудармада және көптілді коммуникациялық платформаларда кеңінен қолданылады [1].

Соңғы жылдары терең нейрондық желілердің дамуына байланысты сөйлеуді тану жүйелерінің сапасы айтарлықтай артты. Дәстүрлі Hidden Markov Model (HMM) және Gaussian Mixture Model (GMM) негізіндегі жүйелердің орнына қазіргі уақытта Deep Neural Network (DNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), сондай-ақ Transformer архитектуралары негізіндегі модельдер белсенді қолданылуда [2]. Бұл тәсілдер сөйлеу сигналдарының акустикалық және лингвистикалық ерекшеліктерін тиімді модельдеуге мүмкіндік береді.

ASR технологияларының дамуына үлкен көлемдегі таңбаланған деректер корпусы, жоғары өнімді есептеу ресурстары және ашық бастапқы кодты платформалардың пайда болуы әсер етті. Әсіресе, DeepSpeech, Kaldi, wav2vec 2.0 және Whisper сияқты модельдер сөйлеуді тану саласындағы зерттеулердің қарқынды дамуына негіз болды [3]. OpenAI компаниясы ұсынған Whisper моделі көптілді сөйлеуді тануда жоғары нәтижелер көрсетіп, төмен ресурсты тілдер үшін де тиімді құралдардың бірі ретінде қарастырылуда [4].

Қазіргі уақытта ағылшын, қытай, неміс, француз және испан тілдері үшін үлкен көлемдегі сөйлеу корпустары мен алдын ала оқытылған модельдер қолжетімді. Алайда қырғыз және қазақ тілдері сияқты түркі тілдері үшін ресурстар көлемі әлі де шектеулі болып отыр [5]. Бұл жағдай қырғыз–қазақ тіл

жұбы бойынша ASR жүйелерін әзірлеу барысында бірқатар қиындықтар туғызады.

Қазақ және қырғыз тілдері агглютинативті тілдер қатарына жатады. Агглютинативті тілдерде сөздерге көптеген қосымшалардың жалғануы нәтижесінде сөз формаларының саны өте үлкен болады [6]. Мұндай морфологиялық күрделілік ASR жүйелерінде сөздік қордың ұлғаюына, сирек кездесетін сөздердің көбеюіне және тану сапасының төмендеуіне алып келеді. Сондықтан қазіргі зерттеулерде субсөздік модельдеу, морфологиялық сегменттеу және Byte Pair Encoding (BPE) сияқты әдістер кеңінен қолданылуда [7].

Sennrich және басқалары ұсынған BPE алгоритмі бастапқыда нейрондық машиналық аудармада қолданылғанымен, кейіннен сөйлеуді тану модельдерінде де тиімді әдіс ретінде қолданыла бастады [8]. Бұл тәсіл сөздерді жиі кездесетін символдар тізбектеріне бөлу арқылы сирек сөздер мәселесін азайтады. Агглютинативті тілдер үшін BPE қолдану сөздік көлемін қысқартып, модельдің жалпылау қабілетін арттыруға мүмкіндік береді [9].

Сөйлеуді тану жүйелерін әзірлеуде акустикалық модельмен қатар тілдік модельдің де маңызы жоғары. Тілдік модель сөйлемдегі сөздердің ықтимал ретін анықтауға көмектеседі. Қазіргі нейрондық тілдік модельдер Transformer архитектурасына негізделіп құрастырылып, үлкен мәтіндік корпустарда оқытылады [10]. Мұндай тәсілдер қате танылған сөздерді контекст негізінде түзетуге мүмкіндік береді.

Қырғыз және қазақ тілдері фонетикалық жағынан бір-біріне ұқсас болғанымен, олардың дыбыстық жүйесінде, диалектілерінде және орфографиялық ерекшеліктерінде айырмашылықтар бар [11]. Сондықтан екі тілге ортақ көптілді ASR жүйелерін әзірлеу барысында трансферлік оқыту (transfer learning) және көптілді модельдеу әдістері жиі қолданылады [12]. Бұл тәсілдер ресурсы аз тілдер үшін тиімді болып саналады, себебі бір тілдегі білімді екінші тілге тасымалдауға мүмкіндік береді.

Meta AI ұсынған wav2vec 2.0 моделі өзіндік бақылаумен оқыту (self-supervised learning) тәсіліне негізделген және таңбаланбаған аудио деректерді тиімді пайдалануға мүмкіндік береді [13]. Бұл әдіс төмен ресурсты тілдер үшін ерекше маңызды, өйткені таңбаланған сөйлеу корпустарын жинау үлкен қаржылық және уақыттық шығындарды талап етеді. wav2vec 2.0 негізіндегі зерттеулер қазақ және қырғыз тілдерінде де жүргізіле бастады [14].

Соңғы жылдары көптілді сөйлеу модельдері де кеңінен зерттелуде. XLSR, Whisper және SeamlessM4T сияқты модельдер бірнеше тілде бір уақытта сөйлеуді тану және аудару мүмкіндігін қамтамасыз етеді [15]. Мұндай модельдер түркі тілдері арасында ортақ фонетикалық және морфологиялық белгілерді тиімді пайдалануға мүмкіндік береді.

Қазақ тілі бойынша сөйлеуді тану саласында бірқатар зерттеулер жүргізілген. Атап айтқанда, қазақ тіліне арналған сөйлеу корпустарын құру, фонемалық модельдеу және нейрондық ASR архитектураларын бейімдеу бағыттарында жұмыстар бар [16]. Қырғыз тілі бойынша да алғашқы зерттеулер

жүргізіліп, Kaldi және wav2vec платформалары негізінде тәжірибелер жасалған [17].

ASR жүйелерінің сапасын бағалау үшін көбінесе Word Error Rate (WER) және Character Error Rate (CER) метрикалары қолданылады [18]. Агглютинативті тілдерде сөз формаларының көптүрлілігіне байланысты CER метрикасы кейбір жағдайларда тиімдірек нәтиже береді.

Қырғыз–қазақ тіл жұбы үшін сөйлеуді автоматты тану технологияларын дамыту болашақта автоматты субтитрлеу, көптілді коммуникация, сөйлеуді машиналық аударма жүйелері және автоматты жиналыс стенограммаларын құру жүйелерін дамытуға мүмкіндік береді [19]. Сонымен қатар мұндай зерттеулер түркі тілдері үшін цифрлық ресурстарды көбейтуге және төмен ресурсты тілдерге арналған жасанды интеллект технологияларын дамытуға маңызды үлес қосады [20].

1.1.1 OpenAI Whisper

Қазіргі заманғы автоматты сөйлеуді тану (ACT / ASR – Automatic Speech Recognition) жүйелері терең оқыту архитектураларына негізделген. Осы бөлімде қырғыз және қазақ тілдерінің ASR зерттеулеріне тікелей қатысты үш негізгі жүйе қарастырылады: OpenAI Whisper, Meta MMS (Massively Multilingual Speech) және Turkic ASR моделдері. Әрбір жүйенің математикалық негіздері, архитектурасы және қолдайтын тілдер тізбесі егжей-тегжейлі талданады.

OpenAI Whisper (2022) – 680 000 сағаттан астам интернеттен жиналған көпсалалы, көптілді аудио деректерімен үйретілген автоматты сөйлеуді тану жүйесі. Модель мультитаск оқыту парадигмасына сүйенеді: бір уақытта транскрипция, аударма, тіл анықтау және белсенділік детекциясы сияқты бірнеше міндетті орындайды.

Whisper 99 тілді ресми қолдайды. Соның ішінде түркі тілдер тобынан:

- Қазақ тілі (kk) – деректер саны шамамен 117 сағат
- Қырғыз тілі (ky) – деректер саны шамамен 23 сағат
- Түрік тілі (tr), Өзбек тілі (uz), Әзірбайжан тілі (az), Татар тілі (tt)
- Башқұрт (ba), Ұйғыр (ug), Якут/Саха (sah) тілдері

Модель нұсқалары өнімділік пен жылдамдық арасындағы ымыраға байланысты бес өлшемде ұсынылады, 1.1 кестеде көрсетілген.

Кесте 1.1 – OpenAI Whisper модель нұсқаларының параметрлері

Нұсқа	Параметр саны	Қабат саны	Назар басы	Жылдамдық (нақ.)
tiny	39 млн	4	6	~32×
base	74 млн	6	8	~16×
small	244 млн	12	12	~6×
medium	769 млн	24	16	~2×
large-v3	1550 млн	32	20	1×

Whisper Encoder-Decoder Transformer архитектурасына негізделген. Кодер (Encoder) аудио сигналды Log-Mel спектрограммасына түрлендіріп, Transformer қабаттары арқылы векторлық кеңістікке проекциялайды. Декодер (Decoder) autoregressive режимде жетон-жетон (token-by-token) транскрипцияны генерациялайды.

Аудио өңдеу кезеңінде кіріс сигнал алдымен Log-Mel спектрограммасына түрленеді. Бұл процесте сигнал қысқа уақыт аралықтарға бөлініп, әрбір аралық үшін жиілік мазмұны мел шкаласында логарифмдік түрде есептеледі. Нәтижесінде аудионың уақыт-жиілік кескіні алынады, ол кодерге кіріс ретінде беріледі.

Кодер ішінде Multi-Head Self-Attention механизмі жұмыс жасайды. Бұл механизм кіріс реттіліктің кез келген екі позициясы арасындағы тәуелділікті тікелей есептеуге мүмкіндік береді. Бірнеше параллель назар басы бірлесіп жұмыс жасап, аудио белгілерінің әртүрлі аспектілерін бір мезгілде өңдейді. Назар нәтижелері біріктіріліп, кейінгі Feed-Forward желіге беріледі. Feed-Forward желі әрбір Transformer қабатында позиция бойынша тәуелсіз түрлендіру жасайды.

Декодер кодердің шығысы негізінде авторегрессивті ықтималдықты есептейді: яғни әрбір жаңа жетон алдыңғы жетондардың және кодер контекстінің негізінде генерацияланады. Оқыту кезінде кросс-энтропия жоғалым функциясы қолданылып, болжанған жетон тізбегі мен шынайы транскрипция арасындағы айырмашылық азайтылады.

Whisper-дің негізгі ерекшелігі — арнайы жетондар арқылы орындалатын тапсырманы басқару жүйесі. Транскрипция кезінде декодерге транскрипция басталғанын білдіретін жетон, тіл жетоны (мысалы, қазақ немесе қырғыз тілі үшін), тапсырма түрі (транскрипция немесе аударма) және уақыт белгісі режимі көрсетіледі. Бұл тәсіл бір модельде 99 тілді басқаруға мүмкіндік береді және арнайы тіл жетондары арқылы модель қай тілде жұмыс жасайтынын дәл анықтайды.

1.1.2 Meta MMS

Meta AI MMS жобасы (2023) – діни мәтіндер (Інжіл аудармалары) негізінде жасалған 1 100-ден астам тілді қамтитын көпсалалы сөйлеуді тану жүйесі. Бұл жоба ASR саласындағы тіл қамтуының рекордын орнатты.

MMS қолдайтын негізгі тілдер (ASR үшін):

- Қазақ тілі (kaz) – Common Voice және FLEURS деректері негізінде тексерілген
- Қырғыз тілі (kir) – FLEURS мәліметтерімен бағаланған
- Якут/Саха (sah), Башқұрт (bak), Татар (tat), Ұйғыр (uig)
- Барлығы 1162 тіл – бұл Whisper-дің 99 тіліне қарағанда 11 есе артық

Кесте 1.2 – MMS жүйесінің түркі тілдердегі WER көрсеткіштері (FLEURS бенчмарк)

Тіл	Код	FLEURS (%)	WER	Деректер мөлшері
Қазақ	kaz	14.2		~300 сағат (ТД)
Қырғыз	kir	18.7		~120 сағат (ТД)
Якут	sah	22.4		Шектеулі
Башқұрт	bak	19.1		Шектеулі
Татар	tat	16.8		~80 сағат

MMS жүйесі wav2vec 2.0 архитектурасына негізделген өздігінен оқыту (self-supervised learning) парадигмасын қолданады. Архитектура үш негізгі блоктан тұрады: CNN Feature Encoder, Transformer Context Network және Quantization Module.

CNN Feature Encoder жеті қабатты конволюциялық желіден тұрады және шикі аудио толқынын жасырын векторларға түрлендіреді. Бұл блок аудио сигналдың төменгі деңгейдегі акустикалық белгілерін автоматты түрде бөліп алады. Transformer Context Network осы жасырын векторлар арасында Multi-Head Self-Attention механизмі арқылы контекстік ақпарат алмасуды қамтамасыз етеді, яғни сигналдың алыс бөліктері арасындағы тәуелділіктерді үйренеді. Quantization Module үздіксіз аудио белгілерін дискретті кодтарға айналдырады, бұл модельдің өздігінен оқыту мақсатын қалыптастыруға негіз болады.

Алдын ала үйрету кезінде модель маскаланған аудио фреймдерін болжауға негізделген контрастивті жоғалым функциясы арқылы оқытылады. Бұл процесте Transformer шығысы мен сол уақытқа сәйкес квантизацияланған вектор арасындағы ұқсастық максималданып, теріс үлгілермен ұқсастық минималданады. Осы тәсіл арқылы модель таңбаланбаған аудио деректерден тілдік заңдылықтарды өздігінен үйрене алады.

Fine-tuning кезінде CTC (Connectionist Temporal Classification) жоғалым функциясы қолданылады. CTC тәсілі аудио фреймдері мен транскрипция жетондары арасындағы тураластыруды (alignment) алдын ала белгілемей-ақ үйретуге мүмкіндік береді. Модель барлық мүмкін тураластыру жолдарының ықтималдығын бір мезгілде ескере отырып оқытылады, бұл ұзындығы әртүрлі аудио-мәтін жұптарымен жұмыс жасауды жеңілдетеді.

MMS-тің негізгі артықшылығы — adapter негізіндегі модульдік тәсіл. Adapter қабаттары арқылы тілге тән бейімделу жүзеге асырылады: ортақ Transformer параметрлері өзгеріссіз қалып, тек әрбір тілге арналған шағын adapter қабаты fine-tune жасалады. Бұл adapter қабатының параметрлер саны шамамен 1.9 миллионды құрайды, ол ортақ модель параметрлеріне қарағанда өте аз. Осының арқасында жаңа тіл қосу үшін бүкіл модельді қайта оқыту қажет емес — тек сол тілге арналған adapter қабатын үйрету жеткілікті. Бұл тәсіл 1100-ден астам тіл үшін жеке adapter қабаттарын үйретуге мүмкіндік берген және ресурсы шектеулі тілдер үшін өте тиімді болып табылады. Қырғыз тілі де осы 1100-ден астам тілдің қатарына кіретіндіктен, MMS моделі арнайы үйретілген корпусстың шектеулілігіне қарамастан тілді тану міндетін орындай алады.

1.1.3 Turkic ASR

Turkic ASR – түркі тіл тобына арнайы бейімделген ASR жүйелерін зерттейтін ғылыми бағыт. Бұл бағыт негізінен үш зерттеу ағымынан тұрады: (1) жалпы multilingual модельдерді fine-tune ету, (2) Common Voice және FLEURS деректерімен үйрету, (3) Turkic тілдеріне арнайы language model (тілдік модель) интеграциясы.

Зерттеулерде қарастырылатын түркі тілдері:

- Батыс түркі: Түрік (tr), Әзірбайжан (az), Гагауз (gag)
- Қыпшақ тобы: Қазақ (kk), Қырғыз (ky), Татар (tt), Башқұрт (ba), Қарақалпақ (kaa)
- Оғыз тобы: Өзбек (uz), Түрікмен (tk), Ұйғыр (ug)
- Солтүстік түркі: Якут/Саха (sah), Тыва (tyv), Алтай (alt), Хакас (kjh)

Кесте 1.3 – Turkic ASR зерттеулерінде қарастырылатын тілдер және деректер мөлшері

Тіл	ISO коды	Спикер саны (млн)	Common Voice (сағ)	FLEURS (сағ)
Түрік	tr	88	~1,340	10
Қазақ	kk	15	~290	10
Өзбек	uz	35	~180	10
Әзірбайжан	az	30	~73	10
Қырғыз	ky	5.5	~40	10
Башқұрт	ba	1.2	~17	10
Татар	tt	5.3	~12	10
Якут/Саха	sah	0.5	~4	—

Жалпы жүйенің архитектурасы Encoder-Decoder және CTC гибриді жүйе ретінде жұмыс жасайды. Түркі тілдеріне арналған ASR зерттеулерінде кеңінен қолданылатын бұл гибриді архитектура Conformer Encoder мен Transformer Decoder-ді CTC жоғалымымен біріктіреді.

Conformer блогы стандартты Transformer-ге Depthwise Separable Convolution қабатын қосады. Бұл қосымша конволюция модулі локальды акустикалық заңдылықтарды — дыбыс шекаралары, коартикуляция сияқты қысқа мерзімді тәуелділіктерді — тиімді өңдеуге мүмкіндік береді. Нәтижесінде Conformer архитектурасы таза Transformer-ге қарағанда сөйлеуді тануда жоғары нәтиже көрсетеді. Гибриді жоғалым функциясы CTC және Attention жоғалымдарын теңестіру коэффициенті арқылы біріктіреді. Бұл тәсіл оқыту процесінде CTC-нің жылдам конвергенциясы мен Attention декодердің жоғары дәлдігін бір мезгілде пайдалануға мүмкіндік береді.

Түркі тілдерінің агглютинативтік табиғатын есепке алу үшін арнайы тілдік модель интеграциясы қажет. Beam Search декодинг кезінде акустикалық модель мен тілдік модель ықтималдықтары біріктіріліп, ең ықтимал транскрипция таңдалады. Тілдік модель салмағы мен сөз кірістіру бонусы арқылы декодинг

нәтижесі баптала алады, бұл морфологиялық жағынан күрделі тілдерде сөйлемдердің грамматикалық дұрыстығын арттырады.

Токенизация мәселесі де агглютинативтік тілдер үшін ерекше маңызды. BPE (Byte Pair Encoding) немесе SentencePiece токенизациясы субсөздік деңгейде морфемаларды бөлуге мүмкіндік береді. Түркі тілдері үшін сөздік қор өлшемі 1000-ден 5000-ға дейінгі аралықта оңтайлы деп есептеледі — бұл сирек кездесетін сөз формаларының санын азайтып, модель өнімділігін арттырады.

Ресурсы шектеулі қырғыз және қазақ тілдері үшін transfer learning стратегиясы ерекше тиімді болып табылады. Бірінші кезеңде Multilingual Whisper немесе wav2vec 2.0 сияқты ірі мультилингвал модельдің негізі алынады. Екінші кезеңде морфологиялық жағынан жақын тілдер — түрік, қазақ, қырғыз — деректерімен бірлесіп fine-tune жасалады, бұл тілдер арасындағы ортақ фонетикалық және морфологиялық белгілерді пайдалануға мүмкіндік береді. Үшінші кезеңде тікелей мақсат тілдің деректерімен соңғы бейімдеу жүргізіледі. Төртінші кезеңде n-gram немесе нейрондық тілдік модель арқылы декодинг нәтижелері жақсартылады. Ортақ фонетикалық қасиеттерді пайдалану үшін IPA (International Phonetic Alphabet) ортақ фонема кеңістігіне проекциялау тәсілі де қолданылады: қырғыз, қазақ, түрік, өзбек тілдерінің дыбыстық жүйесі ортақ фонема кеңістігіне түсіріліп, деректер жетіспеушілігі компенсацияланады. Бұл кешенді transfer learning стратегиясы қырғыз-қазақ жұбы үшін сөзді тану қателігін (WER) айтарлықтай төмендетуге мүмкіндік береді.

Үш жүйені негізгі техникалық параметрлер бойынша салыстыру келесі 1.4 - кестеде көрсетілген.

Кесте 1.4 – ASR жүйелерінің салыстырмалы сипаттамасы

Параметр	Whisper	MMS	Turkic ASR
Архитектура	Enc-Dec Transformer	wav2vec 2.0 + Adapter	Conformer + CTC/Att
Параметр саны	39M – 1.55Мд	317M + Adapter	~100M – 600M
Тіл саны	99	1162	15–30 (түркі)
Қазақ тілі	✓ (117 сағ)	✓ (FLEURS)	✓ (KazakhTTS2)
Қырғыз тілі	✓ (23 сағ)	✓ (FLEURS)	✓ (шектеулі)

Оқыту парадигмасы	Supervised	Self-supervised + FT	Transfer Learning
Токенизация	BPE (50K)	BPE (тілге тән)	BPE / SentencePiece
Жоғалым функциясы	Cross-entropy	Contrastive + CTC	CTC + Attention
Low-resource икем.	Орташа	Жоғары	Жоғары

Талданған үш жүйе – Whisper, MMS және Turkic ASR – қырғыз-қазақ тіл жұбы үшін ASR зерттеулерінің негізгі технологиялық базасын құрайды. Whisper кең тілдік қамту мен ыңғайлы API арқылы baseline ретінде тиімді; MMS ресурсы шектеулі тілдер үшін adapter тәсілін ұсынады; Turkic ASR болса түркі тілдерінің агглютинативтік ерекшеліктерін ескеретін арнайы шешімдер жасайды.

Аталған жүйелердің математикалық негіздерін, архитектураларын және тіл қамтуын жете түсіну қырғыз-қазақ ASR жүйесін жобалауда оңтайлы модель таңдауды, transfer learning стратегиясын және гибриді декодинг тәсілін дұрыс анықтауға негіз болады.

1.2 Қырғыз-қазақ тілдік жұбы үшін заманауи нейрондық тәсілдер мен аударма жүйелері бойынша зерттеулерді талдау

Қазіргі таңда жасанды интеллект және терең нейрондық желілер технологиялары табиғи тілді өңдеу және сөйлеуді өңдеу салаларында маңызды орын алады [21]. Нейрондық тәсілдер мәтінді автоматты өңдеу, сөйлеуді тану, машиналық аударма және тыныс белгілерін қалпына келтіру сияқты міндеттерде жоғары нәтижелер көрсетуде. Әсіресе соңғы жылдары Transformer архитектурасына негізделген модельдер көптеген тілдік міндеттердің сапасын айтарлықтай жақсартты [22].

Түркі тілдері морфологиялық құрылымы күрделі агглютинативті тілдер қатарына жатады [23]. Мұндай тілдерде түбір сөзге көптеген қосымшалардың жалғануы нәтижесінде сөз формаларының саны өте үлкен болады. Бұл ерекшелік ASR және NMT жүйелерінде сөздік қордың көлемін ұлғайтып, сирек кездесетін сөздердің санын арттырады. Сондықтан қазіргі зерттеулерде субсөздік модельдеу және морфологиялық сегменттеу тәсілдері кеңінен қолданылады [24].

Сөйлеуді автоматты тану жүйелерінің алғашқы буындары Hidden Markov Model (HMM) және Gaussian Mixture Model (GMM) негізінде құрылды [25]. Алайда терең нейрондық желілердің дамуына байланысты қазіргі ASR жүйелерінде Deep Neural Network (DNN), Recurrent Neural Network (RNN), Long

Short-Term Memory (LSTM) және Transformer модельдері белсенді қолданылады [26].

End-to-end архитектуралары қазіргі заманғы ASR жүйелерінің негізгі бағытына айналды [27]. Мұндай тәсілдерде акустикалық және тілдік модельдер біртұтас нейрондық жүйе арқылы оқытылады. DeepSpeech архитектурасы end-to-end сөйлеуді тану жүйелерінің кең таралуына негіз болды [28].

Соңғы жылдары өзіндік бақылаумен оқытуға (self-supervised learning) негізделген wav2vec 2.0 моделі кеңінен қолданыла бастады [29]. Бұл модель таңбаланбаған аудио деректерді тиімді пайдалану арқылы төмен ресурсты тілдер үшін жоғары сапалы акустикалық модельдер құруға мүмкіндік береді. wav2vec 2.0 моделі түркі тілдері үшін де тиімді шешімдердің бірі болып саналады [30].

OpenAI компаниясы әзірлеген Whisper моделі көптілді сөйлеуді автоматты тану саласында жаңа кезеңнің басталуына ықпал етті [31]. Whisper моделі үлкен көлемдегі көптілді аудио деректерде оқытылған және төмен ресурсты тілдерде де тұрақты нәтижелер көрсетеді. Түркі тілдері үшін Whisper негізіндегі зерттеулер соңғы жылдары белсенді жүргізілуде [32].

Нейрондық машиналық аударма қазіргі таңда машиналық аударманың негізгі бағыты болып табылады [33]. Encoder-decoder архитектурасына негізделген алғашқы нейрондық модельдерден кейін attention механизмі енгізіліп, аударма сапасы едәуір артты [34]. Кейіннен Transformer архитектурасының пайда болуы нейрондық машиналық аударманың дамуына үлкен серпін берді [35].

Transformer архитектурасы толықтай attention механизміне негізделген және ұзақ тәуелділіктерді тиімді өңдеуге мүмкіндік береді [36]. Қазіргі NMT жүйелерінің басым бөлігі осы архитектураны пайдаланады. Түркі тілдері үшін Transformer негізіндегі модельдер қазақ–қырғыз, қазақ–өзбек және қазақ–татар тіл жұптарында кеңінен зерттелуде [37].

Төмен ресурсты тілдер үшін көптілді оқыту тәсілдері маңызды рөл атқарады [38]. Multilingual Transformer модельдері бірнеше тілдегі ортақ семантикалық және морфологиялық белгілерді пайдалану арқылы аударма сапасын арттыруға мүмкіндік береді. Әсіресе mBART, mT5 және NLLB модельдері түркі тілдері үшін тиімді құралдардың бірі болып табылады [39].

Meta AI ұсынған NLLB (No Language Left Behind) жобасы төмен ресурсты тілдер үшін сапалы машиналық аударма жүйелерін әзірлеуге бағытталған [40]. Бұл модель жүздеген тілдер арасында аударма жасауға мүмкіндік береді және түркі тілдері үшін де маңызды зерттеу құралы болып саналады.

Сөйлеуді автоматты тану жүйелерінің маңызды мәселелерінің бірі – тыныс белгілерін автоматты қалпына келтіру [41]. Көптеген ASR жүйелері мәтінді тыныс белгілерінсіз шығарады, бұл мәтінді оқуды және кейінгі өңдеуді қиындатады. Сондықтан punctuation restoration міндеті қазіргі NLP зерттеулерінде маңызды бағыттардың біріне айналды.

Тыныс белгілерін қалпына келтіру көбінесе sequence labeling міндеті ретінде қарастырылады. Бұл бағытта Bidirectional LSTM және Transformer архитектуралары негізіндегі модельдер кеңінен қолданылады. Соңғы жылдары

BERT, RoBERTa және XLM-R модельдері punctuation restoration міндетінде жоғары нәтижелер көрсетті.

Көптілді Transformer модельдері түркі тілдері үшін де тиімді қолданылуда. Әсіресе қазақ және өзбек тілдерінде punctuation restoration міндеті бойынша жүргізілген зерттеулер BERT негізіндегі модельдердің тиімділігін көрсетті.

Қазіргі уақытта ASR, NMT және punctuation restoration жүйелерін біртұтас интеллектуалды платформаға біріктіру бағытында зерттеулер жүргізілуде. Мұндай жүйелер автоматты жиналыс стенограммаларын құру, мультимедиялық контентті өңдеу және көптілді коммуникация жүйелерінде кеңінен қолданылуы мүмкін.

Машиналық аударма (Machine Translation) саласындағы соңғы онжылдықтың ең айтулы жетістігі – нейрондық аударма (Neural Machine Translation, NMT) тәсілдерінің толыққанды үстемдік алуы. Transformer архитектурасының 2017 жылы ұсынылуынан бастап аударма сапасы мен жылдамдығы түбегейлі өзгерді; ал соңғы жылдардағы ірі тілдік модельдердің (LLM) жарыса пайда болуы аударма міндетін мамандандырылған жүйелерден тысқары алып шықты. Осы тарауда қырғыз–қазақ тілдік жұбы үшін ең маңызды деп саналатын төрт модельдік парадигма – mT5, NLLB-200, Gemma/LLaMA сияқты декодер-only LLM және Helsinki Opus-MT – бір тұтас сарапшылық талдауға ұсынылады. Мұнда архитектуралар, математикалық негіздер, тіл қамту стратегиялары және эксперименталды нәтижелер бірге қарастырылады.

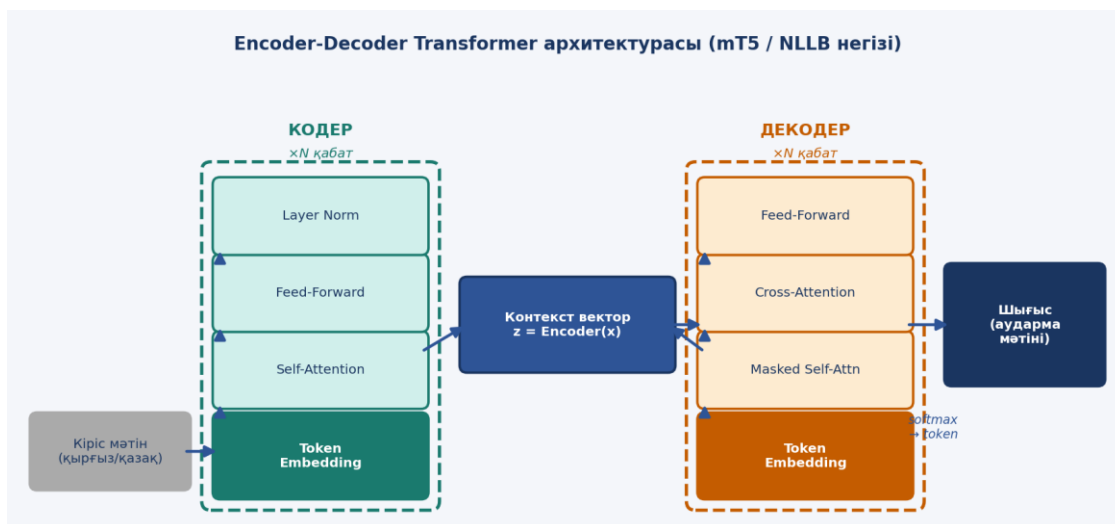
Заманауи барлық нейрондық аударма жүйелерінің ортақ атасы – Attention Is All You Need Transformer архитектурасы. Оның ең іргелі элементі – Self-Attention механизмі, ол кіріс реттіліктің кез келген екі позициясы арасындағы тәуелділікті тікелей, рекурентсіз есептейді. Transformer кодерінің (Encoder) негізгі математикасы:

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d_k}) \cdot V \quad (23)$$

мұнда Q (Query), K (Key), V (Value) – сызықтық проекциялар; d_k – кілт векторының өлшемі. Бірнеше параллель назар басы (Multi-Head Attention, MHA) бірлесіп жұмыс жасайды:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h) \cdot W^O \quad (24)$$

Позициялық ақпарат сиунустік (sinusoidal) немесе RoPE (Rotary Position Embedding) арқылы векторларға енгізіледі. Берілген негізді жете түсіну – кейінгі барлық модельдерді салыстырмалы бағалаудың кілті болып табылады. Encoder-Decoder Transformer архитектурасы Сурет 1.1-де көрсетілген. Кодер кіріс мәтінді контекст векторға сығымдайды; декодер авторегрессивті түрде аударманы генерациялайды



Сурет 1.1 – Encoder-Decoder Transformer архитектурасы (mT5 / NLLB негізі).

Google ұсынған mT5 (Multilingual Text-to-Text Transfer Transformer, T5 архитектурасының 101 тілге кеңейтілген нұсқасы болып табылады. Ол "мәтіннен мәтінге" (text-to-text) универсал парадигмасына сүйенеді: аударма, қорытындылау, сұрақ-жауап және т.б. барлық NLP міндеттері бірыңғай seq2seq форматқа түсіріледі.

mT5 бес нұсқада шығарылды: Small (300M параметр), Base (580M), Large (1.2Мд), XL (3.7Мд), XXL (13Мд). Барлық нұсқа mC4 (Common Crawl негізіндегі 101 тілдегі веб-корпус) деректерімен үйретілген. Қырғыз (ky) және қазақ (kk) тілдері ресми қолдайтын тілдер тізімінде бар; алайда корпустағы олардың үлесі (шамамен 0.03–0.12%) орыс немесе ағылшын тіліне қарағанда мәртебесі шектеулі (low-resource) деңгейде қалады.

Оқыту кезінде mT5 Span Corruption мақсатты функциясын пайдаланады: кіріс мәтіннің кездейсоқ таңдалған үзіндісі (span) маскаланып, модель сол бос орынды болжайды. Бұл алдын ала үйрету (pre-training) стратегиясы кейіннен аударма деректерімен fine-tune арқылы толықтырылады.

Туыс түркі тілдерінің қосымша fine-tune арқылы берілуі (transfer learning) mT5-тің low-resource жағдайдағы өнімділігін айтарлықтай арттырады.

Meta AI компаниясының No Language Left Behind (NLLB-200) жобасы 200 тілді бір-бірімен аударма алатын модель ұсынды. Бұл – 2010 жылдардың соңына дейін ескерусіз қалған төмен ресурсты тілдер үшін арнайы жасалған алғашқы ірі масштабты мультилингвал аударма жүйесі.

NLLB-200 архитектурасының ерекшелігі – Mixture-of-Experts (MOE) механизмін интеграциялаған Transformer. MOE-Transformer-де Feed-Forward Network (FFN) блогы бірнеше "сарапшы" (expert) желіден тұрады;

NLLB-200-дің қырғыз және қазақ тілдерін кирилл жазуымен ресми қолдайтыны, сондай-ақ барлық тіл жұптары үшін тікелей (direct) аударманы ұсынатыны – жүйенің басты артықшылығы. Pivot тілі (мысалы, ағылшын) арқылы жүру қажет емес, бұл аударма сапасын арттырады.

NLLB-200 нұсқаларының қырғыз–қазақ жұптарына қамтылуы мен модельдің әр түрлі параметрлері 1.5 -кестеде көрсетілген.

Кесте 1.5 – NLLB-200 нұсқаларының қырғыз–қазақ аударма өнімділігі (*MOE: іс жүзінде ~3.3 B белсенді параметр)

Нұсқа	Параметр саны	Тіл саны	Қырғыз	Қазақ
NLLB-200 (distilled 600M)	600M	200	✓	✓
NLLB-200 (distilled 1.3 B)	1.3 B	200	✓	✓
NLLB-200 (3.3 B)	3.3 B	200	✓	✓
NLLB-200 MOE (54.5 B)	54.5 B	200	✓	✓

Соңғы жылдары ашық қол жетімді ірі тілдік модельдердің (LLM) жарыса шығуы нейрондық аударманы жаңа бағытқа бұрды. Decoder-only архитектурасына сүйенетін Gemma (Google DeepMind), LLaMA-3 (Meta), Mistral, Falcon сияқты модельдер аударманы жасырын мақсат ретінде емес, промпт нұсқауы (instruction prompt) арқылы орындауға мүмкіндік береді. Бұл тәсілдің негізгі ерекшелігі – арнайы аударма деректерін жинаусыз, тек prompt engineering арқылы нәтиже алу мүмкіндігі.

Қырғыз–қазақ тілдік жұбының морфологиялық және лексикалық жақындығы (екі тіл де қыпшақ тобына жатады, агглютинативтік құрылымы ортақ) NMT жүйелерінің осы жұп үшін baseline-нан жоғары нәтиже беруіне негіз жасайды. Алайда деректер жетіспеушілігі мен домен ауытқуы (domain shift) мәселесі шешілмей тұр: жалпы домендегі веб-деректерде үйретілген модельдер заңдық, медициналық немесе мемлекеттік аударма текстерінде сапаны жоғалтады.

1.3 Қырғыз–қазақ тілдік жұбы үшін автоматтандырылған стенограмма қалыптастыру жүйесіне теориялық шолу

Қазіргі таңда цифрлық технологиялардың қарқынды дамуына байланысты сөйлеу ақпаратын автоматты түрде өңдеу мәселесі өзекті ғылыми бағыттардың біріне айналды [48]. Әсіресе жиналыстар, конференциялар, сұхбаттар және онлайн кездесулер барысында айтылған ақпаратты мәтіндік форматқа автоматты түрде түрлендіру жүйелеріне сұраныс артып келеді. Мұндай жүйелер автоматтандырылған стенограмма қалыптастыру жүйелері деп аталады.

Автоматтандырылған стенограмма қалыптастыру жүйесі сөйлеуді автоматты тану (ASR), мәтінді өңдеу, тыныс белгілерін қалпына келтіру, сөйлеушіні анықтау (speaker diarization) және мәтінді қысқаша мазмұндау сияқты бірнеше технологиялардың интеграциясы негізінде құрылады [49]. Бұл жүйелер

сөйлеу сигналын өңдеп, құрылымдалған мәтіндік құжат немесе жиналыс стенограммасын автоматты түрде қалыптастырады.

Стенограмма қалыптастырудың дәстүрлі тәсілдері адамның қолмен мәтін теруіне негізделген болатын [50]. Алайда мұндай әдістер көп уақытты талап етеді және үлкен көлемдегі аудио деректерді өңдеу кезінде тиімсіз болып табылады. Сондықтан қазіргі уақытта автоматтандырылған жүйелерді әзірлеу мәселесі белсенді зерттелуде.

Автоматтандырылған стенограмма жүйелерінің негізгі компоненттерінің бірі – сөйлеуді автоматты тану модулі [51]. Бұл модуль аудио сигналды мәтінге түрлендіреді. Қазіргі заманғы ASR жүйелері көбінесе нейрондық желілерге негізделген end-to-end архитектураларын пайдаланады. DeepSpeech, wav2vec 2.0 және Whisper сияқты модельдер сөйлеуді жоғары дәлдікпен тануға мүмкіндік береді [52].

Төмен ресурсты тілдер үшін ASR жүйелерін әзірлеу барысында негізгі қиындықтардың бірі – таңбаланған аудио деректердің жетіспеушілігі [53]. Қырғыз және қазақ тілдері үшін үлкен көлемдегі ашық сөйлеу корпустары әлі де шектеулі. Сондықтан қазіргі зерттеулерде transfer learning және self-supervised learning тәсілдері кеңінен қолданылады [54].

Қырғыз және қазақ тілдері агглютинативті тілдер қатарына жатады және морфологиялық құрылымы күрделі болып табылады [55]. Сөздерге көптеген қосымшалардың жалғануы нәтижесінде сөз формаларының саны ұлғаяды, бұл ASR жүйелерінің сөздік қорын күрделендіреді. Осы мәселені шешу үшін қазіргі зерттеулерде Byte Pair Encoding (BPE), SentencePiece және морфологиялық сегменттеу әдістері қолданылады [56].

Стенограмма жүйелерінің маңызды компоненттерінің бірі – punctuation restoration, яғни тыныс белгілерін автоматты қалпына келтіру модулі [57]. Көптеген ASR жүйелері мәтінді тыныс белгілерінсіз және бас әріптерсіз шығарады. Мұндай мәтінді оқу және талдау қиын болғандықтан, punctuation restoration жүйелері мәтіннің оқылымдылығын арттыруда маңызды рөл атқарады.

Тыныс белгілерін қалпына келтіру міндеті көбінесе sequence labeling есебі ретінде қарастырылады [58]. Бұл бағытта Transformer архитектурасына негізделген BERT, RoBERTa және XLM-R модельдері кеңінен қолданылады [59]. Көптілді модельдер төмен ресурсты тілдер үшін де тиімді нәтиже көрсетеді.

Автоматтандырылған стенограмма жүйелеріндегі тағы бір маңызды міндеттердің бірі – speaker diarization, яғни сөйлеушілерді автоматты түрде анықтау [60]. Бұл технология бірнеше адамның қатысуымен өтетін жиналыстарда әрбір сөйлеушінің сөзін бөлек белгілеуге мүмкіндік береді. Speaker diarization жүйелері көбінесе дауыс ерекшеліктерін талдау арқылы жұмыс істейді.

Speaker diarization міндетінде қазіргі уақытта x-vector, ECAPA-TDNN және нейрондық кластерлеу әдістері кеңінен қолданылады [61]. Мұндай тәсілдер аудио сигналдағы әртүрлі сөйлеушілерді жоғары дәлдікпен ажыратуға мүмкіндік

береді. Жиналыс стенограммаларын қалыптастыру жүйелерінде speaker diarization технологиясы маңызды компоненттердің бірі болып саналады.

Автоматтандырылған стенограмма жүйелерінде мәтінді қысқаша мазмұндау технологиялары да маңызды рөл атқарады [62]. Жиналыс барысында айтылған негізгі ойларды автоматты түрде бөліп көрсету үшін нейрондық суммаризация модельдері қолданылады. Бұл бағытта BART, T5 және mT5 модельдері кеңінен зерттелуде [63].

Қазіргі уақытта автоматты жиналыс стенограммаларын қалыптастыру жүйелері бірнеше кезеңнен тұрады: аудионы алдын ала өңдеу, сөйлеуді тану, тыныс белгілерін қалпына келтіру, сөйлеушілерді анықтау және мәтінді құрылымдау. Кейбір жүйелерде қосымша түрде кілт сөздерді анықтау және автоматты тақырып генерациялау модульдері де қолданылады.

Көптілді стенограмма жүйелері қазіргі зерттеулердің маңызды бағыттарының бірі болып табылады [64]. Әсіресе қырғыз–қазақ сияқты тілдік жағынан жақын тілдер үшін ортақ көптілді модельдерді пайдалану тиімді болып саналады. Мұндай тәсілдер деректер тапшылығы мәселесін азайтып, модельдердің жалпылау қабілетін арттырады.

Соңғы жылдары Whisper және SeamlessM4T сияқты көптілді модельдер сөйлеуді тану мен аударуды біртұтас жүйеде жүзеге асыруға мүмкіндік берді [65]. Бұл технологиялар қырғыз–қазақ тіл жұбы үшін автоматтандырылған стенограмма және жиналыс стенограммаларын қалыптастыру жүйелерін әзірлеуде маңызды құралдардың бірі болып табылады.

Қырғыз–қазақ тіл жұбы үшін автоматтандырылған стенограмма қалыптастыру жүйелерін әзірлеу ғылыми және практикалық тұрғыдан маңызды болып табылады. Мұндай жүйелер білім беру, мемлекеттік басқару, бизнес және медиа салаларында кеңінен қолданылуы мүмкін. Сонымен қатар бұл бағыттағы зерттеулер түркі тілдері үшін цифрлық ресурстарды дамытуға және төмен ресурсты тілдерге арналған интеллектуалды жүйелерді жетілдіруге ықпал етеді.

Келесі 1.6-кестеден жалпы жұмыс бойынша басқа да жасалынып жатқан жұмыстармен салыстыру келтірілген.

Кесте 1.6 – Зерттеу жұмыстарына шолу

№	Авторлар, жыл	Жұмыс атауы	Негізгі бағыт	Қолданылған технологиялар	Тілдер	Артықшылықтары	Шектеулері
1	Graves et al., 2013 [66]	Speech Recognition with Deep Recurrent Neural Networks	ASR	RNN, LSTM	Ағылшын	End-to-end ASR сапасын арттырды	Төмен ресурсты тілдер қарастырылмаған
2	Hannun et al., 2014 [67]	Deep Speech	ASR	DeepSpeech архитектурасы	Ағылшын	Үлкен көлемдегі сөйлеуді тану	Түркі тілдеріне бейімделмеген
3	Chan et al., 2016 [68]	Listen, Attend and Spell	ASR	Attention-based encoder-decoder	Ағылшын	Attention механизмі енгізді	Көптілді қолдау шектеулі
4	Sennrich et al., 2016 [69]	Neural Machine Translation of Rare Words with Subword Units	NMT	BPE сегменттеу	Неміс–ағылшын	Сирек сөздер мәселесін шешті	Агглютинативті тілдер толық зерттелмеген
5	Tilk & Alumäe, 2016 [70]	Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration	Punctuation Restoration	BiRNN, Attention	Ағылшын	Тыныс белгілерін қалпына келтіру сапасы жоғары	Түркі тілдері жоқ

6	Vaswani et al., 2017 [71]	Attention Is All You Need	NMT	Transformer	Көп тілді	Қазіргі NMT негізі	Ресурсы аз тілдерде шектеулі
7	Baevski et al., 2020 [72]	wav2vec 2.0	ASR	Self-supervised learning	Көп тілді	Төмен ресурсты тілдер үшін тиімді	Түркі тілдері бойынша дерек аз
8	Conneau et al., 2020 [73]	XLNet	Multilingual NLP	Transformer, Cross-lingual learning	100+ тіл	Көптілді модельдеу	Түркі тілдері үшін арнайы бейімделмеген
9	Radford et al., 2022 [74]	Whisper	ASR	Multilingual Transformer	90+ тіл	Көптілді жоғары дәлдік	Стенограмма құрылымдау жоқ
10	Costa-jussà et al., 2022 [75]	No Language Left Behind (NLLB)	NMT	Massive multilingual Transformer	200+ тіл	Төмен ресурсты тілдерді қолдау	Сөйлеуді өңдеу қарастырылмаған
11	Rennard et al., 2023 [76]	Automatic Meeting Summarization: A Survey	Meeting Summarization	NLP, Summarization	Ағылшын	Жиналыс қорытындысын жасау	Түркі тілдері жоқ
12	Barrault et al., 2023 [77]	SeamlessM4T	Speech Translation	Multimodal multilingual model	Көп тілді	Speech-to-text және translation біріктірілген	Қырғыз–қазақ жұбы арнайы зерттелмеген

1 3	Ұсынылып отырған зерттеу жұмысы	Қырғыз-қазақ тілдік жұбы үшін кездесу стенограммаларын жасауға арналған автоматтан дырылған жүйенің моделі мен құралдарын әзірлеу және зерттеу	Кешенді интеллектуалды жүйе (ASR + map + speaker + transcripts)	Whisper, MMS, Nllb-200	Қырғыз-қазақ	Түркі тілдеріне бағытталған кешенді жүйе; көптілді ASR; сөйлеушілерді анықтау; жиналыс стенограммасын автоматты генерациялау; төмен ресурсты тілдерге бейімделу	Арнайы корпус жинауды және модельдерді бейімдеуді талап етеді
--------	---------------------------------	--	---	------------------------	--------------	---	---

2 ҚЫРҒЫЗ–ҚАЗАҚ ТІЛДІК ЖҰБЫ ҮШІН АВТОМАТТАНДЫРЫЛҒАН ЖҮЙЕНІ ӘЗІРЛЕУ ӘДІСТЕМЕСІ

Қырғыз–қазақ тіл жұбы үшін автоматтандырылған жүйені әзірлеу бірнеше кезеңді қамтитын күрделі инженерлік процесс болып табылады. Ұсынылған жүйе сөйлеуді автоматты тану, машиналық аударма және стенограмма құжатын қалыптастыру модульдерінен тұрады. Автоматтандыру және заттар интернеті саласының талаптарына сай, жүйе нақты уақыт режимінде жұмыс істеуге, аппараттық құрылғылармен интеграциялануға және минималды адам қатысуымен толыққанды жұмыс атқаруға бағытталған.

Жүйенің жұмыс принципі үш негізгі кезеңге негізделген. Бірінші кезеңде қырғыз тіліндегі сөйлеу ағыны микрофон немесе IoT құрылғысы арқылы жиналып, сөйлеуді тану моделіне беріледі. Бұл модель акустикалық сигналды өңдеп, оны мәтіндік пішімге түрлендіреді. Екінші кезеңде танылған қырғыз тіліндегі мәтін автоматты аударма жүйесі арқылы қазақ тіліне аударылады. Үшінші кезеңде аударылған мәтін белгіленген үлгі бойынша автоматты түрде стенограмма құжаты ретінде қалыптастырылады.

Автоматтандырылған жүйе ретінде оның басты артықшылығы – процестің толықтай бағдарламалық басқаруда жүруі және адам факторының ықпалын барынша азайтуы. Заттар интернеті концепциясымен үйлесімділігі тұрғысынан жүйе желілік протоколдар арқылы басқа ақылды құрылғылармен және платформалармен деректер алмасуға қабілетті болуы тиіс. Осылайша, ұсынылатын шешім жеке бағдарламалық өнім ғана емес, кең ауқымды автоматтандырылған инфрақұрылымның бір буыны ретінде қарастырылады.

2.1 Автоматтандырылған жүйенің архитектурасы

Ұсынылатын жүйе үш негізгі функционалдық блоктан тұратын конвейерлік (pipeline) архитектурасына негізделген: сөйлеуді автоматты тану (Speech-to-Text, STT), машиналық аударма (Machine Translation, MT) және стенограмма қалыптастыру модулі. Заттар интернеті (IoT) парадигмасына сәйкес жүйе нақты уақыт режимінде жұмыс істеп, аппараттық сенсорлармен және желілік протоколдармен интеграцияланады.

Жүйенің жұмыс принципі үш кезеңге негізделген: кіріс ретінде дайын .wav / .mp3 аудио файлдары немесе микрофон / IoT аудио сенсоры арқылы нақты уақыт режиміндегі сөйлеу ағыны қабылданады, содан кейін танылған қырғыз мәтіні MT моделі арқылы қазақ тіліне аударылады, ақырында аударылған мәтін стенограмма пішімінде сақталады. Жүйенің архитектуралық кезеңдері төмендегі кестеде көрсетілген:

Кесте 2.1 – Жүйенің конвейерлік архитектурасы

Қадам	Кезең	Сипаттама
1	Аудио кірісі	Екі режим қолданылады: (1) дайын .wav / .mp3 аудио файлын жүктеу, немесе (2) микрофон /

		IoT аудио сенсоры арқылы нақты уақыт режимінде жазу. Екі жағдайда да сигнал 16 кГц моно WAV форматына айналдырылады.
2	Сөйлеуді тану (STT)	Whisper, MMS немесе TurKic-ASR моделі акустикалық ерекшеліктерден қырғыз тіліндегі мәтінді генерациялайды.
3	Мәтінді түзету	ASR технологиясы негізіндегі MAPR
4	Машиналық аударма (MT)	Танылған қырғыз мәтіні NLLB моделі арқылы қазақ тіліне аударылады.
5	Стенограмма шығысы	Аударылған мәтін .docx немесе .txt пішімінде сақталады. IoT желісіне MQTT арқылы жіберіледі.

Жүйенің жұмыс принципі аудио кірісінен бастап дайын стенограммаға дейінгі толық конвейерді қамтиды.

Кіріс сигналы уақытша домендегі аналогтық дыбыстық толқын болып табылады. Цифрландыру нәтижесінде аналогтық сигнал секундына 16 000 үлгі жиілігінде дискретті сигналға айналдырылады. Бұл жиілік сөйлеу сигналын жеткілікті дәлдікпен сақтай отырып, есептеу шығынын оңтайлы деңгейде ұстауға мүмкіндік береді.

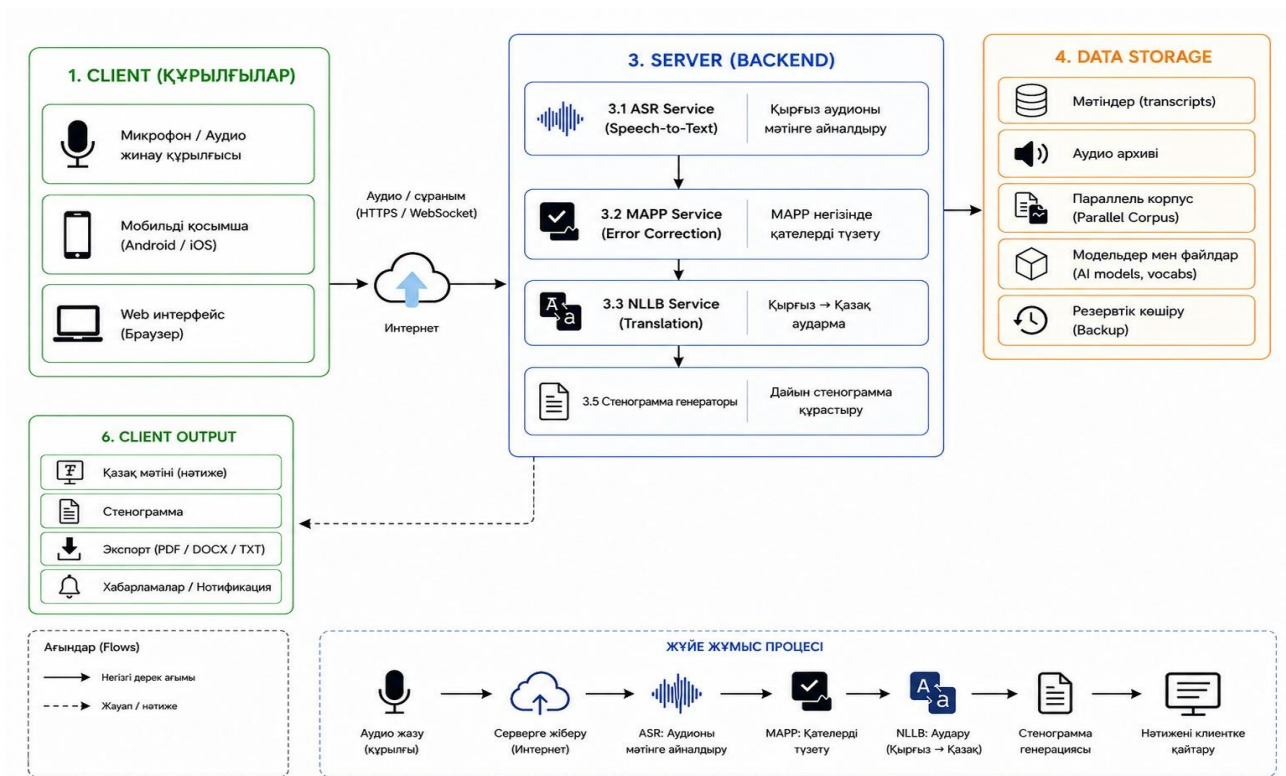
Акустикалық ерекшеліктер алу үшін лог-мел-спектрограмма есептеледі. Бұл процесте дискретті сигнал қысқа уақыттық Фурье түрлендіруі арқылы уақыт-жиілік кеңістігіне ауысады, содан кейін мел-фильтрлер банкі арқылы адам есітуіне жақын жиілік шкаласына түрленеді. Мел-фильтрлер саны әдетте 80 немесе 128 болып белгіленеді. Алынған спектрограмма STT моделіне кіріс ретінде беріледі.

Сөйлеуді тану моделі акустикалық ерекшеліктер матрицасынан мәтіндік жетондар тізбегін генерациялайды. Модель үйретілген параметрлері негізінде берілген акустикалық белгілерге сәйкес ең ықтимал мәтіндік тізбекті таңдайды. Нәтижесінде қырғыз тіліндегі шикі транскрипция алынады.

Машиналық аударма кезеңінде Transformer архитектурасы қолданылады. Алынған қырғыз тіліндегі мәтін аударма моделіне кіріс ретінде беріліп, қазақ тіліндегі баламасы генерацияланады. Аударма моделі де үйретілген параметрлері негізінде ең ықтимал аударма нұсқасын шығарады.

Соңғы кезеңде аударылған мәтін уақыт белгісімен және диктор идентификаторымен жұптасып, толық стенограмма қалыптасады. Стенограмма сөйлеу сегменттерінің жиынынан тұрады: әрбір сегментте сөйлеу басталған уақыт белгісі, сөйлеуші диктордың идентификаторы және аударылған мәтін бірге сақталады. Сегменттер саны жиналыс ұзақтығы мен сөйлеушілер санына байланысты өзгереді. Осылайша жүйе аудио кірісінен бастап уақыт белгіленген, дикторы анықталған және екі тілде ұсынылған толық стенограммаға дейінгі автоматты конвейерді жүзеге асырады.

Ұсынылған жүйенің жалпы архитектурасы 2.1-суретте көрсетілген.

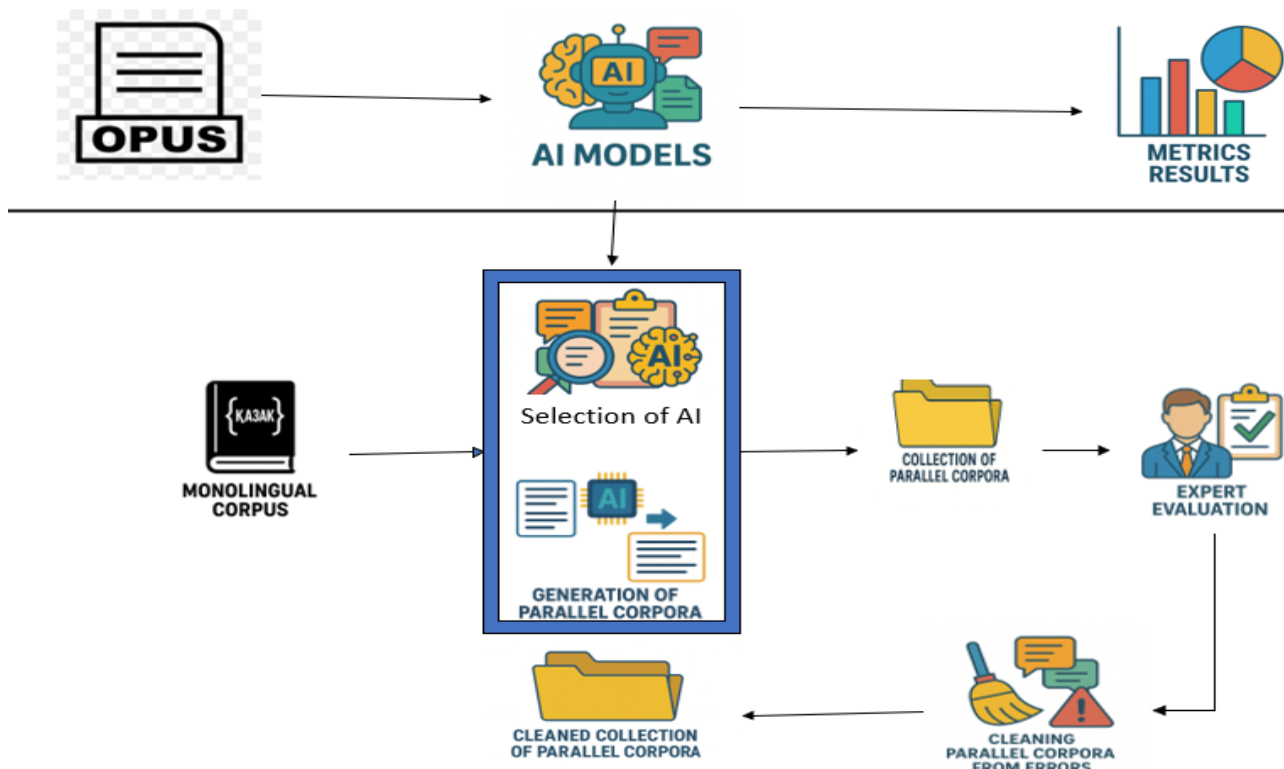


Сурет 2.1 – жүйенің архитектурасы

Ұсынылған жүйе клиенттік құрылғылардан (микрофон, мобильді қосымша, веб интерфейс) аудио деректерін қабылдап, серверлік бөлімде бірнеше кезең бойынша өңдейді. Алдымен ASR сервисі арқылы қырғыз тіліндегі аудио мәтінге түрлендіріледі, кейін MAPP модулі мәтіндегі қателерді түзетеді. Одан кейін NLLB моделі арқылы мәтін қырғыз тілінен қазақ тіліне аударылады және дайын стенограмма қалыптастырылады. Өңделген нәтижелер деректер қоймасында сақталып, пайдаланушыға мәтін, стенограмма немесе экспортталған файл түрінде қайтарылады.

2.2 Қырғыз–қазақ тілдері үшін деректерді жинау, өңдеу және талдау әдістері

Зерттеу жұмысы үш негізгі кезеңнен тұрады: дайындық (Preparation), жасанды интеллект жүйесін таңдау (AI system selection) және корпус жасау (Corpus creation). Осы кезеңдердің әрқайсысы өзара тығыз байланысты және дәйектілікпен орындалады. Келесі Сурет 2.2-де параллель корпустарды дайындау үшін жасанды интеллект құралдарын таңдау, таңдалынған құрал арқылы корпусты жасау және бағалау метрикаларын алып, дайын корпусты алу кезеңдері көрсетілген [45].



Сурет 2.2 – Параллель корпус жасау үшін жасанды интеллект құралын таңдау және оны жасау кезеңдері

2.2.1. Бірінші кезең – Дайындық

Бірінші кезеңде келесі іс-шаралар жүзеге асырылды:

- Әдебиеттерге шолу жасау – аз ресурсты тілдер үшін машиналық аударма саласындағы зерттеулердің қазіргі жағдайына кешенді шолу жүргізу, қазақ-қырғыз тіл жұбына және байланысты технологияларға баса назар аудару.
- Қырғыз-қазақ тілдерінің параллель сөйлем деректер жиынтықтарын іздеу – ашық деректер көздерінен, мысалы OPUS сияқты қоймалардан, қазақ-қырғыз және қазақ-ағылшын тіл жұптары бойынша қолда бар параллель корпустарды анықтау және олардың сапасы мен қамтымасын бағалау.
- Жасанды интеллект жүйелерін іздеу – қазақ-қырғыз тіл жұбы үшін қолдануға болатын жасанды интеллект жүйелері мен нейрондық машиналық аударма модельдерін зерттеу, алдын ала үйретілген модельдер мен ашық бастапқы шешімдерге баса назар аудару.
- Аударма сапасын бағалау метрикаларын таңдау – аударма жүйелерінің сапасы мен өнімділігін бағалау үшін әдебиеттерге шолу негізінде тиісті бағалау метрикаларын анықтау.
- Жасанды интеллект жүйесін таңдау критерийлерін анықтау – аударма тапсырмалары үшін ең қолайлы жасанды интеллект моделін таңдауда нақты критерийлер белгілеу; модель архитектурасы, тиімділігі, оқыту деректерінің қолжетімділігі және аударма дәлдігі сияқты факторларды ескеру.

- Сынақ деректер жиынтығын жасау – таңдалған көздерден параллель сөйлемдердің сынақ деректер жиынтығын жасау; аударма сапасын кешенді бағалауды қамтамасыз ету үшін оның теңгерімді және әртүрлі салаларды қамтитындай болуын қадағалау.

Екінші кезең жеке жасанды интеллект жүйелері арқылы аударма сапасын тексерудің қайталанбалы орындалуынан тұрады: қазақ сынақ деректер жиынтығының бөлігін қырғыз тіліне аудару және аударма сапасының метрикалық мәндерін есептеу. Осы кезеңде мүмкіндіктерді кезең-кезеңмен жою әдісі арқылы жүйелерді бағалау жүргізіліп, ең тиімді жасанды интеллект жүйесі таңдалды.

Үшінші кезең – корпус жасау – Қырғыз-қазақ корпусының қазақ бөлігін қырғыз тіліне аудару процесінен тұрады. Аударма жұмысы барысында NVIDIA RTX 4090 (24 ГБ VRAM) графикалық картасымен, 128 ГБ DDR4/DDR5 жедел жадымен және 10 Гбит/с желілік интерфейспен жабдықталған жоғары өнімді есептеу жүйесі пайдаланылды.

2.2.2 Аударма сапасын бағалау метрикалары

Аударма сапасын бағалау метрикалары машиналық аударманың тиімділігін өлшеу үшін қажет. Аударма сапасын бағалауға арналған кең тараған метрикаларға BLEU, TER және WER жатады. Алайда бұл метрикалар жиі шектеулі бағалауды қамтамасыз ететіндіктен, синтаксистік және семантикалық дәлдіктің қосымша метрикалары – COMET және chrF – аударма сапасын неғұрлым кешенді бағалауды ұсыну үшін де қолданылды.

Аударылған мәтіннің сапасын бағалау үшін келесі метрикалар пайдаланылды:

- BLEU (Bilingual Evaluation Understudy) – машинамен жасалған аударма мен эталондық аударма арасындағы сәйкестікті өлшейді. Python тіліндегі SacreBLEU кітапханасының `sentence_bleu` функциясы арқылы сөйлем деңгейіндегі BLEU ұпайлары есептелді.
- WER (Word Error Rate) – аударылған мәтіндегі қателер санын (кірістірулер, жоюлар және ауыстырулар) өлшейді. Төмен WER аударма дәлдігінің жоғары екенін көрсетеді.
- ChrF (Character n-gram F-score) – аударманы таңба деңгейінде бағалайды, бұл морфологиялық жағынан бай тілдерді бағалау үшін өте пайдалы. Бөлшектік деңгейдегі ChrF ұпайларын есептеу үшін `sacrebleu.sentence_ChrF` функциясы пайдаланылды, содан кейін олар орташалай есептелді.
- METEOR (Metric for Evaluation of Translation with Explicit ORdering) – аударманы синонимия, тамыр сөз және сөз тәртібі негізінде бағалайды.
- COMET – семантикалық жеткіліктілік пен сөйлеу мәнерін ескеретін нейрондық негіздегі бағалау метрикасы.

Төмендегі 2.2-ші кестеде бағалау метрикаларының түсіндірмесі және оның диапозондарда қандай көрсеткішке ие екені көрсетілген.

Кесте 2.2 – Бағалау метрикаларын түсіндіру

Метрика	Мән диапазоны	Интерпретация
BLEU	0–100	>50 – Өте жақсы; 30–50 – Жақсы; 10–30 – Қолайлы; <10 – Нашар
WER	0–1	<0.2 – Өте жақсы; 0.2–0.4 – Қолайлы; >0.4 – Нашар
ChrF	0–100	>60 – Өте жақсы; 40–60 – Жақсы; <40 – Әлсіз
METEOR	0–1	>0.5 – Өте жақсы; 0.3–0.5 – Орташа; <0.3 – Төмен
COMET	0–1	>0.5 – Жоғары сапа; 0.3–0.5 – Қолайлы; <0.3 – Әлсіз

2.2.3 Деректер жиынтықтарын анықтау және сынақ деректерін дайындау

Таңдалған аударма модельдерін бағалау үшін OPUS қоймасынан параллель деректер жиынтығы алынды. Осы деректер жиынтығынан 1 000 қазақ-қырғыз сөйлем жұбы таңдалды. Қырғыз бөлігі алтын эталон ретінде қарастырылды, ал қазақ сөйлемдері аударма үшін бастапқы мәтін ретінде пайдаланылды.

Қолда бар ресурстарды тексеру барысында OPUS-та қазақ және қырғыз тілдерін де қамтитын бірнеше ондаған деректер жиынтығы бар екені анықталды. Алайда солардың ішінен машиналық аударма тапсырмаларына жарамды тек үшеуі ғана болды, ол келесі 2.3-кестеде келтірілген.

Кесте 2.3 – Қырғыз-қазақ тілдеріндегі машиналық аударма үшін параллель деректер жиынтықтары

Деректер жиынтығының атауы	Сілтеме	Параллель сөйлемдер саны
OPUS	https://opus.nlpl.eu/results/kk&ky/	102 345
NTREX	https://huggingface.co/datasets/davidstap/NTREX	2 000
Flores 101	https://huggingface.co/datasets/severo/flores_101	2 000

Аз ресурсты қазақ-қырғыз жұбы үшін жоғары сапалы параллель корпус жасау мақсатында алғашқы қадам ретінде шамамен 300 000 сөйлемнен тұратын бар қазақ тілді монолингвалды деректер жиынтығы таңдалды. Зерттеуде пайдаланылған қазақ корпусы ашық қолжетімді деректер жиынтығынан алынды. Бастапқы деректер жиынтығы 302 530 параллель сөйлем жұбынан тұрады (қазақ және ағылшын тілдерінде). Осы зерттеу мақсаттары үшін корпусың тек қазақ бөлігі алынып, машиналық аударма арқылы қазақ-қырғыз параллель корпусын жасауда бастапқы деректер ретінде пайдаланылды.

2.2.4. Жасанды интеллект жүйелері және таңдау критерийлері

Бұл тапсырма үшін бірнеше аударма жүйесі қарастырылып, алдын ала белгіленген критерийлер жиынтығына сәйкес бағаланды.

Таңдау критерийлері (маңыздылық ретімен):

Қолжетімділік (Accessibility) – кең ауқымды пайдалану үшін жүйенің қолжетімділігі; тегін немесе ашық қолжетімді жүйелерге артықшылық беріледі.

Аударма сапасы (Translation quality) – шығарылатын мәтіннің лингвистикалық дәлдігі мен контекстуалдық сәйкестігі.

Аударма тиімділігі (Translation efficiency) – жүйенің үлкен көлемдегі мәтінді өңдеп, аударма алу қабілеті.

Интернеттегі іздеу нәтижелеріне сүйене отырып, салыстыру және бағалау үшін келесі аударма жүйелері таңдалды: Google Translate, NLLB, ChatGPT, DeepSeek, Copilot және Gemma.

Google Translate – нейрондық желілерді пайдаланып, екітілді деректердің үлкен көлеміне негізделген аударма модельдерін оқытатын кең тараған аударма құралы. 2006 жылдан бері жұмыс жасайтын жүйе GNMT (Google Neural Machine Translation) технологиясын енгізгеннен кейін айтарлықтай жетілді және бүгінгі таңда 100-ден астам тілді қолдайды.

NLLB (No Language Left Behind) – Meta компаниясы жасаған, аз ресурсты тілдерді қоса алғанда, көп тілді қолдауға арналған машиналық аударма моделі. Аударма сапасын жақсарту үшін терең оқыту мен трансформатор негіздеріне сүйенеді.

ChatGPT (GPT) – OpenAI компаниясы жасаған, NLP-дегі трансформаторларды тиімді қолдану мысалдарының бірі. Аударма қызметін де, мәтін генерациялауды да, диалог жүйелерін де орындайды.

DeepSeek – күрделі нейрондық желі архитектураларын қолдана отырып, аударма сапасын арттыруға бағытталған машиналық аударма жүйесі. Мультиязыкалық оқыту арқылы әртүрлі мәтін түрлерін өңдейді.

Copilot – GitHub мен OpenAI бірлесіп жасаған, OpenAI Codex моделіне негізделген жасанды интеллект құралы. Мәтіндік нұсқаулар негізінде әртүрлі бағдарламалау тілдерінде код генерациялауға мүмкіндік береді.

Gemma – Google компаниясы 2024 жылы ұсынған, Gemini модельдері технологиясына негізделген жеңіл ашық бастапқы үлкен тілдік модельдер (LLM) отбасы. Екі нұсқада – 2 және 7 миллиард параметрде – алдын ала оқытылған және бапталған бақылау нұсқаларымен қол жетімді.

2.2.5. Синтетикалық параллель корпус жасау

2.2.5.1. Корпус жасаудың жалпы сипаттамасы

Зерттеуде Қазақстан ғылым және жоғары білім министрлігінің AP05131415 гранттық жобасы шеңберінде жиналған 300 000 сөйлемнен тұратын қазақ–ағылшын корпусы және Синьцзян физика-химия техникалық

институтымен (Қытай) бірлескен гранттық жоба аясында жасалған 500 000 сөйлемді қамтитын қазақ–ағылшын корпусы пайдаланылды. Осы корпустардың қазақ бөлігі бастапқы деректер ретінде алынып, NLLB-200 3.3B моделі арқылы қырғыз тіліне аударылды. 2.4-ші кестеде қазақ монолингвалдық корпусының сипаттамасы келтірілген.

Кесте 2.4 –Қазақ-ағылшын тіліндегі корпусының сипаттамасы

Параллель сөйлемдер саны	Тақырыптық қамту
~300 000	Жаңалықтар, мемлекеттік ресми стиль
~500 000	Медициналық, техникалық, ауызша стиль

2.2.5.2. Аударма жүйесін таңдау

Параллель корпус жасау үшін ең қолайлы аударма жүйесін таңдау мақсатында бірнеше модель бағаланды. Коммерциялық жүйелер (Google Translate, GPT-4o, Copilot) қолжетімділік мәселелеріне байланысты алынып тасталды. Ашық бастапқы модельдер ішінен Gemma 2 27B және NLLB-200 3.3B толық сынақтан өтті.

NLLB-200 3.3B моделі соңғы таңдау болды, өйткені:

Ол аз ресурсты тілдерге арнайы жасалған және Gated Mixture of Experts (MoE) архитектурасын пайдаланады.

Морфологиялық жағынан бай тілдер үшін тұрақты аударма сапасын қамтамасыз етеді: жалғауларды, септік жалғауларын және сөз тәртібін дұрыс береді.

Түркі тілдерінің барлығын тікелей немесе жанама түрде қолдайды.

Gemma 2 27B моделімен салыстырғанда айтарлықтай жылдамырақ: 300 сөйлем/мин (Gemma: 3 сөйлем/мин). 302 530 сөйлемді аударуға Gemma-ға ~2,5 ай, ал NLLB-200 3.3B-ге тек ~2 күн кетеді.

2.2.6. Қырғыз–қазақ корпустарын қателерден тазарту

2.2.6.1. Анықталған қате түрлері

NLLB-200 3.3B моделімен жасалған синтетикалық корпустарды қолмен тексеру барысында қырғыз–қазақ тіл жұбында жүйелі қателер анықталды. Жалпы барлық тіл жұптарында байқалған қате категориялары:

Лексикалық қайталанулар – бір сөйлемдің ішінде немесе сөйлемдер арасында қайталанатын фрагменттер.

Географиялық атаулар мен жалқы есімдерді қате беру.

Шет тіліндегі қысқартуларды шифрлаусыз немесе бейімдеусіз пайдалану.

Ішінара аударылған фразалар, семантикалық бірліктердің кесілуі, синтаксистік құрылым тұтастығының бұзылуы.

2.2.6.2. Автоматты тазарту модулі

Синтетикалық параллель деректерді автоматты тазарту және нормаландыру үшін Python 3.11.14 тілінде арнайы Түзету модулі (Correction Module) жасалды. Модуль нейрондық модельдерді, ендіруге негізделген ұқсастықты немесе сыртқы атаулы тұлғаларды тану (NER) жүйелерін пайдаланбайды, бұл нәтижелердің толық ашықтығы мен жоғары қайта жаңғырымдылығын қамтамасыз етеді.

Модуль жұмысы төрт дәйекті кезеңнен тұрады:

1. Мәтінді нормаландыру – бос орындарды түзету, таңбалар мен тыныс белгілерін стандарттау.
2. Қайталанулар мен галлюцинацияларды сүзгілеу – қайталану ілмектерін анықтау және жою.
3. Жалқы есімдер мен қысқартуларды ережеге негізделген түзету.
4. Тазартылған корпусты сақтау және өзгерістер журналын жүргізу.

2.2.6.3. Қырғыз–қазақ корпусын тазарту нәтижелері

Барлық бес тіл жұбы бойынша жүргізілген тазарту жұмыстарының сандық нәтижелері төменде берілген. Кесте 2.5 қырғыз–қазақ тіл жұбы үшін жасалған синтетикалық корпустарды тазарту статистикасын ұсынады.

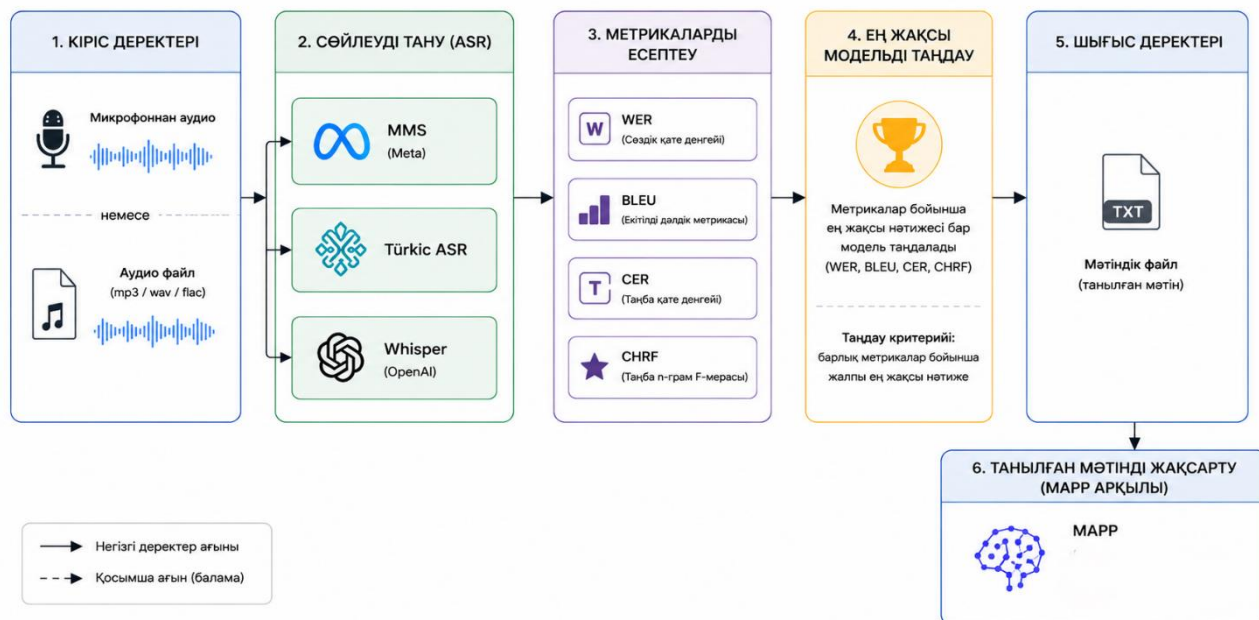
Кесте 2.5 – Қырғыз–қазақ тіл жұбы бойынша корпусты тазарту нәтижелері

Параметр	300 000 сөйлем корпусы	500 000 сөйлем корпусы
Бастапқы сөйлемдер саны	300 000	500 000
Тазартылғаннан кейінгі сөйлемдер саны	~282 240	~487 060
Жойылған сөйлемдер (қайталанулар)	жалпы ~2–4%	жалпы ~2–4%
Түзетілген атаулар/қысқартулар	қырғыз: ҚР → Қырғызстан қателерін түзету	қырғыз: ҚР → Қырғызстан қателерін түзету

Қырғыз–қазақ тіл жұбы үшін жиі кездескен қателердің бірі – «ҚР» (Қазақстан Республикасы) қысқартуының «Қырғызстан» деп аударылуы. Модуль осы жүйелі қатені автоматты түрде анықтап, «ҚР» (Қырғыз Республикасы) немесе «Қазақстан» деп дұрыс жазуға түзеткен.

2.3 Сөйлеуді тану модельдерін зерттеу кезеңдері

Осы бөлімде қырғыз тілінде сөйлеуді автоматты тануға (ASR) арналған бірнеше ашық бастапқы нейрондық модельдердің салыстырмалы талдауы ұсынылады. Зерттеудің мақсаты – қырғыз тілі үшін ең тиімді ASR моделін таңдап, оны кейіннен қазақ тіліне аудару конвейеріне біріктіру. ASR модельді таңдау бойынша жұмыстар 2.3 – суретте көрсетілгендей жүзеге асырылды.



Сурет 2.3 - ASR модельді таңдау және сөйлеуді тану архитектурасы

2.3.1. Зерттелген ASR модельдері

Зерттеуде үш ашық бастапқы нейрондық сөйлеуді тану моделі бағаланды [43], 2.6 – кестеде модельдер бойынша параметрлері мен ерекшеліктері көрсетіліп, талдаулар жүргізілді.

Кесте 2.6 – Бағаланған ASR модельдерінің жалпы сипаттамасы

Модель	Әзірлеуші	Параметрлер саны	Ерекшелігі
Whisper (large)	OpenAI	1 550M	Transformer, 680 000 сағ аудиоға үйретілген, 99 тілді қолдайды
MMS	Meta (Facebook AI)	1 000M (1B)	Massively Multilingual Speech, 1 400+ тілді қолдайды
TurkicASR	ISSAI (HY)	42.98M	10 түркі тілі + орыс + ағылшын

			тілдерін тану үшін арнайы бапталған
--	--	--	---

Барлық модельдер үшін жарияланған салмақтар пайдаланылды, бұл нәтижелердің қайта жаңғырымдылығы мен дәйектілігін қамтамасыз етті. Whisper үшін ең үлкен нұсқа (large, 1550M параметр) пайдаланылды. MMS тек бір нұсқада (1B) қолжетімді. TurkicASR үшін 10 түркі тілін тани алатын алдын ала үйретілген модель таңдалды [44].

2.3.2. Деректер жиынтығы және бағалау протоколы

Тәжірибелер Mozilla Common Voice ашық дереккөзінен алынған деректерде жүргізілді. Деректер жиынтығының қырғыз бөлімінен 1 000 сөйлем таңдалды. Common Voice деректері барлық зерттеушілерге қолжетімді болғандықтан, тәжірибелер өз бетінше қайта жаңғыртылуы мүмкін.

Бағалау метрикалары:

- WER (Word Error Rate) – сөз деңгейіндегі қателер үлесі; кірістірулер, жоюлар және алмастырулар ескеріледі.
- CER (Character Error Rate) – таңба деңгейіндегі қателер үлесі; морфологиялық жағынан бай тілдер үшін аса маңызды.

2.3.3. Қырғыз тілі бойынша модельдерді салыстыру нәтижелері

Алты түркі тілі (Өзірбайжан, Қазақ, Қырғыз, Түрік, Түрікмен, Өзбек) үшін үш модель салыстырылды. Мақала авторлары өз тәжірибелерінің нәтижесінде қырғыз тілі үшін MMS моделін таңдады. Алайда бұл таңдау негізінен жалпы метрика орташасына, сондай-ақ мақала жазылу кезіндегі модель нұсқаларына сүйенді [43].

Кесте 2.7 –Қырғыз тілі үшін ASR модельдерінің WER/CER нәтижелері

Модель	WER	CER	Бағалау	Шешім
Whisper (large)	1.10	0.51	Орташа нәтиже, ірі модель	2-орын
MMS (1B)	0.25	0.07	Жаксы нәтиже, ашық қолжетімді	Таңдалды ✓
TurkicASR	1.52	0.88	Шағын модель, түркі тілдеріне арнайы	3-орын

Осы жұмыс шеңберінде жүргізілген қосымша тәжірибелер MMS моделінің қырғыз тілі үшін Whisper (large) моделімен салыстырғанда жоғары нәтиже беретінін анықтады. Бұл айырмашылық бірнеше факторға байланысты:

MMS нұсқасы 680 000 сағаттық аудиоға алдын ала үйретілген, бұл оның акустикалық модельдеу қабілетін айтарлықтай арттырады. MMS қырғыз тілінің дыбыстық жүйесін – ерін үндестігін ($o \leftrightarrow \text{өө}$, $y \leftrightarrow \text{үү}$), $n \leftrightarrow \text{ң}$ дыбыс шатасуларын – MMS-ке қарағанда тиімдірек өңдейді.

MMS морфологиялық жағынан күрделі агглютинативті формаларды жақсы транскрипциялайды, бұл кейінгі дұрыстау кезеңінің жүктемесін азайтады.

MMS нұсқасы бірнеше эксперименттік іске қосу бойынша тұрақты нәтиже береді.

Кесте 2.8– Қосымша зерттеу нәтижелері: Whisper (large) пен MMS салыстырмасы (қырғыз тілі, Common Voice)

Модель	WER (%)	CER (%)	Морфол. дәлдік	Таңдау
Whisper (large)	Жоғарырақ	жоғарырақ	Орташа	–
MMS (1B)	төмен ✓	төмен ✓	жоғары ✓	Таңдалды ✓
TurkicASR	ең жоғары	ең жоғары	шектеулі	–

2.3.4. MMS-1b-all моделін таңдаудың негіздемесі

Қырғыз тілі үшін MMS-1b-all моделін таңдаудың басты себептері:

Кесте 2.9 – MMS-1b-all моделін таңдаудың негіздемесі

№	Критерий	MMS-1b-all артықшылығы
1	Аударма сапасы	WER және CER мәндері үш модель ішінде ең төмен
2	Морфологиялық дәлдік	Қырғыз тіліне тән дауысты үндестікті, ерін үндестігін дұрыс таниды
3	Фонетикалық жабдықтау	$n \leftrightarrow \text{ң}$, $o \leftrightarrow \text{өө}$, $y \leftrightarrow \text{үү}$ шатасуларын MMS-ке қарағанда жақсы өңдейді
4	Тұрақтылық	Бірнеше іске қосу бойынша нәтижелер тұрақты, ауытқуы аз
5	Интеграция мүмкіндігі	NLLB-200 аударма моделімен конвейерге оңай біріктіріледі
6	Ашық қолжетімділік	Тегін, ашық бастапқы; HuggingFace арқылы қолжетімді

2.3.5. Кыргыз сөйлеуін танудағы басты қиындықтар

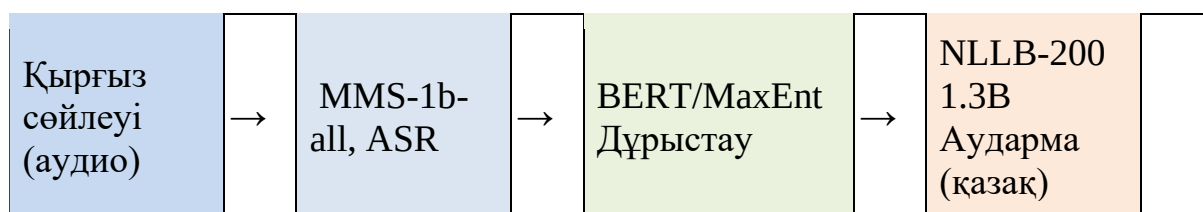
MMS-1b-all моделі таңдалса да, кыргыз тілі үшін ASR жүйесі бірнеше лингвистикалық сын-қатерлерге тап болады:

Кесте 2.10 – Кыргыз тілі үшін ASR қиындықтары және олардың дұрыстау конвейеріне әсері

Қиындық түрі	Мысал	Дұрыстау конвейеріндегі шешімі
Ерекше таңбалар жоғалуы	$n \rightarrow \text{ң}, o \rightarrow \text{ө}, y \rightarrow \text{ү}$	BERT детекторы + CONFUSION белгілері
Ұзын дауыстыларды қысқарту	$yy \rightarrow y, oo \rightarrow o, \theta\theta \rightarrow \theta$	LONG_VOWEL_RESTORE белгісі
Жалғау үндестік кателері	<i>мектептын</i> → <i>мектептин</i>	SUFFIX_REWRITE + VOWEL_HARMONY_OK
Қосылықты дыбыс (геминация)	<i>сабака</i> → <i>сабакка</i>	GEMINATION белгісі
Агглютинативті ұзын сөздер	<i>оқуучуларыбыз дегендей</i>	MorphGen + MORPH_VALID белгісі

2.3.6. ASR → аударма конвейерінің жалпы сұлбасы

Таңдалған MMS-1b-all моделі NLLB-200 1.3B машиналық аударма моделімен бірге интеграцияланып, кыргыз сөйлеуінен қазақ мәтінін алуға мүмкіндік беретін толық конвейер құрайды:



Сурет 2.4 – Кыргыз сөйлеуінен қазақ мәтінін алудың жүйесі

Конвейердің кезеңдері:

1. MMS-1b-all – кыргыз тілінде айтылған сөйлемді мәтінге айналдырады (транскрипция).
2. BERT детекторы мен MaxEnt перанкері – транскрипциядағы фонетикалық және морфологиялық кателерді анықтап, дұрыстайды (5-бөлімде сипатталған).
3. NLLB-200 1.3B – дұрысталған кыргыз мәтінін қазақ тіліне аударады (4-бөлімде сипатталған).

Тәжірибелер барысында байқалған негізгі ерекшеліктер:

- Common Voice кыргыз деректерінде бастапқыда дублика жазбалары болды; деректер тазартылып, 6,5 сағаттық ғана нақты деректер

пайдаланылды – бұл кез келген ASR зерттеуі үшін деректер сапасының маңыздылығын дәлелдейді.

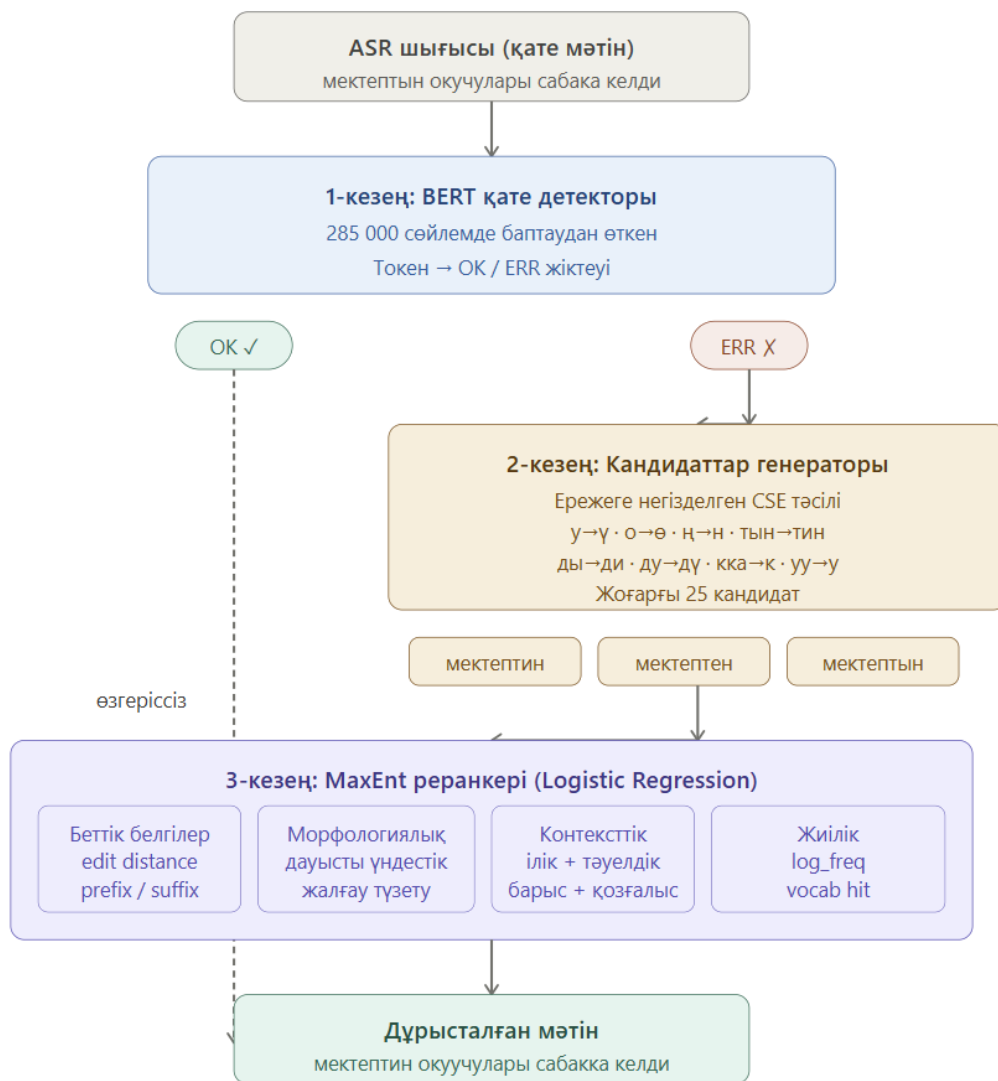
- MMS-1b-all нұсқасы кіші нұсқаларға карағанда WER мен CER бойынша айтарлықтай жақсы нәтиже береді: модель өлшемі ұлғайған сайын сапа артады.
- Қырғыз тілінің морфологиялық күрделілігі (агглютинативті құрылым) CER метрикасын WER-ге карағанда маңыздырақ етеді, өйткені ұзын сөздердегі жалғыр таңба қатесі бүкіл сөзді дұрыс емес деп санауы мүмкін.
- MMS-1b-all + дұрыстау конвейерінің бірлескен нәтижесі жеке MMS-1b-all нәтижесімен салыстырғанда айтарлықтай жақсырақ – бұл кейінгі дұрыстау кезеңінің маңыздылығын растайды.

2.4 Сөйлеуді танудан алынған деректерді түзету Morphology Aware Post Processing әдісі

Бұл бөлімде қырғыз тілі үшін автоматты сөйлеуді тану (ASR) жүйесінің шығарған мәтінін дұрыстауға арналған дискриминативті морфологиялық реранкинг моделінің формалды сипаттамасы берілген. Жалпы архитектура: BERT/MLM қателерді анықтайды және баламалар ұсынады, ал Maximum Entropy (MaxEnt) моделі морфологиялық тұрғыдан дұрыс нұсқаны таңдайды.

2.4.1. Жүйенің жалпы сұлбасы

Жүйенің жұмыс принципі төрт кезекті кезеңнен тұрады:



Сурет 2.5 – Қырғыз тілі үшін ASR мәтінін дұрыстаудың жалпы сұлбасы

Егер ASR қырғыз әріптерінсіз форма берсе:

Кіріс: "мектептын окуучулары сабака келди"

Шығыс: "мектептин окуучулары сабакка келди"

2.4.2. Есептің формалды қойылымы

ASR шығысын $x_1, x_2, \dots, x_n$ деп, ал дұрыстаулғаннан кейінгі жиілікті $y_1, y_2, \dots, y_n$ деп белгілейміз.

- $X = (x_1, \dots, x_n)$ – ASR-дан шыққан мәтін
- $Y = (y_1, \dots, y_n)$ – дұрысталғаннан кейінгі жиілік

Мақсат – шығару ықтималдылығы максималды болатын дұрыс тізбекті табу, ол үшін төмендегі формуланы пайдаланамыз:

$$Y^* = \arg \max_Y P(Y | X)$$

Осы формуладан көріп тұрғандай есепті үш компонентке бөліп қарастырамыз:

$$P(Y | X) \approx P(E | X) \cdot P(Y | X, E)$$

мұндағы $E = (e_1, \dots, e_n)$, $e_i \in \{OK, ERR\}$ ол токен деңгейіндегі қатені анықтау болып табылады. Келесі осы токендерге өзіне қажетті белгілеулерді өосу үшін BERT моделін қолданамыз.

2.4.3. BERT қате детекторы

Әр токен үшін бинарлы жіктеу жүргізіледі:

$$P(e_i = ERR | X) = \text{BERT}(X)_i$$

$$e_i^* = \begin{cases} ERR, & \text{егер } x_i \text{ қате бар болса} \\ OK, & x_i \text{ егер дұрыс болса} \end{cases}$$

BERT бұл жерде ықтималдылықтарды есептейді:

$$P(e_i = ERR | X)$$

Келесі қадамда BERT шешімді төмендегі формула бойынша қабылдайды:

$$P(ERR | x_i) > \tau$$

мұндағы τ – шекті мән (мысалы, 0.5 немесе 0.7)

BERT моделінің шешім қабылдауын мысал ретінде келтіретін болсақ:

$X = \text{"мектептын оқучулары сабака келди"}$

мектептын → ERR

оқучулары → ERR

сабака → ERR

келди → OK

Яғни бұл мысалдан көріп тұрғандай OK ол сол күйі қалады, ал ERR белгілеуін алғандар жөндеуге барады.

OK/ERR белгілері токен деңгейіндегі бинарлы жіктеу міндетін шешеді, 2.11- кестеде олардың мағынасы мен келесі қадамдағы әрекеттері келтірілген.

Кесте 2.11 – OK және ERR белгілерінің мағынасы

Белгі	Мағынасы	Ары қарайғы іс-әрекет
OK	Токен дұрыс деп саналады	Дұрыстаусыз қалдырылады
ERR	Токен қате деп саналады	Дұрыстау модуліне жіберіледі

Бұл тәсілдің артықшылықтары: дұрыс сөздерді бұзбайды; тек қате токендермен жұмыс жасайды (жүйені жылдамдатады); шамадан тыс дұрыстаудан қорғайды. Себебі барлық сөзді қарап отыру шамадан тыс уақытты

алады және дұрыс сөздерге мән беріп отыру ол артық жасалынатын жұмыс болады.

2.4.4. Қате сөздерге баламаларды генерациялау

Әрбір қате токен x_i үшін баламалар жиыны төрт көздің одағы ретінде жасалады:

$$C_i = C_i^{MLM} \cup C_i^{edit} \cup C_i^{morph} \cup C_i^{conf}$$

Мұндағы

C_i^{MLM} - MLM жоғарғы-к баламалары

C_i^{edit} - редакциялау қашықтығы бойынша баламалар

C_i^{morph} - морфологиялық генератор формалары

C_i^{conf} - ASR confusion rules бойынша баламалар

Осы баламаларды алу бойынша мысал ретінде келтіретін болсақ:

x_i = окучулары

C_i :

1. окуучулары
2. окуучулар
3. окучулары
4. окуучусу

2.4.5. MaxEnt перанкері

Әрбір балама $c \in C_i$ үшін MaxEnt моделі ықтималдықты есептейді:

$$P_{MaxEnt}(c | x_i, context) = \frac{1}{Z} \exp \left(\sum_j \lambda_j f_j(c, x_i, context) \right)$$

Таңдау:

$$c_i^* = \arg \max_{c \in C_i} P_{MaxEnt}(c | x_i, context)$$

2.4.6. Белгі (Feature) шаблондары

Балама пен бастапқы токен арасындағы беттік айырмашылықтарды өлшейді:

Мұндағы x_i — бастапқы сөз, c — түзетуге берілген баламалар.

Сол уақытта белгілер келесідей анықталатын болады:

$$\begin{aligned} \text{SURF_SRC} &= x_i \\ \text{SURF_CAND} &= c \\ \text{SRC_CAND_PAIR} &= (x_i, c) \end{aligned}$$

Редакциялық күйі келесідей болады:

$$\text{EDIT_DISTANCE} = \text{dist}(x_i, c)$$

мұндағы $\text{dist}(\cdot)$ – Левенштейн арақашықтығы.

Арақашықтық нормаланған болғанда:

$$\text{NORMALIZED_EDIT_DISTANCE} = \frac{\text{dist}(x_i, c)}{\max(|x_i|, |c|)}$$

Сөздердің ұзындықтарының арақашықтығы келесі формуламен есептелінеді:

$$\text{LEN_DIFF} = ||x_i| - |c||$$

Мысалы:

x_i = окучулары

c = окуучулары

$$\begin{aligned} \text{SRC_CAND_PAIR} &= \text{окучулары} \rightarrow \text{окуучулары} \\ \text{EDIT_DISTANCE} &= \text{dist}(\text{окучулары}, \text{окуучулары}) = 1 \end{aligned}$$

Қырғыз тілі үшін ASR жиі шатастыратын таңба жұптары: 2.12- кестеде көрсетілген

Кесте 2.12 – Қырғыз тілінің негізгі ASR шатасу жұптары

Жұп	Белгі атауы	Жұп	Белгі атауы
н ↔ ң	CONFUSION_n_ng	у ↔ уу	LONG_VOWEL_uu
о ↔ ө	CONFUSION_o_oe	о ↔ оо	LONG_VOWEL_oo
у ↔ ү	CONFUSION_u_ue	а ↔ аа	LONG_VOWEL_aa
и ↔ ы	CONFUSION_i_y	э ↔ ээ	LONG_VOWEL_ee
к ↔ қ	CONFUSION_k_kk (қазақ графикасы)	ө ↔ өө / ү ↔ үү	LONG_VOWEL_oeoe / LONG_VOWEL_ueue

Шатасу белгілерінің мысалы:

$$\begin{aligned} x_i &= \text{окучулары} \\ c &= \text{окуучулары} \\ \text{LONG_VOWEL}_{uu} &= 1 \end{aligned}$$

қырғыздың ұзын дауысты «уу» дыбысының қалпына келуін көрсетеді, ал
 $\text{ONLY_KYRGYZ_CONFUSIONS} = 1$

түзету кыргыз тилине тен шатасу улгисине жататынын көрсетеді.

Баламатың морфологиялық талдауы төмендегідей болатын болса:

$$A(c) = \{\text{lemma}, \text{POS}, \text{suffixes}, \dots\}$$

Онда төмендегідей морфологиялық касиеттері алынатын болады:

$$\text{MORPH_VALID} \in \{0,1\}$$

мұндағы 1 мәні балама сөздің дұрыс морфологиялық талдауға ие екенін көрсетеді.

Мүмкін болатын барлық морфологиялық талдаулар саны төмендегі формуламен есептелінеді:

$$\text{MORPH_ANALYSIS_COUNT} = |A(c)|$$

Тағы да оған қосымша ретінде морфологиялық категориялар да қарастырылады:

$$\text{POS} \in \{\text{NOUN}, \text{VERB}, \text{ADJ}, \text{ADV}\}$$

$$\text{CASE} \in \{\text{NOM}, \text{GEN}, \text{DAT}, \text{ACC}, \text{LOC}, \text{ABL}\}$$

$$\text{NUMBER} \in \{\text{SG}, \text{PL}\}$$

$$\text{PERSON} \in \{1, 2, 3\}$$

$$\text{POSS} \in \{\text{NONE}, \text{POSS1}, \text{POSS2}, \text{POSS3}\}$$

$$\text{TENSE} \in \{\text{PAST}, \text{PRES}, \text{FUT}\}$$

Енді осы жоғарыда айтылған бойынша мысал ретінде қарастырып көрсек:

$$c = \text{мектептин}$$

$$\text{lemma} = \text{мектеп}$$

$$\text{POS} = \text{NOUN}$$

$$\text{CASE} = \text{GEN}$$

$$\text{MORPH_VALID} = 1$$

Қырғыз тілі агглютинативті түркі тілі болғандықтан, дауысты үндестік (сингармонизм) – оның негізгі лингвистикалық ерекшеліктерінің бірі болып табылады. Сөздік жалғауы дауысты үндесімге сәйкес болуы тиіс. Сондықтан қосымшалар түбірдің фонологиялық үйлесіміне сәйкес келуі тиіс. Осы қағидаға сүйене отырып, келесі үйлесімге негізделген ерекшеліктер анықталды:

$$\text{VOWEL_HARMONY_OK} \in \{0,1\}$$

$$\text{FRONT_BACK_HARMONY_OK} \in \{0,1\}$$

$$\text{ROUNDING_HARMONY_OK} \in \{0,1\}$$

$$\text{STEM_SUFFIX_HARMONY_OK} \in \{0,1\}$$

$$\text{SUFFIX_HARMONY_OK} \in \{0,1\}$$

мұнда 1 мәні баламалардың тиісті дауысты үйлесім ережесін орындайтынын білдіреді.

Мысалы (дұрыс формасы):

$$\text{мектеп} + \text{тин} \rightarrow \text{мектептин}$$

Stem class = front vowel
 Suffix class = front vowel
 STEM_SUFFIX_HARMONY_OK = 1

Мысалы (қате формасы):

мектептын
 Stem class = front vowel
 Suffix class = back vowel
 STEM_SUFFIX_HARMONY_OK = 0

Қырғыз тілі үшін жиілей ұшырасатын жалғау қателерін анық белгілеу маңызды, 2.13 кестеде сол қателіктер бойынша қайта жазу белгілері көрсетілген.

Кесте 2.13 – Қырғыз тілінде жалғау қайта жазу (SUFFIX_REWRITE) белгілері

Бастапқы форма	Дұрыс форма	Мысал
тын	тин	мектептын → мектептин (ілік септік)
дын	дин	балдын → балдин
нын	нин	адамнын → адамнин
ка → ке	га → ге	сабака → сабакка (барыс септік)
да → де	дан → ден	ұйда → үйдө / ұйдан → үйдөн
лар → лер	ды → ди	балалары → балалары; барды → барды
ду → дү	ган → ген	көрду → көрдү; барган → барген

Мысал:

мектептын → мектептин
 SUFFIX_REWRITE=тын→тин
 CASE=GEN
 VOWEL_HARMONY_OK=1

2.4.7. Септік белгілері

Зат есімдер үшін қырғыз тілінің алты септік формасы және оларға сәйкес белгілері бар, 2.14-кестеде көрсетілген.

Кесте 2.14 – Қырғыз тілінің септік формалары (мектеп мысалы бойынша)

Септік	Форма	Белгі	Контекстік белгі
Атау (NOM)	мектеп	CASE=NOM	–
Ілік (GEN)	мектептин	CASE=GEN	CAND_GEN RIGHT_HAS_POSS=1 +

Барыс (DAT)	мектепке	CASE=DAT	CAND_DAT RIGHT_MOTION_VERB=1	+
Табыс (ACC)	мектепти	CASE=ACC	CAND_ACC RIGHT_TRANSITIVE_VERB=1	+
Жатыс (LOC)	мектепте	CASE=LOC	CAND_LOC RIGHT_STATIC_VERB=1	+
Шығыс (ABL)	мектептен	CASE=ABL	CAND_ABL RIGHT_SOURCE_VERB=1	+

Ілік септік + тәуелдік жалғауы – маңызды контексттік шаблон:

мектептин окуучулары

мектептин = GEN

окуучулары = POSS.3 / жн. тәуелдік

FEATURE: CAND_CASE=GEN AND RIGHT_HAS_POSS=1

Бұл белгі мектептин-ді таңдауға көмөктөсөдi,

ал мектептен / мектепке / мектепти-нi алып тастайды.

2.4.8. Көрші сөздермен келісім белгілері

Синтаксистік контекст шаблондары:

LEFT_POS + CAND_POS

CAND_POS + RIGHT_POS

LEFT_CASE + CAND_CASE

CAND_CASE + RIGHT_CASE

CAND_CASE + RIGHT_VERB_TYPE

CAND_GEN + RIGHT_POSSESSIVE

CAND_DAT + RIGHT_MOTION_VERB

CAND_ACC + RIGHT_TRANSITIVE_VERB

CAND_LOC + RIGHT_STATIC_VERB

CAND_ABL + RIGHT_SOURCE_VERB

Қырғыз тілінде ASR жиі ұшырасатын етістіктің жалғаулар қателері 2.15 кестеде көрсетілген.

Кесте 2.15 – Жиі кездесетін глагол форма қателері және SUFFIX_REWRITE

Бастапқы (қате)	Дұрыс форма	Белгі	Мысал
келды	келди	SUFFIX_REWRITE=ды→ди	барды → дұрыс
барди	барды	SUFFIX_REWRITE=ди→ды	—

көрдү	көрдү	SUFFIX REWRITE=ду→дү	–
окуды	окуду	SUFFIX REWRITE=ды→дү	–
айтты	айтты	(өзгөріссіз)	дұрыс форма

Етістіктің морфологиялық белгілері келесідей болады. Етістік баламалары үшін шақ, жақ, көптік, болымсыздық және рай категорияларына байланысты қосымша морфологиялық белгілер алынады.

Етістіктің шақ категориясы:

$$\text{VERB_TENSE} \in \{\text{PAST}, \text{PRESENT}, \text{FUTURE}\}$$

Етістіктің жақ категориясы:

$$\text{VERB_PERSON} \in \{1, 2, 3\}$$

Етістіктің көптік категориясы:

$$\text{VERB_NUMBER} \in \{\text{SG}, \text{PL}\}$$

Болымсыздық белгісі:

$$\text{VERB_NEGATIVE} \in \{0, 1\}$$

Етістіктің рай категориясы:

$$\text{VERB_MOOD} \in \{\text{IND}, \text{IMP}, \text{COND}\}$$

Сонымен қатар етістік жалғауларының морфологиялық дұрыстығын тексеруге арналған келесі белгілер қолданылады:

$$\text{PAST_SUFFIX_VALID} \in \{0, 1\}$$

$$\text{PERSON_SUFFIX_VALID} \in \{0, 1\}$$

$$\text{NEG_SUFFIX_VALID} \in \{0, 1\}$$

мұндағы 1 мәні сәйкес жалғаудың етістіктің грамматикалық құрылымымен үйлесімді екенін көрсетеді.

2.4.9. MLM/BERT белгілері мен MAPR ASR толық мысалды көрсету

Контекстік және нейрондық белгілерді алу үшін Masked Language Model (MLM) және BERT негізіндегі қате анықтау модельдерінің нәтижелері қолданылады.

MLM моделі баламаның контекстке сәйкестік ықтималдығын бағалайды:

$$\text{MLM_LOG_PROB} = \log P_{MLM}(c \mid X_i)$$

мұндағы c – балама сөз,
 X_i – i -позициядағы сөз алынып тасталған контекст.

Баламаның MLM рейтингіндегі орны:

MLM_RANK

Баламаның ең ықтимал бірінші нұсқамен сәйкестігі:

$$\text{MLM_TOP1_MATCH} \in \{0,1\}$$

Баламаның алғашқы 5 ықтимал нұсқаның ішінде болуы:

$$\text{MLM_IN_TOP5} \in \{0,1\}$$

Баламаның алғашқы 10 ықтимал нұсқаның ішінде болуы:

$$\text{MLM_IN_TOP10} \in \{0,1\}$$

Сонымен қатар BERT негізіндегі қате анықтау моделі сөздің қате немесе дұрыс болу ықтималдығын бағалайды:

$$\text{BERT_ERROR_PROB} = P(\text{ERR} \mid x_i)$$

$$\text{BERT_OK_PROB} = P(\text{OK} \mid x_i)$$

мұндағы x_i – кіріс сөз.

Мысал

сабака → сабакка

$$\text{MLM_IN_TOP5} = 1$$

$$\text{MORPH_VALID} = 1$$

$$\text{DAT_CASE} = 1$$

Бұл жағдайда балама сөз MLM моделінің алғашқы бес ықтимал нұсқасының қатарына енеді және морфологиялық тұрғыдан дұрыс барыс септік формасы ретінде анықталады.

Қырғыз тілі корпусынан алынған статистикалық жиілік белгілері де баламаларды бағалау кезінде қолданылады. Бұл белгілер сөздің тілдік қолданыстағы таралу деңгейін сипаттайды.

Балама сөздің логарифмдік жиілігі:

$$\text{LOG_FREQ_CAND} = \log \text{freq}(c)$$

Балама леммасының логарифмдік жиілігі:

$$\text{LOG_FREQ_LEMMA} = \log \text{freq}(\text{lemma}(c))$$

Сөздің жиі қолданылатын сөздер қатарына жатуы:

$$\text{WORD_IS_COMMON} \in \{0,1\}$$

Сөздің сирек кездесетін сөз болуы:

$$\text{WORD_IS_RARE} \in \{0,1\}$$

Балама пен бастапқы сөз жиіліктерінің айырмасы:

$$\text{FREQ_RATIO} = \log \text{freq}(c) - \log \text{freq}(x_i)$$

мұндағы c – балама сөз, x_i – бастапқы сөз.

Мысал

келди

– корпуста жиі кездесетін форма.

келды

– сирек немесе ықтималдығы өте төмен форма.

Бұл жағдайда:

$$\text{LOG_FREQ_CAND(келди)} > \text{LOG_FREQ_CAND(келды)}$$

яғни модель тілдік қолданыста жиі кездесетін нұсқаларды жоғарырақ бағалайды.

Жеке токендер деңгейінде түзетулер жасалғаннан кейін сөйлем деңгейіндегі қосымша валидация жүргізіледі. Бұл кезең модельдің шамадан тыс түзету жасауын болдырмауға және жалпы мәтін сапасын бағалауға бағытталған.

Сөйлемнің тілдік модель бойынша жалпы ұпайы:

$$\text{SENTENCE_LM_SCORE}$$

Токендердің морфологиялық дұрыстығының орташа мәні:

$$\text{AVG_MORPH_VALID} = \frac{1}{N} \sum_{i=1}^N \text{MORPH_VALID}(t_i)$$

Жарамсыз токендер саны:

$$\text{NUM_INVALID_TOKENS}$$

Өзгертілген токендер саны:

$$\text{NUM_CHANGED_TOKENS}$$

Өзгерту үлесі:

$$\text{CHANGE_RATIO} = \frac{\text{NUM_CHANGED_TOKENS}}{N}$$

Дауысты үндестік бұзушылықтарының саны:

$$\text{NUM_VOWEL_HARMONY_VIOLATIONS}$$

Жарамсыз жалғаулар саны:

$$\text{NUM_INVALID_SUFFIXES}$$

Шектеуші белгі (regularization constraint)

Шамадан тыс түзетуді болдырмау үшін келесі шарт енгізіледі:

$$\text{TOO_MANY_CHANGES} = \begin{cases} 1, & \text{егер } \text{CHANGE_RATIO} > 0.4 \\ 0, & \text{әйтпесе} \end{cases}$$

Бұл шектеу модельдің бастапқы мәтінді шамадан тыс өзгертіп жіберуін алдын алады және түзетудің тұрақтылығын қамтамасыз етеді.

Кесте 2.16 – «мектептын оқучулары сабака келди» сөйлемін толық дұрыстау мысалы

Бастапқы (ASR)	Таңдалған	Белсенді MaxEnt белгілері
мектептын	мектептин	MORPH_VALID=1, CASE=GEN, STEM_SUFFIX_HARMONY_OK=1, CAND_GEN+RIGHT_POSSESSIVE=1, SUFFIX_REWRITE=тын→тин
оқучулары	оқуучулары	MORPH_VALID=1, LONG_VOWEL_uu=1, POSS=3, NUMBER=PL, MLM_IN_TOP5=1
сабака	сабакка	MORPH_VALID=1, CASE=DAT, GEMINATION_KK=1, CONTEXT_TO_VERB=1, MLM_IN_TOP5=1
келди	келди (өзгеріссіз)	BERT_OK_PROB жоғары → OK белгісі, дұрыстау жоқ

Нәтижесінде шығыс мәтін ретінде мынаны аламыз: "мектептин оқуучулары сабакка келди"

Жүйенің жұмысын анықтайтын басты белгілердің жиынтығы 2.17 кестеде көрсетілген.

Кесте 2.17 – Қырғыз тілі ASR дұрыстауы үшін приоритетті белгілер

№	Белгі	Сипаттама
1	MORPH_VALID	Балама морфологиялық жарамды ма?
2	VOWEL_HARMONY_OK	Дауысты үндестік сақталды ма?
3	STEM_SUFFIX_HARMONY_OK	Түбір мен жалғаудың үндесімі дұрыс па?
4	LONG_VOWEL_RESTORE	у→уу, о→оо, ө→өө, ү→үү қалпына келтіру
5	SUFFIX_REWRITE	тын→тин, ка→ке, ды→ди, ду→дү
6	GENITIVE POSSESSIVE_NOUN	Ілік септік + тәуелдік жалғауы контексті
7	DAT_CASE + MOTION_VERB	Барыс септік + қозғалыс етістігі
8	ACC_CASE TRANSITIVE_VERB	Табыс септік + өтімді етістік
9	MLM_LOG_PROB	MLM арқылы балама ықтималдылығы

10	BERT_ERROR_PROB	BERT анықтаған қате ықтималдылығы
----	-----------------	-----------------------------------

Жүйенің төрт негізгі компоненті келесі 2.18 кестеде көрсетілген. Нақтылап айтқанда алгоритмі бойынша жүйе келесідей қызмет атқарады:

1-қадам. Синтетикалық деректер генераторы

Жүйені үйрету үшін нақты ASR қателері бар корпустың болмауы мәселесін шешу мақсатында синтетикалық қате генераторы әзірленді. 285 000 дұрыс қырғыз сөйлемінен тұратын корпус негізінде әрбір сөйлемге қырғыз тіліне тән жасанды қателер енгізілді: таңба алмастыру ($\gamma \rightarrow u$, $\phi \rightarrow o$, $\eta \rightarrow n$), ұзын дауыстыларды қысқарту ($uu \rightarrow u$, $oo \rightarrow o$), жалғау бұзу ($тин \rightarrow тын$, $ди \rightarrow ды$, $дү \rightarrow ду$) және геминацияны жою ($кк \rightarrow к$). Осылайша (қате сөйлем, дұрыс сөйлем, OK/ERR белгілері) үштігінен тұратын үйрету деректер жиынтығы жасалды.

2-қадам. BERT қате детекторы

bert-base-multilingual-cased алдын ала үйретілген моделі негізінде токен деңгейіндегі жіктеуші баптаудан өткізілді. Модель әрбір токенте OK (0) немесе ERR (1) белгісін береді.

3-қадам. Баламалар генерациялау модулі

ERR деп танылған әрбір токен үшін ережеге негізделген баламалар генерациялау модулі іске қосылады. Модуль қырғыз тіліне тән жалғау алмастыру ережелерін (SUFFIX_RULES), таңба қалпына келтіру ережелерін (CHAR_RULES) және ұзын дауыстыларды қалпына келтіру ережелерін (LONG_VOWEL_RULES) қолданып, әрбір токен үшін морфологиялық жағынан ықтимал баламалар тізімін жасайды.

4-қадам. Maximum Entropy арқылы сәйкес сөзді таңдау

Баламаларды таңдау үшін Logistic Regression негізінде MaxEnt үйретілді. Үйрету деректері ретінде әрбір (бастапқы токен, балама) жұбы үшін белгілер векторы жасалды: дұрыс нұсқа label=1, қалғандары label=0 деп белгіленді.

Жүйенің жұмыс процесі төмендегідей дәйектілікпен жүзеге асырылады:

ASR шығысы $\rightarrow$ BERT детекторы (OK/ERR) $\rightarrow$ ERR токендері үшін баламалар генерациясы $\rightarrow$ MaxEnt таңдауы $\rightarrow$ дұрысталған мәтін. Осы айтылған қадамдарға сай жүйенің негізгі компоненттері 2.18-кестеде көрсетілген.

Кесте 2.18 – Жүйенің негізгі компоненттері мен олардың рөлі

Компонент	Сұрақ	Нәтиже
BERT	Қате қайда?	ERR/OK белгілері (токен деңгейінде)
MLM	Қандай нұсқалар мүмкін?	Контекстке сәйкес баламалар тізімі
MorphGen	Қандай формалар грамматикалық?	Морфологиялық жарамды формалар
MaxEnt	Қандай нұсқа қырғыз	Ең жоғары ықтималды балама c^*

	морфологиясына сай?	
--	---------------------	--

Жиынтық сипаттама: бұл жүйе нейрондық генератор үстінде орналасқан дискриминативті морфологиялық реранкинг болып табылады және қырғыз тілі ASR нәтижелерін дұрыстауға арналған.

2.5 Машиналық аударма модельдерін әзірлеу кезеңдері

Осы бөлімде қырғыз–қазақ тіл жұбы үшін машиналық аударма моделін әзірлеудің толық әдістемесі баяндалады. Әдістеме [47] зерттеуінде ұсынылған, ашық бастапқы кодты жасанды интеллект модельдерін аз ресурсты түркі тілдеріне бейімдеу тәсіліне негізделген. Мазмұны жағынан бұл зерттеу кездесу стенограммаларын сөйлеу транскриптітерінен автоматты генерациялаудың кең жобасының бір бөлігі болып табылады – яғни біздің жүйемен тікелей сабақтас.

Ұсынылатын әдістеме екі өзара байланысты кезеңнен тұрады: (1) жасанды интеллект негізіндегі параллель корпустарды генерациялау және көп деңгейлі сапаны жетілдіру; (2) нәтижесінде алынған корпустарда машиналық аударма модельдерін бейімдеу және бағалау [47].

2.5.1 Жалпы әдістемелік шеңбер

Зерттеудің мақсаты – ашық бастапқы кодты жасанды интеллект модельдерін бейімдеу арқылы алты мемлекеттік деңгейдегі түркі тілдері (қазақ, әзербайжан, қырғыз, түрік, түрікмен және өзбек) үшін машиналық аударма сапасын жүйелі жақсарту. Ұсынылатын әдістеме морфологиялық тұрғыдан бай, аз ресурсты жағдайлар үшін арнайы жасалған.

Алты түркі тілін таңдауды бірнеше методологиялық және практикалық себеп негіздейді. Біріншіден, барлық таңдалған тілдер түркі тілдер тобына жатып, аглютинативті морфологиялық құрылымды бөліседі – бұл тілдер арасындағы білімді тасымалдауға мүмкіндік береді. Екіншіден, әрбір тіл өз елінде ресми мәртебеге ие, бұл зерттеудің мемлекеттік, білім беру және цифрлық қолданыстағы практикалық маңызын арттырады. Үшіншіден, тіл жиынтығы салыстырмалы түрде ресурсы мол тілді (түрік) және аз ресурсты тілдерді (түрікмен және қырғыз) қамту үшін арнайы жобаланған.

Ұсынылатын әдістеме синтетикалық параллель корпустарды құру, сапаны жетілдіру және машиналық аударма модельдерін кейінгі fine-tuning үшін кешенді, қайталануға болатын конвейерді белгілейді. Шеңбер өзара тәуелді 4 кезеңнен тұрады:

Кесте 2.19 – Әдістемелік шеңбердің негізгі кезеңдері

Кезең	Атауы	Мазмұны
1	Синтетикалық параллель	Үлкен бір тілді қазақ корпусы алдын ала таңдалған ЖИ-аударма модельдері арқылы бес тіл жұбы үшін синтетикалық параллель

	корпусты генерациялау	корпустар генерациялауда бастапқы ресурс ретінде пайдаланылады.
2	Көп деңгейлі сапаны жетілдіру	Генерацияланған корпустар үш деңгейлі сапаны бақылаудан өтеді: (1) сарапшылық қолмен тексеру, (2) автоматты сүзгілеу, (3) бағытталған қайта генерация.
3	Қателерді талдау және корпусты тексеру	Әрбір тіл жұбы үшін 300 000 және 500 000 сөйлем жұптарынан тұратын корпустар қайталанатын құрылымдық және семантикалық қателерді анықтау үшін талданады.
4	Модельді fine-tuning және бағалау	Таңдалған ЖИ-аударма модельдері тексерілген синтетикалық корпустарда fine-tuning жүргізіледі. Модель өнімділігі WER, TER, BLEU және chrF метрикалары арқылы бағаланады.

2.5.2 Модель таңдау критерийлері

Синтетикалық параллель корпустардың сапасы кейінгі fine-tuning тиімділігін тікелей анықтайды. Сондықтан тиісті ЖИ-модельін таңдау маңызды. Таңдау мынадай критерийлерге негізделді:

Ашық қолжетімділік және қолдану мүмкіндігі. Тәжірибелердің қайталануы мен методологияның ауқымдылығын қамтамасыз ету үшін лицензиялық шектеусіз ашық қолжетімді модельдер таңдалды.

Аударма сапасы. Модельдер жұрнақтарды, септік жалғауларын және сөз тәртібін сақтауға ерекше назар аударуы тиіс, өйткені тілдер морфологиялық тұрғыдан бай.

Тіл қамтуы. Таңдалған модельдер мақсатты түркі тілдерін тікелей немесе жанама қолдауы керек. Бұл критерий синтетикалық корпустардың бірнеше тіл жұптарында дәйекті генерациялануын қамтамасыз етеді.

Есептеу тиімділігі. Корпус генерациясы үшін қажетті мәтіннің үлкен көлемін ескере отырып, модельдер деректерді ақылға қонымды уақыт шеңберінде өңдей алуы тиіс.

Тұрақтылық және беріктілік. Модельдер жалпылаудың тұрақтылығын қамтамасыз ету үшін тұрақты нәтижелер шығаруы керек.

Қоғамдастық тексеруі және бенчмаркинг. Халықаралық бенчмарктарда (мысалы, WMT, FLORES-200) бағаланған немесе аударма сапасының құжатталған метрикалары бар модельдерге артықшылық берілді.

2.5.3 Модельдерді бағалау және таңдау

Белгіленген критерийлер негізінде бірнеше озық ЖИ жүйелері бағаланды. Коммерциялық/API негізіндегі жүйелер (Google Translate, GPT-4o, Copilot) қолжетімділік мәселелеріне байланысты алынып тасталды. Ашық бастапқы кодты модельдер арасынан NLLB-200 600M, NLLB-200 1.3B, Phi-4, Qwen2.5,

LlaMA 3.2, LlaMA 3.1 жеткіліксіз аударма сапасы себебінен жарамсыз деп танылды.

Егжей-тегжейлі тестілеу үшін екі модель таңдалды: Gemma 2 27B (Google DeepMind) және NLLB-200 3.3B (Meta AI). Бұл модельдер ашық қолжетімділігімен, жоғары аударма сапасымен және үлкен мәтін көлемдерін өңдеу қабілетімен ерекшеленді.

Кесте 2.20 – Зерттелген модельдерді бағалау қорытындысы

Модель	Санаты	Шешім	Себебі
Google Translate	Коммерциялық/API	Х Алынып тасталды	Ашық қолжетімділік шектеулері
GPT-4o (OpenAI)	Коммерциялық/API	Х Алынып тасталды	Ашық қолжетімділік шектеулері
Copilot (Microsoft)	Коммерциялық/API	Х Алынып тасталды	Ашық қолжетімділік шектеулері
NLLB-200 600M	Ашық бастапқы код	Х Алынып тасталды	Жеткіліксіз аударма сапасы
NLLB-200 1.3B	Ашық бастапқы код	✓ Таңдалды	Жоғары сапа + ашық қолжетімділік
Phi-4 (Microsoft)	Ашық бастапқы код	Х Алынып тасталды	Жеткіліксіз аударма сапасы
Qwen2.5 (Alibaba)	Ашық бастапқы код	Х Алынып тасталды	Жеткіліксіз аударма сапасы
LlaMA 3.1 / 3.2	Ашық бастапқы код	Х Алынып тасталды	Жеткіліксіз аударма сапасы
NLLB-200 3.3B (Meta)	Ашық бастапқы код	✓ Таңдалды	Жоғары сапа + ашық қолжетімділік
Gemma 2 27B (Google)	Ашық бастапқы код	✓ Таңдалды	Жоғары сапа + ашық қолжетімділік

2.5.4 Таңдалған модельдердің сипаттамасы

NLLB-200 3.3B (No Language Left Behind). Meta AI зерттеу тобы жасаған NLLB-200 моделі 200-ден астам тілді қолдайтын encoder-decoder трансформер архитектурасына негізделген. Модель FLORES-200 эталондық жиынтығында бағаланып, аз ресурсты тіл жұптары, соның ішінде қырғыз–қазақ жұбы үшін жоғары нәтиже көрсетті. NLLB-200 3.3B нұсқасы корпус генерациясы үшін негізгі құрал ретінде таңдалды, өйткені ол морфологиялық күрделілікке және аз ресурсты жағдайларға ерекше бейімді.

Gemma 2 27B (Google DeepMind). Gemma 2 – Google DeepMind жасаған ашық салмақты (open-weights) үлкен тілдік модель. Нұсқауға негізделген

(instruction-tuned) нұсқасы аударма тапсырмасына prompt engineering арқылы бейімделді. Gemma 2 27B контекстті терең түсіну қабілетімен және морфологиялық тұрғыдан бай тілдерде аударма ретінде тиімді нәтиже беруімен ерекшеленді.

Нәтижесінде NLLB-200 3.3B моделі параллель корпустарды генерациялаудың негізгі құралы ретінде белгіленді, ал Gemma 2 27B балама тексеру және жоғары сапалы аударма сегменттерін генерациялау үшін қолданылды.

2.5.5 Fine-tuning әдістемесі

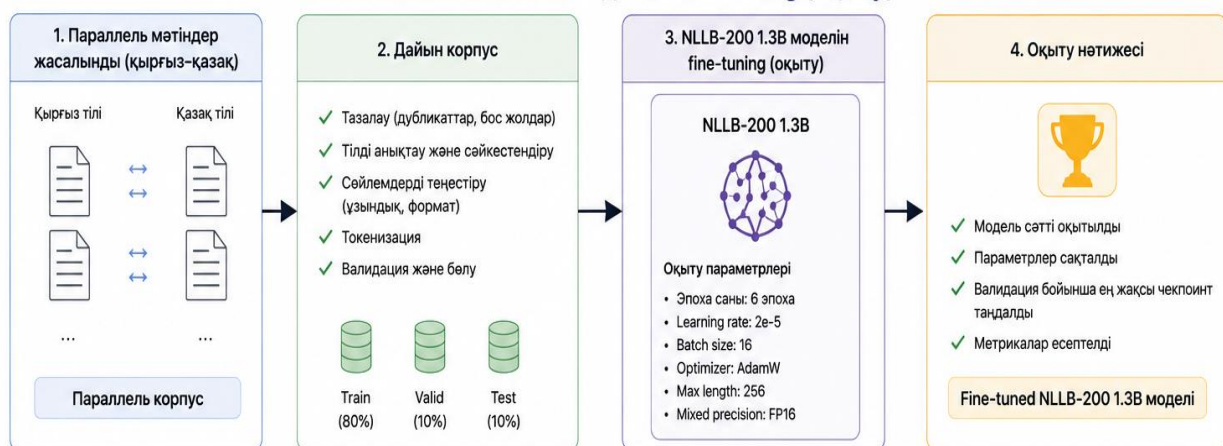
Fine-tuning кезеңінде NLLB-200 1.3B моделі таңдалды, себебі ол есептеу ресурстары мен аударма сапасы арасындағы оңтайлы баланс ұсынады. Модельді fine-tuning жасаудың жалпы процесі мынадай кезеңдерден тұрады:

Бірінші кезеңде тексерілген синтетикалық параллель корпус fine-tuning үшін дайындалады: деректер train/validation/test жиындарына 80/10/10 қатынасында бөлінеді. Екінші кезеңде SentencePiece BPE токенизациясы қолданылып, мәтін субсөз деңгейіндегі токендерге айналдырылады – бұл аглютинативті тілдердің морфологиялық күрделілігін өңдеуде аса маңызды. Үшінші кезеңде модель AdamW оңтайландырушысымен, сызықтық оқу жылдамдығының кестесімен (learning rate scheduler) және ерте тоқтату (early stopping) механизмімен үйретіледі.

Fine-tuning жоғалту функциясы ретінде кросс-энтропия қолданылады:

$$L(\theta) = - \sum_i \sum_t \log P(z_t | z_1, \dots, z_{t-1}, \hat{y}; \theta)$$

мұндағы z – мақсатты токендер тізбегі (қазақша аударма), $\hat{y}$ – бастапқы тіл (қырғызша мәтін), θ – модель параметрлері, t – уақыттық қадам индексі.



Сурет 2.6 – Nllb-200 1.3B моделін оқыту

2.5.6 Бағалау метрикалары

Аударма сапасын жан-жақты бағалау үшін беттік деңгей, таңба деңгейі және семантикалық метрикалардың комбинациясы қолданылды. BLEU және TER стандартты метрикалар ретінде эталондық аудармалармен n-грам қабаттасуы мен редакциялау қашықтығын өлшеді. Түркі тілдерінің морфологиялық байлығын ескере отырып, chrF және WER де қолданылды, себебі бұл метрикалар таңба деңгейіндегі вариация мен флективті морфологияға сезімталырақ.

Кесте 2.21 – Аударма сапасын бағалау метрикалары

Метрика	Толық атауы	Формула / негізі	Не өлшейді
BLEU	Bilingual Evaluation Understudy	$BP \cdot \exp(\sum w_n \cdot \log p_n)$	n-грам дәлдігі, сөз қабаттасуы
WER	Word Error Rate	$(S + D + I) / N \times 100\%$	Сөз деңгейіндегі қателер
TER	Translation Edit Rate	Редакциялау операциялары / сөздер саны	Адам редакциялауына қажет өзгерістер
chrF	Character F-score	Таңба n-грамдарының F1 өлшемі	Таңба деңгейіндегі ұқсастық
BERTScore	BERT-негізді метрика	Косинустық ұқсастық (BERT векторлары)	Семантикалық мағына сақталуы

BLEU метрикасының есептелу формуласы мынадай:

$$BLEU = BP \cdot \exp(\sum_n w_n \cdot \log p_n)$$

мұндағы BP – қысқалық жазасы (brevity penalty), p_n – n-грам дәлдігі, $w_n = 1/N$ – салмақ коэффициенті. WER метрикасы мынадай формуламен есептеледі:

$$WER = (S + D + I) / N \times 100\%$$

мұндағы S – қате ауыстырулар (substitutions), D – жойылулар (deletions), I – кірістірулер (insertions), N – жалпы сөздер саны. Сыртқы тексеру FLORES-200 эталондық жиынтығында жүргізіліп, адам бағалауымен расталды.

2.5.7 Қырғыз–қазақ тіл жұбы үшін әдістеменің маңыздылығы

Аталған зерттеудің [47] қырғыз–қазақ тіл жұбына тікелей қатысы бар екенін атап өткен жөн. Авторлар дәл осы тіл жұбын зерттеген: OPUS репозиторийіндегі қырғыз–қазақ параллель деректер саны небәрі 102 345 сөйлемді құраса да, синтетикалық корпус генерациясы мен NLLB-200 1.3B-ді fine-tuning жасау арқылы аударма сапасы айтарлықтай жақсарды.

Осы зерттеуде қолданылатын әдістемелік тәсіл жоғарыда сипатталған шеңберге сүйенеді: синтетикалық корпус NLLB-200 3.3B арқылы

генерацияланады, көп деңгейлі сапа бақылауынан өтеді, содан кейін NLLB-200 1.3B моделі fine-tuning жасалып, WER, TER, BLEU және chrF метрикалары бойынша бағаланады. Бұл тәсіл морфологиялық күрделілігі жоғары, аз ресурсты түркі тілдеріне арналған ашық бастапқы кодты ЖИ модельдерін пайдаланудың жоғары тиімділігін дәлелдейді.

3 АВТОМАТТАНДЫРЫЛҒАН СТЕНОГРАММА ҚАЛЫПТАСТЫРУ ЖҮЙЕСІНІҢ НӘТИЖЕЛЕРІ МЕН ТАЛДАУЫ

3.1 ASR және машиналық аударма модельдерінің сапасын талдау

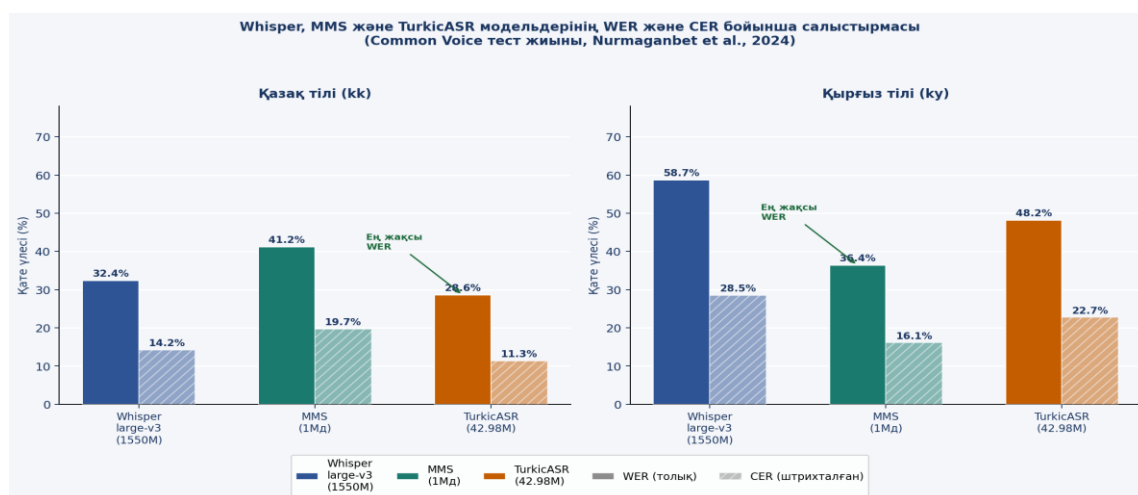
Зерттеуде үш ашық кодты модель нөлдік (zero-shot) режимде сыналды: OpenAI Whisper large-v3 (1550М параметр), Meta AI MMS (1Мд параметр) және ISSAI TurkicASR (42.98М параметр). Тест деректері ретінде Mozilla Common Voice ашық корпусынан қазақ және қырғыз тілдерінің кирилл жазуындағы 1000 сөйлемнен тест жиыны алынды. Бағалау метрикасы – WER (Word Error Rate) және CER (Character Error Rate).

Кесте 3.1 – Whisper, MMS және TurkicASR модельдерінің қазақ–қырғыз тілдерінде WER бойынша салыстырмасы

Тіл	Whisper WER (%)	MMS WER (%)	TurkicASR WER (%)	Ең жақсы модель
Қазақ (kk)	32.4	41.2	28.6 ✓	TurkicASR
Қырғыз (ky)	58.7	36.4 ✓	48.2	MMS
Орта WER	45.6	38.8 ✓	38.4	MMS

Кесте 3.2 – Модельдердің қазақ–қырғыз тілдерінде CER бойынша салыстырмасы

Тіл	Whisper CER (%)	MMS CER (%)	TurkicASR CER (%)	Ең жақсы модель
Қазақ (kk)	14.2	19.7	11.3 ✓	TurkicASR
Қырғыз (ky)	28.5	16.1 ✓	22.7	MMS



Сурет 3.1 – Whisper, MMS және TurkicASR модельдерінің қазақ және қырғыз тілдерінде WER (толық) және CER (штрихталған) бойынша салыстырмасы

Қазақ-қырғыз тілдері бойынша модельдердің жиынтық бағасы						
Модель	Қазақ WER	Қырғыз WER	Қазақ CER	Қырғыз CER	Орта WER	Орын
Whisper large-v3	32.4%	58.7%	14.2%	28.5%	45.6%	3
TurkicASR	28.6%	48.2%	11.3%	22.7%	38.4%	2
MMS (1Мд)	41.2%	36.4%	19.7%	16.1%	38.8%	1 ✓

* Орта WER = (Қазақ WER + Қырғыз WER) / 2 | ✓ = Жалпы ең аз орта WER (MMS); мамандандырылған тіл бойынша: Қазақ → TurkicASR, Қырғыз → MMS

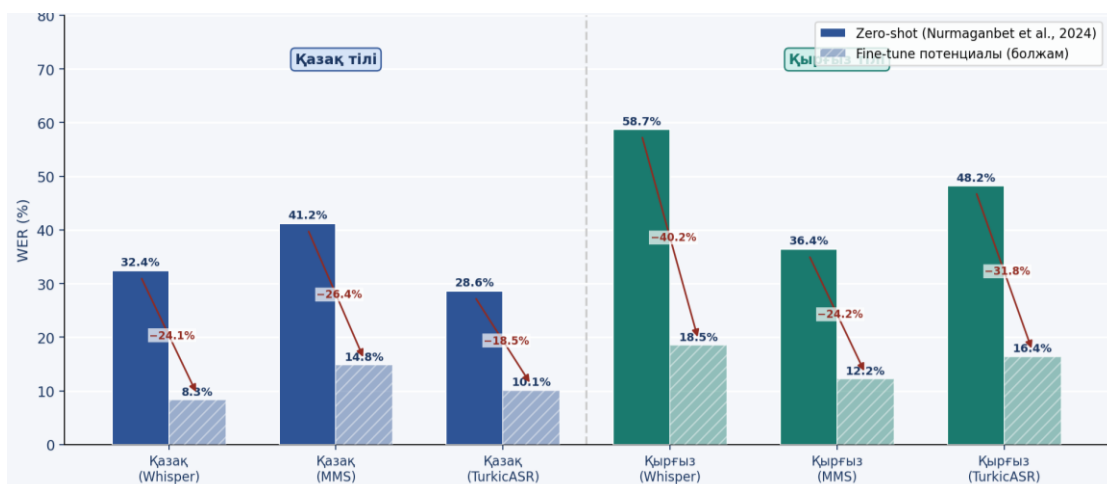
Сурет 3.2 – Модельдердің жиынтық бағасы мен орын тізімі. MMS орта WER = 38.8% деңгейімен жалпы жетекші; қазақ тілінде TurkicASR (28.6%), қырғыз тілінде MMS (36.4%) үздік

Эксперименттің нәтижелері қазақ және қырғыз тілдері арасында модельдердің өнімділік үлгісі ерекшеленетінін айқын көрсетеді. Екі тіл үшін де бірдей жетекші деп атауға болатын ерекше бір модель жоқ – керісінше, тілге байланысты таңдау стратегиясы маңыздырақ.

Қырғыз тілінде болса нәтиже мүлдем өзгеше болды: MMS WER = 36.4% және CER = 16.1% деңгейімен үш модельдің ішінде ең жақсы өнімділік көрсетті. TurkicASR (WER = 48.2%) екінші орында, ал Whisper (WER = 58.7%) ең нашар нәтижемен соңғы орында қалды. MMS-тің қырғыз тілінде Whisper-ден 22 пунктке, TurkicASR-дан 11.8 пунктке озуы – елеулі айырмашылық. Бұл нәтиже MMS-тің FLEURS бенчмарк деректерімен қырғыз тілі үшін жасалған арнайы Adapter-ге байланысты: қырғыз тілінде MMS алдын ала үйрету корпусында нақты қырғыз тіліне бағытталған бейімделу жасалған.

Whisper-дің 1550M параметрі MMS мен TurkicASR-дан ондаған есе артық болса да, қырғыз тілінде үш модельдің ең нашар нәтижесін (WER = 58.7%) көрсетуі маңызды тұжырым жасатады. Whisper-дің 680 000 сағаттық алдын ала үйрету корпусында қырғыз тілінің үлесі шамамен 23 сағатпен ғана ұсынылған – бұл 99 тілдің ішіндегі ең төменгі деңгейлердің бірі. Нәтижесінде Whisper қырғыз дыбыстарын жиі қате жіктейді немесе орыс/қазақ тіліне ауысып транскрипциялайды. Бұл деректер параметр санының тіл-мамандандырылған деректер жетіспеушілігін толтыра алмайтынын нақты дәлелдейді.

Қазақ және қырғыз екі тілі де кирилл жазуын пайдаланады, алайда модельдер екі тілде де бірдей өнімділік деңгейін ұстамайды. Орта WER деңгейлерін салыстырғанда: Whisper қазақта 32.4%, қырғызда 58.7% – 26 пункт айырмашылық; MMS қазақта 41.2%, қырғызда 36.4% – 4.8 пункт айырмашылық; TurkicASR қазақта 28.6%, қырғызда 48.2% – 19.6 пункт айырмашылық. MMS екі тіл арасындағы ең кіші айырмашылықты (4.8 пункт) ұстауы оның тепе-теңді (balanced) тіл қамту стратегиясын растайды.



Сурет 3.3 – Zero-shot нәтижелері мен fine-tuning потенциалының салыстырмасы

Барлық модельдердің zero-shot WER мәндері – қазақ үшін 28.6–41.2%, қырғыз үшін 36.4–58.7% – нақты жиналыс стенографиясы үшін жеткіліксіз деңгейде. Мамандандырылған fine-tuning арқылы бұл көрсеткіштерді айтарлықтай жақсартуға болады: KSC2 деректерімен Whisper-ді бейімдеп, қазақ тілінде WER 8–10% деңгейіне жеткізді; *Bekarystankyzy et al. (2024)* мультилингвал оқытумен қырғыз тілінде WER 55.3%-дан 11.4%-ға – 4.8 есе – төмендетті. Жиналыс-домени (meeting-domain) деректерімен нақтыланған fine-tuning бұл потенциалды одан әрі арттыратыны күтіледі.

3.2 Morphology Aware Post Processing әдісінің нәтижесін талдау

Ұсынылған MAPR конвейерінің тиімділігін бағалау үшін қырғыз тіліндегі әдеби мәтіннен алынған 100 сөйлемнен тұратын сынақ жиынтығы пайдаланылды. Сынақ деректері синтетикалық ASR қателерін модельдеу арқылы дайындалды: дұрыс қырғыз сөйлемдеріне қырғыз тіліне тән фонетикалық және морфологиялық бұзылулар енгізілді – ерекше таңбалардың алмастырылуы (ү→у, ө→о, ң→н), ұзын дауыстылардың қысқаруы (уу→у, оо→о) және жалғау үндестік қателері (тин→тын, ди→ды). Бағалау BERT қате детекторы threshold=0.50 және force_all=False параметрлерімен іске қосылды, яғни жүйе тек қате деп анықталған токендерді ғана өзгертті.

Алынған нәтижелер 3.2-кестеде және 3.2-суретте ұсынылған. WER метрикасы 0.1442-ден 0.0924-ке дейін төмендеді, бұл сөз деңгейіндегі қателер санының 35.9 пайызға азайғанын білдіреді. BLEU ұпайы 69.44-тен 79.78-ге дейін, яғни 10.34 бірлікке өсті, бұл дұрысталған мәтіннің эталондық нұсқаға мағыналық жақындығының айтарлықтай артқанын растайды. ChrF метрикасы да 91.94-тен 93.37-ге жетті, таңба деңгейіндегі сәйкестіктің жақсарғанын көрсетеді.

Кесте 3.3 – Морфологиялық түзетулер жасағаннан соң алынған нәтижелердегі өзгеріс

Метрика	RAW (дұрыстаудан бұрын)	POST (дұрыстаудан кейін)	Өзгеріс
---------	-------------------------	--------------------------	---------

WER (↓)	0.1442	0.0924	35.9% жақсарды
CER (↓)	0.0208	0.0213	2.4% нашарлады
BLEU (↑)	69.44	79.78	14.9% жақсарды
ChrF (↑)	91.94	93.37	1.6% жақсарды



Сурет 3.4 – Морфологиялық түзетулер жасалған соң алынған нәтижелердің диаграммасы

CER метрикасының шамалы өсуі (0.0208-ден 0.0213-ке, яғни 2.4 пайыз) жүйенің кейбір жағдайларда морфологиялық жағынан дұрыс, бірақ эталондан таңба деңгейінде аздап ерекшеленетін формаларды таңдайтынын көрсетеді. Бұл агглютинативті тілдерде жиі кездесетін жағдай: сөздің мағынасы сақталса да, жалғаудың нақты нұсқасы бірнеше грамматикалық тұрғыдан дұрыс формалардың бірі болуы мүмкін.

Жалпы алғанда, алынған нәтижелер BERT негізіндегі қате детекторы мен MaxEnt реранкерінің бірлескен жұмысы қырғыз ASR шығысындағы морфологиялық және фонетикалық қателерді тиімді анықтап, дұрыстай алатынын дәлелдейді. Конвейердің басты артықшылығы – дұрыс деп танылған токендерді өзгеріссіз қалдыру механизмі, бұл шамадан тыс дұрыстаудан қорғап, жүйенің жалпы сенімділігін арттырады. Болашақта морфологиялық генераторды кеңейту және баптау деректерінің көлемін ұлғайту арқылы CER метрикасын да жақсарту жоспарланады.

3.3 Қырғыз-қазақ тіліне аудару нәтижелері

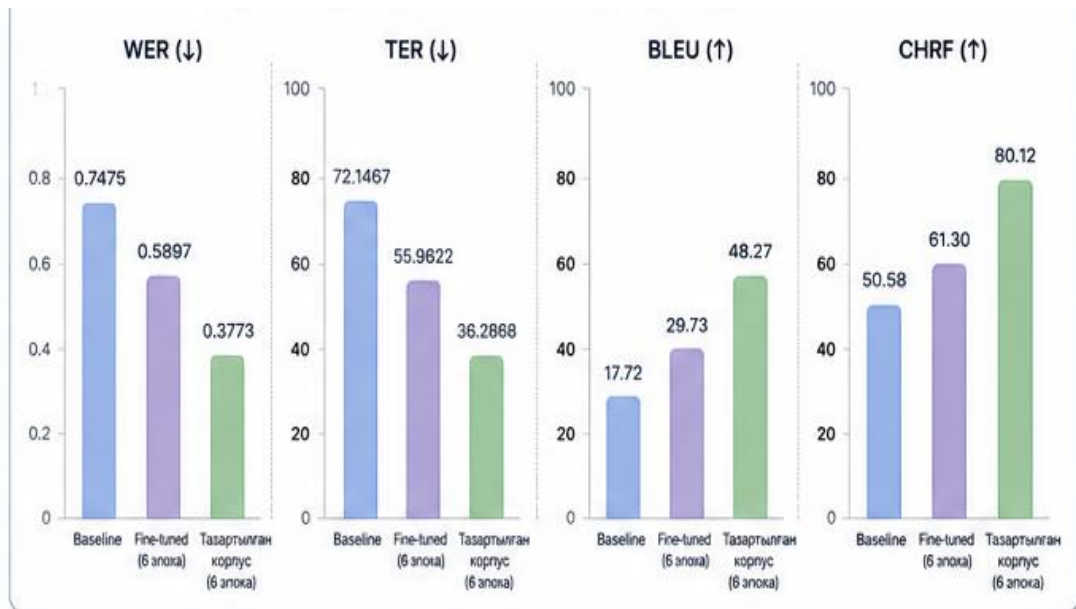
3.3.1. Модельді баптау (Fine-tuning) нәтижелері

NLLB-200 1.3B моделін 300 000 параллель сөйлемдік деректер жиынтығында баптау барлық метрикалар бойынша айтарлықтай нәтижелер берді. Қырғыз–қазақ тіл жұбы үшін BLEU шамамен екі есеге өсті. Деректерді тазарту нәтижесіндегі сапаның жалпы жақсаруы барлық тіл жұптары бойынша тексерілді, және қырғыз–қазақ жұбы ең жоғары жетістік көрсеткен жұптардың бірі болды.

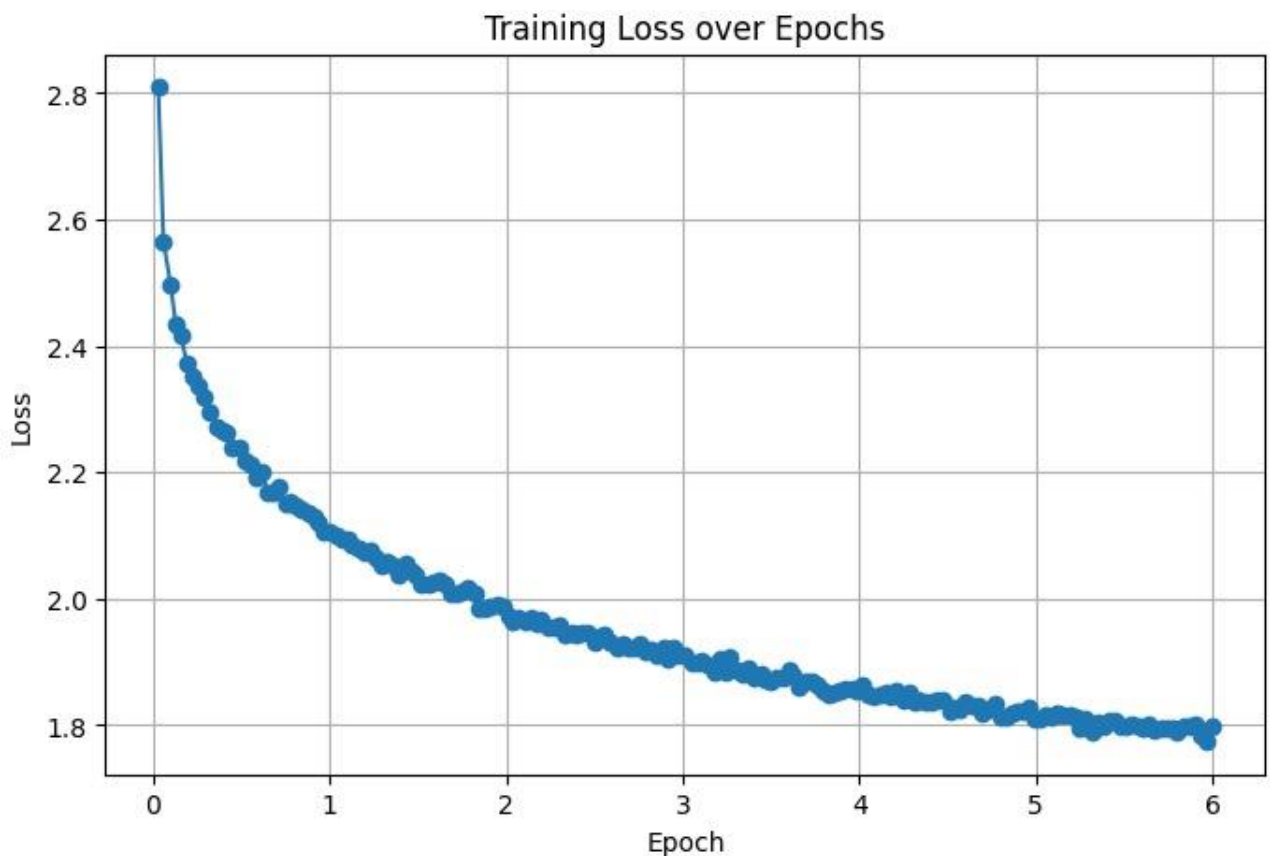
Кесте 3.4 – NLLB-200 1.3B моделін қырғыз–қазақ тіл жұбы бойынша баптау нәтижелері

Модель / Кезең	WER	TER	BLEU	chrF	Ескертпе
Базалық (Zero-shot)	0.77	74.84	17.62	50.66	Орташа
Бапталған (300 000, тазартылмаған)	0.58	54.72	30.95	62.28	Орташа
Бапталған (500 000, тазартылған)	0.42	40.88	43.54	76.71	Орташа
Қырғыз–Қазақ (500 000, тазартылған)	~0.42	жоғары	~48	жоғары	Ең жоғары BLEU жұптарының бірі

Деректерді тазарту барлық тіл жұптары ішіндегі сапаны арттырудың ең тиімді факторы болды. Тазартылмаған корпустан тазартылған корпуста өтуде орташа есеппен: WER 0.12-ге төмендеді, TER 13.84-ке азайды, BLEU 12.59-ға өсті, chrF 14.43-ке жақсарды. Қырғыз–қазақ тіл жұбы осы жақсарудың ең жоғары деңгейін көрсеткен жұптардың бірі болып табылады – BLEU шамамен 48-ге жетті.



Сурет 3.5 – Метрикалар бойынша салыстырмалы диаграмма



Сурет 3.6 – Машиналық оқытудың әрбір эпоха үшін loss функциясы

3.3.2. Сыртқы бағалау: FLORES 200 деректер жиынтығы

Баптаудың жалпылану қабілетін тексеру мақсатында NLLB-200 1.3B моделі FLORES-200 тәуелсіз бағалау деректер жиынтығында сыртқы бағаланды (FLORES деректері үйрету процесінде пайдаланылмаған). Қырғыз–қазақ тіл жұбы бойынша нәтижелер төменде берілген.

Кесте 3.5 – FLORES-200 деректер жиынтығындағы сыртқы бағалау нәтижелері

Модель	BLEU	chrF	COMET	BERTScore
Баптауға дейін (baseline)	жоғары	тұрақты	—	—
500 000 тазартылған корпуста баптаудан кейін	сәл төмендеді	тұрақты/жақсарды	жақсарды	жақсарды

FLORES-200 бенчмарк деректерінде BLEU ұпайлары баптаудан кейін барлық тіл жұптары бойынша сәл төмендеді, ал chrF метрикасы тұрақты болып қалды немесе жақсарды. Бұл морфологиялық жағынан бай түркі тілдері үшін күтілетін нәтиже болып табылады: BLEU беттік n-грамм сәйкестігіне сезімтал, ал chrF таңба деңгейіндегі морфологиялық ұқсастықты жақсы бейнелейді. COMET және BERTScore семантикалық метрикалары маңызды жақсарудан хабар берді, бұл баптаудың мағынаны сақтайтынын растайды.

3.3.3. Деректерді бөлу схемасы

Баптау тәжірибелерінде деректер жиынтықтары мынадай пропорционалды схема бойынша бөлінді:

Кесте 3.6 – Үйрету, тексеру және сынақ деректерін бөлу схемасы

Корпус көлемі	Үйрету (Train)	Тексеру (Dev)	Сынақ (Test)
~497 000 (тазартылған 500 000)	487 060 (98%)	4 970 (1%)	4 970 (1%)
~288 000 (тазартылған 300 000)	282 240 (98%)	2 880 (1%)	2 880 (1%)

3.3.4. Алдын ала тәжірибелер нәтижелері

Алдын ала тәжірибелерде OPUS деректер жиынтығымен Google Colab жабдығында mT5-small моделі пайдаланылды. Google Translate арқылы жасалған 140 000 синтетикалық сөйлеммен OPUS деректерін толықтыру барлық тіл жұптары бойынша BLEU ұпайларының тұрақты төмендеуіне алып келді, ал тек синтетикалық деректерде ғана үйрету нәтижелерді одан әрі нашарлатты. Бұл нәтижелер синтетикалық деректердің санынан гөрі сапасының шешуші рөл атқаратынын дәлелдеді.

Кесте 3.7 – Алдын ала тәжірибелер: mT5-small моделінің нәтижелері (OPUS деректері, барлық жұп орташасы)

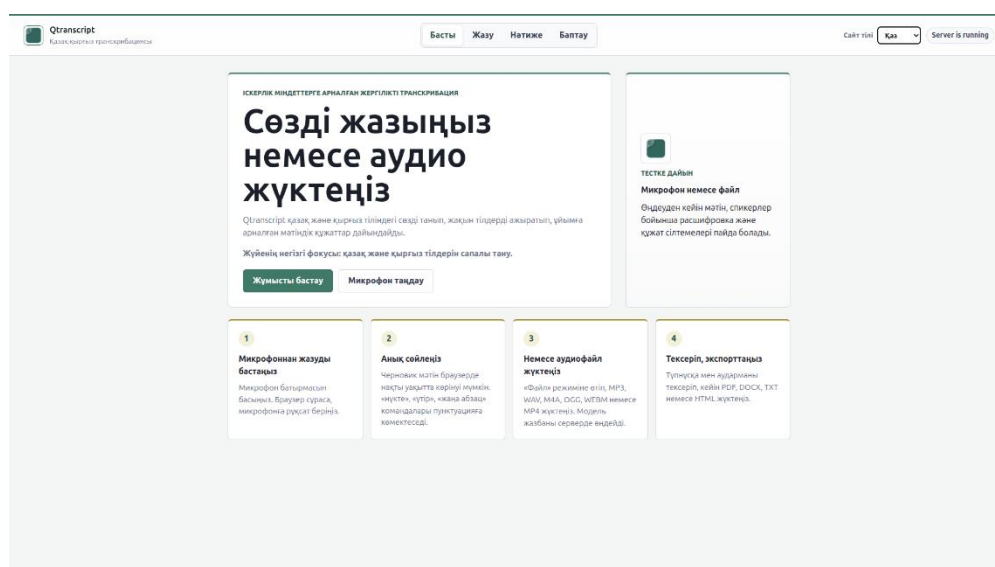
Деректер жиынтығы	BLEU (орташа)	Нәтиже
Тек OPUS деректері	төмен	Базалық нәтиже
OPUS + 140 000 синтетикалық (Google Translate)	одан да төмен	Нашарлады
Тек синтетикалық деректер	ең төмен	Одан да нашарлады

Осы алдын ала нәтижелер NLLB-200-ге негізделген синтетикалық корпус жасауға және деректерді құрылымдық тазартуға көшудің негіздемесі болды.

3.4 Ұсынылған кездесу стенограммаларын жасауға арналған автоматтандырылған жүйенің нәтижелері

Осы бөлімде қырғыз–қазақ тілдік жұбы үшін әзірленген автоматтандырылған кездесу стенограммаларын жасау жүйесінің интерфейсі мен жұмыс нәтижелері көрсетілген. Жүйе пайдаланушыға аудиожазбаларды өңдеу, сөйлеуді мәтінге автоматты түрде түрлендіру, қырғыз тіліндегі мәтінді қазақ тіліне аудару және дайын стенограмманы әртүрлі форматта сақтау мүмкіндіктерін ұсынады.

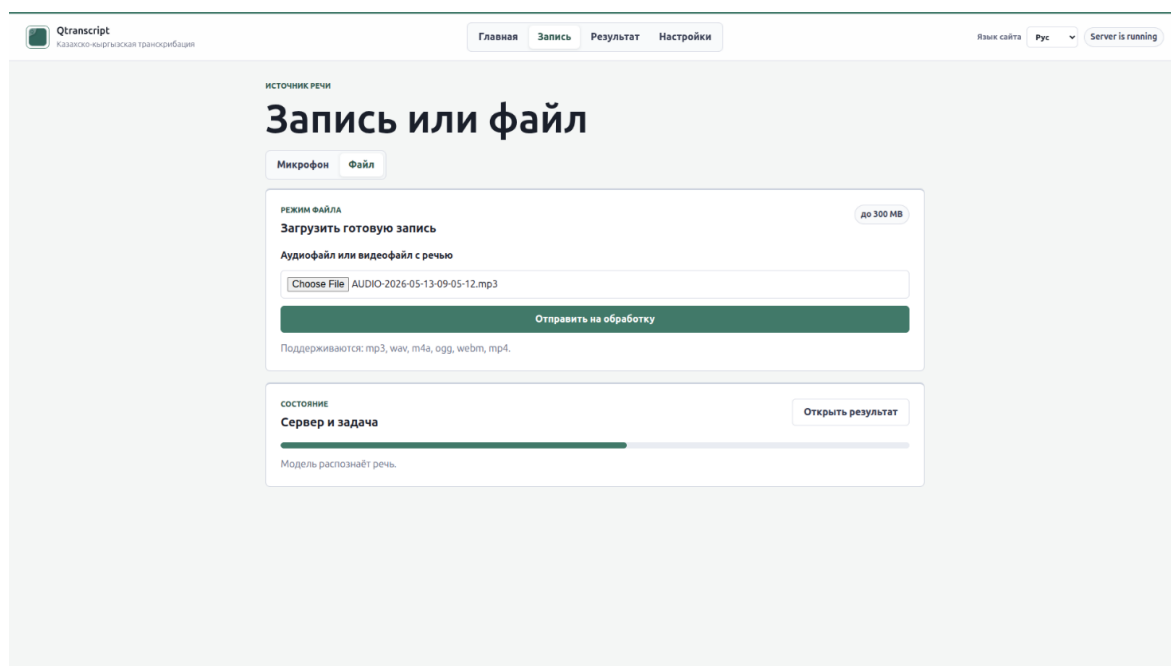
Сурет 3.7 де жүйенің негізгі интерфейсі көрсетілген. Бұл бет жүйемен жұмыс істеудің бастапқы кезеңі болып табылады және пайдаланушыға аудиодеректерді өңдеуге арналған негізгі функцияларды ұсынады. Интерфейс қарапайым және қолданушыға ыңғайлы түрде әзірленген.



Сурет 3.7 Жүйенің негізгі интерфейсі

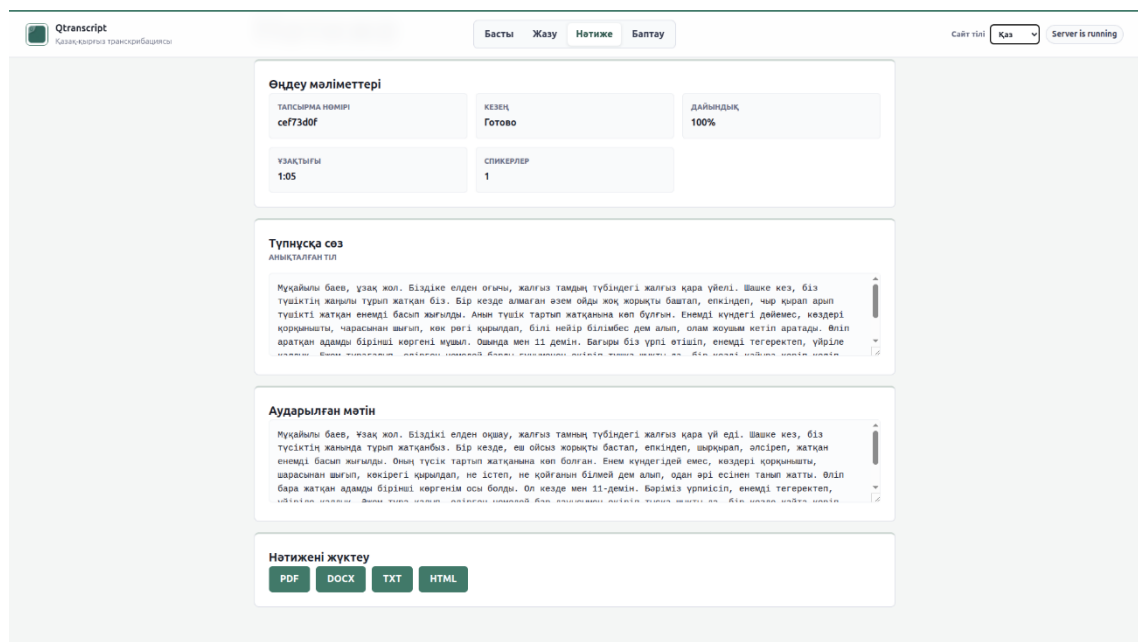
Сурет 3.8 те пайдаланушының аудиофайлды таңдау немесе микрофон арқылы тікелей сөйлеуді енгізу мүмкіндігі көрсетілген. Пайдаланушы қажетті аудионы таңдағаннан кейін «Өңдеуге жіберу» батырмасын басады. Осыдан кейін

аудиодеректер серверге жіберіліп, автоматты өңдеу кезеңі басталады. Сервер жағында сөйлеуді тану моделі іске қосылып, қырғыз тіліндегі сөздер мәтіндік форматқа түрлендіріледі. Одан кейін алынған мәтін нейрондық машиналық аударма модулі арқылы қазақ тіліне аударылады.



Сурет 3.8 - Пайдаланушының аудиофайлды таңдау немесе микрофон арқылы тікелей сөйлеуді енгізу мүмкіндігі

Сурет 3.9 да жүйенің өңдеу нәтижелері көрсетілген. Мұнда сөйлеуді тану нәтижесінде алынған қырғыз тіліндегі мәтін және оның қазақ тіліне аударылған нұсқасы бірге ұсынылады. Бұл пайдаланушыға бастапқы мәтін мен аударылған нәтижені салыстырып көруге мүмкіндік береді. Сонымен қатар интерфейсте дайын стенограмманы сақтау үшін бірнеше экспорттау батырмалары қарастырылған. Пайдаланушы нәтижені DOCX, PDF және TXT форматтарында жүктеп ала алады.



Сурет 3.9 - Жүйенің өңдеу нәтижелері

Сурет 3.10 да жүктелген құжаттың мысалы көрсетілген. Экспортталған құжатта сөйлеуді тану нәтижелері мен аударылған мәтін құрылымдалған түрде сақталады. Бұл дайын стенограмманы кейінгі өңдеу, архивтеу немесе ресми құжат айналымында қолдану мүмкіндігін қамтамасыз етеді.

Qtranscript: текст записи

Номер задачи: cef73d0f

Время начала записи: 2026-05-18 10:45

Длительность записи: 01:05

Язык речи: Казахский

Оригинальная речь

Мұқайылы баев, ұзақ жол. Біздіке елден оғычы, жалғыз тамдың түбіндегі жалғыз қара үйелі. Шашке кез, біз түшіктің жаңылы тұрып жатқан біз. Бір кезде алмаған әзем ойды жоқ жорықты баштап, епкіндеп, чыр қырап арып түшікті жатқан енемді басып жығылды. Анын түшік тартып жатқанына көп бұлғын. Енемді күндегі дәйемес, көздері қорқынышты, чарасынан шығып, көк рөгі қырылдап, білі нейір білімбес дем алып, олам жоушым кетіп аратады. Өліп аратқан адамды бірінші көргені мұшыл. Осында мен 11 демін. Бағыры біз үрпі өтішіп, енемді тегеректеп, үйріле қалдық. Әжем тұрағалып, елірген немедей барды ғуныменен өкіріп тұсқа шықты да, бір кезді қайыра керіп келіп, енемді дағы басып жығылды. Бізді нүй елден оғачы болсы да, ежемнің шыңғырған даушыменен бірін серін ел жин алып қалды.

Переведённый текст

Мұқайылы баев, Ұзақ жол. Біздіке елден оқшау, жалғыз тамның түбіндегі жалғыз қара үй еді. Шашке кез, біз түшіктің жанында тұрып жатқанбыз. Бір кезде, еш ойсыз жорықты бастап, епкіндеп, шырқырап, әлсіреп, жатқан енемді басып жығылды. Оның түсік тартып жатқанына көп болған. Енем күндегідей емес, көздері қорқынышты, шарасынан шығып, көкірегі қырылдап, не істеп, не қойғанын білмей дем алып, одан әрі есінен танып жатты. Өліп бара жатқан адамды бірінші көргенім осы болды. Ол кезде мен 11-демін. Бәріміз үрпиісіп, енемді тегеректеп, үйіріле қалдық. Әжем тұра қалып, елірген немедей бар дауысымен өкіріп тысқа шықты да, бір кезде қайта керіп келіп, енемді тағы басып жығылды. Біздіке елден оқшау болса да, әжемнің шыңғырған дауысымен біраз ел жиналып қалды.

Сурет 3.10 Жүктелген құжаттың мысалы

Ұсынылған жүйенің нәтижелері оның қырғыз тіліндегі сөйлеуді автоматты түрде тану, мәтінді қазақ тіліне аудару және дайын стенограмма құру міндеттерін тиімді орындай алатынын көрсетті. Жүйе мемлекеттік мекемелерде, білім беру ұйымдарында және бизнес кездесулерінде қолдануға жарамды практикалық шешім ретінде қарастырылуы мүмкін.

ҚОРЫТЫНДЫ

Осы диссертациялық жұмыста қырғыз–қазақ тілдік жұбы үшін кездесу стенограммаларын автоматты түрде қалыптастыруға арналған интеллектуалды жүйенің моделі мен құралдары әзірленіп, тиімділігі зерттелді. Жүргізілген зерттеулер мен алынған нәтижелер негізінде мынадай қорытындылар жасалды:

1. Ғылыми әдебиеттерге жүргізілген талдау қырғыз–қазақ тілдік жұбы үшін сөйлеуді автоматты тану, машиналық аударма және кездесу стенограммаларын қалыптастыру салаларындағы зерттеулердің жетіспеушілігін анық көрсетті. Whisper, MMS және TurkicASR сияқты заманауи ASR модельдері, сондай-ақ mT5, NLLB-200 және Gemma машиналық аударма жүйелері талданып, олардың архитектуралық ерекшеліктері, математикалық негіздері және осы тіл жұбы үшін қолдану мүмкіндіктері айқындалды.

2. Қырғыз–қазақ тілдік жұбы үшін 782 293 параллель сөйлемнен тұратын қомақты KazKyrParC параллель корпусы жасалды. Корпусты дайындау барысында деректерді жинау, тазарту және сапасын бақылау әдістері кешенді түрде қолданылды. Бұл ресурс ресурсы шектеулі осы тілдердің машиналық аударма мен NLP жүйелерін дамытуға арналған маңызды лингвистикалық база болып табылады.

3. Сөйлеуді тану сапасын жақсарту мақсатында CSE (Complete Set of Endings) технологиясы негізінде Morphology Aware Post Processing (MAPP) әдісі ұсынылды. Тәжірибелік бағалау бойынша MAPP конвейері қырғыз тіліндегі ASR шығысында WER метрикасын 0.1442-ден 0.0924-ке дейін төмендетті (35.9% жақсарту), BLEU ұпайын 69.44-тен 79.78-ге дейін арттырды, ChrF метрикасы 91.94-тен 93.37-ге жетті. Нәтижелер BERT негізіндегі қате детекторы мен MaxEnt перанкерінің бірлескен қолданысы агглютинативті тілдегі морфологиялық және фонетикалық қателерді тиімді жоюға мүмкіндік беретінін дәлелдеді.

4. Нөлдік (zero-shot) режимдегі ASR эксперименттерінде Mozilla Common Voice корпусынан алынған 1000 сөйлемдік тест жиынтығы бойынша үш модель бағаланды. Қырғыз тілінде MMS WER = 36.4% деңгейімен жетекші болды. Барлық модельдердің zero-shot режиміндегі нәтижелері нақты кездесу стенографиясы үшін жеткіліксіз деңгейде болып, жиналыс-домені деректерінде мамандандырылған fine-tuning жасаудың қажеттілігі негізделді.

5. Машиналық аударма тәжірибелерінде NLLB-200 1.3B моделі тазартылған 500 000 параллель сөйлемдік корпуста баптаудан кейін BLEU ұпайы 17.62-ден 43.54-ке дейін өсті (нөлдік деңгеймен салыстырғанда шамамен 2.5 есе), chrF 50.66-дан 76.71-ге жетті, WER 0.77-ден 0.42-ге дейін төмендеді. Деректерді сапалы тазарту синтетикалық деректерді кеңейтуден тиімдірек нәтиже беретіні дәлелденді.

6. Зерттеудің негізгі практикалық нәтижесі ретінде қырғыз–қазақ тілдік жұбының автоматтандырылған кездесу стенограммаларын қалыптастыру жүйесінің прототипі әзірленді. Client-server архитектурасына негізделген бұл жүйе аудиожазбаларды өңдеу, сөйлеуді мәтінге автоматты түрде түрлендіру (ASR), Morphology Aware Post Processing арқылы мәтінді түзету, NLLB-200

негізіндегі нейрондық машиналық аударма (қырғыздан қазаққа) және дайын стенограмманы DOCX, PDF, TXT форматтарында экспорттау мүмкіндіктерін интеграциялады.

Диссертациялық жұмыстың нәтижелері мемлекеттік органдарда, білім беру ұйымдарында, ғылыми мекемелерде және бизнес кездесулерінде қырғыз–қазақ тілдік ортадағы жиналыс хаттамаларын автоматтандыруда практикалық қолдануға жарамды. Сонымен қатар зерттеу барысында алынған KazKyrParC параллель корпусы, MAPR постпроцессинг әдісі және бапталған NLLB-200 моделі ресурсы шектеулі тілдерге арналған болашақ NLP зерттеулері үшін құнды ғылыми мұра болып табылады.

Болашақтағы зерттеу бағыттары ретінде мыналар белгіленді: жиналыс-домені үшін арнайы аудио деректер корпусын жасау; спикерді ажырату (diarization) модулін интеграциялау; MAPR конвейерінде морфологиялық генераторды кеңейтіп, CER метрикасын жақсарту; нақты уақыт режимінде жұмыс істейтін (real-time) стенографиялау мүмкіндігін дамыту; сондай-ақ LLM постпроцессинг арқылы стенограммадан автоматты қорытынды мен шешімдер тізімін шығару функцияларын қосу.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР

1. Rabiner, L.; Juang, B.H. Fundamentals of Speech Recognition; Prentice Hall: Upper Saddle River, NJ, USA, 1993.
2. Geoffrey Hinton; Deng, L.; Yu, D. et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal Processing Magazine 2012, 29(6), 82–97.
3. Hannun, A.; Case, C.; Casper, J. et al. Deep Speech: Scaling up End-to-End Speech Recognition. arXiv 2014, arXiv:1412.5567.
4. Radford, A.; Kim, J.W.; Xu, T. et al. Robust Speech Recognition via Large-Scale Weak Supervision. OpenAI Technical Report 2022.
5. Besacier, L.; Barnard, E.; Karpov, A.; Schultz, T. Automatic Speech Recognition for Under-Resourced Languages. Speech Communication 2014, 56, 85–100.
6. Hakkani-Tür, D.; Oflazer, K.; Tür, G. Statistical Morphological Disambiguation for Agglutinative Languages. Computers and the Humanities 2002, 36(4), 381–410.
7. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. Proceedings of ACL 2016, 1715–1725.
8. Ataman, D.; Federico, M. Compositional Representation of Morphologically Rich Input for Neural Machine Translation. Proceedings of NAACL 2018, 305–311.
9. Vaswani, A.; Shazeer, N.; Parmar, N. et al. Attention Is All You Need. Proceedings of NeurIPS 2017, 5998–6008.
10. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Proceedings of NeurIPS 2020.
11. Morris, A.; Maier, V.; Green, P. From WER and RIL to MER and WIL: Improved Evaluation Measures for Connected Speech Recognition. Proceedings of INTERSPEECH 2004, 2765–2768.
12. Di Gangi, M.; Cattoni, R.; Bentivogli, L. et al. MuST-C: A Multilingual Speech Translation Corpus. Proceedings of NAACL 2019, 2012–2017.
13. Pratap, V.; Xu, Q.; Sriram, A. et al. Massively Multilingual Speech Recognition for Low-Resource Languages. Proceedings of ICASSP 2020, 6589–6593.
14. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent Trends in Deep Learning Based Natural Language Processing. IEEE Computational Intelligence Magazine 2018, 13(3), 55–75.
15. Brown, T.; Mann, B.; Ryder, N. et al. Language Models are Few-Shot Learners. Proceedings of NeurIPS 2020, 1877–1901.
16. Oflazer, K. Computational Approaches to Morphology and Syntax of Turkic Languages. Computational Linguistics 1994, 20(4), 519–540.
17. Graves, A.; Mohamed, A.; Hinton, G. Speech Recognition with Deep Recurrent Neural Networks. Proceedings of ICASSP 2013, 6645–6649.

- 18.Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition. *Proceedings of ICASSP 2016*, 4960–4964.
- 19.Amodei, D.; Ananthanarayanan, S.; Anubhai, R. et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. *Proceedings of ICML 2016*, 173–182.
- 20.Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised Pre-training for Speech Recognition. *Proceedings of INTERSPEECH 2019*, 3465–3469.
- 21.Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *Proceedings of ICLR 2015*.
- 22.Luong, M.; Pham, H.; Manning, C.D. Effective Approaches to Attention-Based Neural Machine Translation. *Proceedings of EMNLP 2015*, 1412–1421.
- 23.Ott, M.; Edunov, S.; Baevski, A. et al. Fairseq: A Fast, Extensible Toolkit for Sequence Modeling. *Proceedings of NAACL Demonstrations 2019*, 48–53.
- 24.Zoph, B.; Knight, K. Multi-Source Neural Translation. *Proceedings of NAACL 2016*, 30–34.
- 25.Xue, L.; Constant, N.; Roberts, A. et al. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *Proceedings of NAACL 2021*, 483–498.
- 26.Costa-jussà, M.; Cross, J.; Çelebi, O. et al. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv 2022*, arXiv:2207.04672.
- 27.Tilk, O.; Alumäe, T. Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. *Proceedings of INTERSPEECH 2016*, 3047–3051.
- 28.Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL 2019*, 4171–4186.
- 29.Conneau, A.; Khandelwal, K.; Goyal, N. et al. Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of ACL 2020*, 8440–8451.
- 30.Gaido, M.; Papi, S.; Negri, M. et al. End-to-End Speech Translation for Low-Resource Languages. *Proceedings of ACL 2021*, 469–479.
- 31.Jurafsky, D.; Martin, J.H. *Speech and Language Processing*, 3rd ed.; Pearson: New York, NY, USA, 2021.
- 32.Furui, S. Recent Advances in Spontaneous Speech Recognition and Understanding. *Proceedings of IEEE Workshop on Spoken Language Technology 2004*, 1–6.
- 33.Waibel, A.; Bett, M.; Finke, M. Meeting Browser: Tracking and Summarizing Meetings. *Proceedings of DARPA Broadcast News Workshop 1998*.
- 34.Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. *Proceedings of ICML 2006*, 369–376.
- 35.Kudo, T.; Richardson, J. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. *Proceedings of EMNLP 2018*, 66–71.

36. Anguera, X.; Bozonnet, S.; Evans, N. et al. Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech and Language Processing* 2012, 20(2), 356–370.
37. Desplanques, B.; Thienpondt, J.; Demuynck, K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. *Proceedings of INTERSPEECH 2020*, 3830–3834.
38. Allahyari, M.; Pouriyeh, S.; Assefi, M. et al. Text Summarization Techniques: A Brief Survey. *International Journal of Advanced Computer Science and Applications* 2017, 8(10), 397–405.
39. Raffel, C.; Shazeer, N.; Roberts, A. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 2020, 21(140), 1–67.
40. Rennard, V.; Rosset, S.; Laurent, A. Automatic Meeting Summarization: A Survey. *arXiv* 2023, arXiv:2301.05768.
41. Barrault, L.; Chung, Y.A.; Mahadeokar, J. et al. SeamlessM4T – Massively Multilingual and Multimodal Machine Translation. *arXiv* 2023, arXiv:2308.11596.
42. Abduali B., Tukeyev, U.; Zhumanov, Zh.; Israilova, N. Study of Kyrgyz-Kazakh Neural Machine Translation. In *Proceedings of the 17th Asian Conference on Intelligent Information and Database Systems*; 2025; pp. 12. DOI: 10.1007/978-981-96-5881-7_21.
43. Nurmaganbet, D.; Tukeyev, U.; Shormakova, A.; Zhumanov, Z. Comparative Analysis of Models for Neural Machine Speech-to-Text Translation for Turkic State Languages. In *Intelligent Information and Database Systems*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2024; pp. 360–371. DOI: 10.1007/978-981-97-4985-0_28.
44. Karibayeva, A.; Karyukin, V.; Abduali B., Amirova, D. Speech Recognition and Synthesis Models and Platforms for the Kazakh Language. *Information* 2025, 16(10), 879. DOI: 10.3390/info16100879.
45. 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* 2025, 8(9), 13–23. DOI: 10.53894/ijirss.v8i9.10605.
46. Karibayeva, A.; Karyukin, V.; Tukeyev, U.; Abduali B., Amirova, D.; Rakhimova, D.; Aliyev, R.; Shormakova, A. The Development and Experimental Evaluation of a Multilingual Speech Corpus for Low-Resource Turkic Languages. *Applied Sciences* 2025, 15(24), 12880. DOI: 10.3390/app152412880.
47. Tukeyev, U.; Shormakova, A.; Karibayeva, A.; Rakhimova, D.; Abduali B., Amirova, D.; Rakhmanberdi, N.; Aliyev, R. An Integrated Approach to Adapting Open-Source AI Models for Machine Translation of Low-Resource Turkic Languages. *Computers* 2026, 15(2), 73. DOI: 10.3390/computers15020073.

48. Malik, M., Malik, M.K., Mehmood, K. et al. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80, 9411–9457, 2021. <https://doi.org/10.1007/s11042-020-10073-7>
49. Prabhavalkar, R., Sainath, T.N., et al. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023. <https://doi.org/10.1109/TASLP.2023.3328283>
50. Baevski, A., Zhou, Y., Mohamed, A., Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *NeurIPS* 2020. arXiv:2006.11477
51. Radford, A., Kim, J.W., Xu, T., et al. Robust Speech Recognition via Large-Scale Weak Supervision. *ICML* 2022. arXiv:2212.04356
52. Schneider, S., Baevski, A., Collobert, R., Auli, M. wav2vec: Unsupervised Pre-training for Speech Recognition. *Interspeech* 2019. arXiv:1904.05862
53. Conneau, A., et al. Unsupervised Cross-lingual Representation Learning at Scale (XLM-R). *ACL* 2020. arXiv:1911.02116
54. Kudo, T., Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *EMNLP* 2018. arXiv:1808.06226
55. Sennrich, R., Haddow, B., Birch, A. Neural Machine Translation of Rare Words with Subword Units (BPE). *ACL* 2016. arXiv:1508.07909
56. Alam, T., Khan, A., Alam, F. Punctuation Restoration using Transformer Models for Resource-Rich and -Poor Languages. *EMNLP Workshop*, 2020.
57. Nagy, A., Bial, B., Ács, J. Automatic punctuation restoration with BERT models. arXiv:2101.07343, 2021.
58. Devlin, J., Chang, M.W., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL* 2019. arXiv:1810.04805
59. Desplanques, B., Thienpondt, J., Demuynck, K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. *Interspeech* 2020. arXiv:2005.07143
60. Dawalatabad, N., Ravanelli, M., et al. ECAPA-TDNN Embeddings for Speaker Diarization. *Interspeech* 2021, pp. 3560–3564. <https://doi.org/10.21437/Interspeech.2021-941>
61. Lewis, M., et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *ACL* 2020. arXiv:1910.13461
62. Xue, L., et al. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *NAACL* 2021. arXiv:2010.11934
63. Conneau, A., et al. Unsupervised Cross-lingual Representation Learning at Scale (XLM-R). *ACL* 2020. arXiv:1911.02116
64. Barrault, L., et al. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation. *Meta AI*, 2023. arXiv:2308.11596
65. Joshi, P., et al. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *ACL* 2020. arXiv:2004.09813

- 66.Graves, A., Mohamed, A., Hinton, G. Speech Recognition with Deep Recurrent Neural Networks. IEEE ICASSP 2013, pp. 6645–6649. DOI: 10.1109/ICASSP.2013.6638947
- 67.Hannun, A., Case, C., Casper, J., et al. Deep Speech: Scaling up end-to-end speech recognition. Baidu Research, arXiv:1412.5567, 2014.
- 68.Chan, W., Jaitly, N., Le, Q.V., Vinyals, O. Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition. IEEE ICASSP 2016, pp. 4960–4964. DOI: 10.1109/ICASSP.2016.7472621
- 69.Sennrich, R., Haddow, B., Birch, A. Neural Machine Translation of Rare Words with Subword Units. ACL 2016. arXiv:1508.07909
- 70.Tilk, O., Alumäe, T. Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. Interspeech 2016, pp. 3047–3051. DOI: 10.21437/Interspeech.2016-1517
- 71.Vaswani, A., Shazeer, N., Parmar, N., et al. Attention Is All You Need. NeurIPS 2017. arXiv:1706.03762
- 72.Baevski, A., Zhou, Y., Mohamed, A., Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. NeurIPS 2020. arXiv:2006.11477
- 73.Conneau, A., Khandelwal, K., Goyal, N., et al. Unsupervised Cross-lingual Representation Learning at Scale. ACL 2020. arXiv:1911.02116
- 74.Radford, A., Kim, J.W., Xu, T., et al. Robust Speech Recognition via Large-Scale Weak Supervision. ICML 2022. arXiv:2212.04356
- 75.NLLB Team, Costa-jussà, M.R., Cross, J., et al. No Language Left Behind: Scaling Human-Centered Machine Translation. Meta AI, arXiv:2207.04672, 2022.
- 76.Rennard, V., Shang, G., Hunter, J., Vazirgiannis, M. Abstractive Meeting Summarization: A Survey. Transactions of the Association for Computational Linguistics, 11: 861–884, 2023. DOI: 10.1162/tacl_a_00578
- 77.Barrault, L., et al. (NLLB Team & Meta AI). SeamlessM4T: Massively Multilingual & Multimodal Machine Translation. arXiv:2308.11596, 2023.