Жоба туралы қысқаша ақпарат

Жоба аты	AP23487816 «Түркі тілдеріндегі сөйлеу транскрипциясы мен кездесу хаттамаларын жасаудың нейрондық модельдерін зерттеу»
Жоба өзектілігі	Түркі мемлекеттері арасындағы ынтымақтастықтың белсенді дамуын ескере отырып, түркі тілдерінің өзара әрекеттесуін дамыту мәселесі өте өзекті. Түркі тілдерінің өзара әрекеттесуін дамытудың маңызды бағыттарының бірі - түркі тілдерінде өткізілген кездесулердің хаттамаларын жедел дайындауды қамтамасыз ету.
Жоба мақсаты	Жобаның мақсаты - түркі тілдерінде сөйлеу транскрипциялары мен кездесу хаттамаларын жасауға арналған нейрондық модельдерге арналған кешенді технологияны әзірлеу. Зерттеліп жатқан тілдер - қазақ, әзірбайжан, қырғыз, түрік, түрікмен және өзбек тілдері.
Жоба міндеттері	Түркі тілдеріндегі сөйлеу транскрипциялары мен кездесу хаттамаларын қалыптастырудың нейрондық модельдерінің күрделі технологиясын зерттеу мақсатына жету үшін жоба келесі міндеттерді қояды: - түркі тілдеріне арналған транскрипция мәтініндегі сөйлеуді тану технологиясын әзірлеу (іріктеу) және зерттеу; - түркі тілдерінің мәтіндік транскрипцияларын қазақ тіліне машиналық аударма жасау және зерттеу; - қазақ тіліндегі стенограмма мәтінінен отырыстардың қысқаша хаттамаларын қалыптастыруды әзірлеу және зерттеу; - отырыстың қысқаша хаттамасын қазақ тілінен әзірбайжан, қырғыз, түрік, түрікмен және өзбек тілдеріне машиналық аударманы әзірлеу және зерттеу.
Күтілетін және қол жеткізілген нәтижелер	Жобаның күтілетін нәтижелері: 1) түркі тілдерінде сөйлеу транскрипттері мен кездесу хаттамаларын жасауға арналған интеграцияланған технологияның модельдері, әдістері мен құралдары ; 2) түркі тілінің транскрипттерін қазақ тіліне машиналық аудару модельдері, әдістері мен құралдары; 3) нейрондық модельдерді оқытуға арналған деректер жиынтығы; 4) транскрипттерден қазақ тілінде қысқа кездесу хаттамаларын жасауға арналған модельдер, әдістер мен құралдар. 2025 жылға дейін қол жеткізілген жоба нәтижелері: Әрқайсысы 780 000 параллель сөйлемнен тұратын бес жұп ресми түркі тілдері үшін параллель корпустар жасалды. Алты ресми түркі тілінің көптілді параллель корпусы жасалды, онда 3 885 542 параллель сөйлем бар. Бес түркі тілінен транскрипттерді қазақ тіліне машиналық аударуға арналған нейрондық модельдер әзірленді және әзірленген параллель корпустарды пайдаланып зерттелді. Әзірленген параллель корпустарға авторлық құқық сертификаттары және бес түркі тілін қазақ тіліне машиналық аударуға арналған NLLB 200 1.3B нейрондық моделіне арналған жетілдіру бағдарламасы алынды. Абстрактілі тәсілді қолдана отырып, транскрипттерден қазақ тілінде қысқаша кездесу хаттамаларын жасауға арналған модельдер, әдістер және құралдар әзірленді. 116 000 жазбадан тұратын синтетикалық қазақша қорытындылау корпусы әзірленді. Қазақ жаңалықтар веб-сайттарынан деректер жиынтығын талдауға арналған бағдарлама әзірленді. Бұл бағдарлама baq.kz веб-сайтынан 74 726 жазбадан тұратын қорытындылау деректер жиынтығын жинады . Gemma-3-4B- it, Sherkala (Llama -3.1 Sherkala8BChat) және mT5 негізіндегі нейрондық модельдер әзірленген қорытындылау корпустарында одан әрі оқытылды .
Зерттеу тобы мүшелеріні	1. Тукеев Уалшер Scopus ID: 55701639900 https://www.scopus.com/authid/detail.uri?authorId=55701639900 Orcid

ң аты-жөндері, олардың идентификаторлары (Scopus авторының идентификаторы, зерттеушінің идентификаторы, ORCID, егер бар болса) және тиісті профильдерге сілтемелер	0000-0001-9878-981X https://orcid.org/0000-0001-9878-981X Web of Science ResearcherID: O-1614-2014 RG https://www.researchgate.net/profile/Ualsher-Tukeyev GoogleScholar https://scholar.google.com/citations?view_op=list_works&hl=ru&user=VwnDwusAAAAJ&user=VwnDwusAAAAJ 2. Шормакова Асем Ноябрьевна Scopus 0000-0002-1637-4643 https://www.scopus.com/authid/detail.uri?authorId=55786942400 Orcid 0000-0002-1637-4643 https://orcid.org/0000-0002-1637-4643 WebOfScience LNF-7076-2024 https://www.webofscience.com/wos/author/record/64005178 RG https://www.researchgate.net/profile/Assem-Shormakova GoogleScholar https://scholar.google.com/citations?user=8dzErTAAAAAJ&hl=ru 3. Кәрібаева Айдана Сейілғазықызы Scopus Author ID – 57196004542 https://www.scopus.com/authid/detail.uri?authorId=57196004542 ResearcherID Web of Science - AAR-4134-2020 https://www.webofscience.com/wos/author/record/AAR-4134-2020 ORCID – 0000-0002-2023-1573 https://orcid.org/0000-0002-2023-1573 4. Рахимова Диана Рамазановна Scopus Author ID – 55682794500 https://www.scopus.com/authid/detail.uri?authorId=55682794500 ResearcherID Web of Science - https://www.webofscience.com/wos/author/record/34409631 - 0 ORCID – 0000-0003-1427-198X https://orcid.org/0000-0003-1427-198X 5. Абдуали Балжан Абдирасилқызы Scopus Author ID - 57191430929 https://www.scopus.com/authid/detail.uri?authorId=57191430929 ResearcherID Web of Science - P-4962-2017 https://www.webofscience.com/wos/author/record/P-4962-2017 ORCID – 0000-0003-0140-4181 https://orcid.org/0000-0003-0140-4181 GoogleScholar https://scholar.google.com/citations?user=IKyuPX4AAAAJ&hl=ru 6. Әмірова Дина Тарихатқызы Scopus: 57196009653, https://www.scopus.com/authid/detail.uri?authorId=57196009653 Orchid: 0000-0002-0728-905X, https://orcid.org/0000-0002-0728-905X Web of Science ResearcherID: P-5668-2017, https://www.webofscience.com/wos/author/record/39381626 Google Scholar https://scholar.google.com/citations?user=8KZLCwYAAAAJ&hl=ru&oi=ao ResearchGate https://www.researchgate.net/profile/Dina-Amirova 7. Алиев Рашид Шамильевич ResearcherID Web of Science - MFH-7654-2025 https://www.webofscience.com/wos/author/record/MFH-7654-2025 ORCID - 0009-0003-5432-0480 https://orcid.org/0009-0003-5432-0480 8. Рахманберді Назым Алмазқызы 9. Жабаев Талгат Рыспекович 10. Сегизбаева Жансая Елдибаевна 11. Омархалиева Назерке Галымжанқызы 12. Ахметова Дильназ Шухратовна 13. Очеретин Максим Павлович

Қолдану	Түркі тілдерінде өткізілген кездесулердің хаттамаларын дайындау кезінде.
Нақты аналогтарға қарағанда басымдық	2025 жылдың соңына дейін нақты аналогтары табылған жоқ.
Жарияланымдар тізімі (URL, DOI көрсетілген)	1. Тукеев, У.; Шормакова, А.; Карибаева, А.; Рахимова, Д.; Абдуали, Б.; Амирова, Д.; Рахманберди, Н.; Алиев, Р. Ресурстары аз түркі тілдерін машиналық аудару үшін ашық бастапқы кодты жасанды интеллект модельдерін бейімдеудің кешенді тәсілі. Computers 2026, 15, 73. https://doi.org/10.3390/computers15020073 **(1-тоқсан, CS=8.1, 84%, 2025)** 2. Тукеев, У.; Рысбек, Б. Түркі тілдеріне арналған морфологияға негізделген сегменттеу және тонкеизация: CSE басшылыққа алынған құрылым (Қазақ ісі). Information 2026, 17, 128. https://doi.org/10.3390/info17020128 **(2-тоқсан, CS=6.5, 74%, 2024)** 3. Жансая Сегізбаева, Марек Милош Қазақ дауысты дыбыстары мен дауыссыз дыбыстар үндестігінің есептеу моделі. Journal of Problems in Computer Science and Information Technologies. ҚазҰУ. (Жариялауға қабылданды). Авторлық куәліктер. 4. 2025 жылғы 22 қыркүйектегі № 62307 авторлық куәлік. ШОРМАҚОВА ӘСЕМ НОЯБРЕВНА, Тукеев Уалшер, Рахманберді Назым Алмазқызы. 5. 2025 жылғы 14 қазандағы № 62995 авторлық куәлік. «KazKyrParC» қазақ-қырғыз мәтіндерінің параллельді корпусы. ӘБДУӘЛИ БАЛЖАН ӘБДІРӘСІЛҚЫЗЫ, Тукеев Уалшер. 6. 2025 жылғы 1 қазандағы № 62585 авторлық куәлік. "KazTrParC" қазақ-түрік параллель мәтіндер корпусы. РАХИМОВА ДИАНА РАМАЗАНҚЫЗЫ, Тукеев Уалшер. 7. Авторлық куәлік 2025 жылғы 14 қазандағы № 62997. «KazTurkmParC» қазақ-түрікмен мәтіндерінің параллель корпусы ӘМІРОВА ДИНА ТАРИХАТҚЫЗЫ, Тукеев Уалшер. 8. 2025 жылғы 14 қазандағы № 62996 авторлық куәлік. "KazUzParS" қазақ-өзбек параллель корпусы. Қазақ-өзбек тілінің авторы, КӘРІБАЕВА АЙДАНА СЕЙІЛҒАЗЫҚЫЗЫ, Тукеев Уалшер. 2025 жылғы 13 қазандағы № 62915 авторлық куәлік. NLLB-200-1.3B нейрондық үлгісін әзірбайжан-қазақ, қырғыз-қазақ, түрік-қазақ, түркімен-қазақ және өзбек-қазақ тілдерінің параллельді корпусында дәл баптау бағдарламасы. Авторлары: Назым Рахманберді, Уалшер Тукеев, Рашид Шамильұлы Әлиев.
Патент туралы ақпарат	