Краткая информация о проекте

Наименование	AP23487816 «Исследование нейронных моделей формирования стенограммы речи и протоколов совещаний на тюркских языках»
Актуальность	Проблема развития взаимодействия тюркских языков в связи активным развитием сотрудничества тюркских государств стоит весьма актуально. Одним из важных направлений развития взаимодействия тюркских языков является обеспечение оперативного оформления протоколов совещаний, проводимых на тюркских языках.
Цель	Целью проекта является создание комплексной технологии нейронных моделей формирования стенограммы речи и протоколов совещаний на тюркских языках. В качестве объектов исследования берутся языки: казахский, азербайджанский, кыргызский, турецкий, туркменский, узбекский.
Задачи	Для достижения поставленной цели исследования комплексной технологии нейронных моделей формирования стенограммы речи и протоколов совещаний на тюркских языках в проекте ставятся следующие задачи: - разработка (выбор) и исследование технологии распознавания речи тюркских языков в текст стенограммы; - разработка и исследование машинного перевода текста стенограмм тюркских языков на казахский язык; - разработка и исследование формирования из текста стенограмм кратких протоколов совещаний на казахском языке; - разработка и исследование машинного перевода краткого протокола совещания с казахского языка на азербайджанский, кыргызский, турецкий, туркменский, узбекский языки.
Ожидаемые и достигнутые результаты	Ожидаемые результаты проекта: 1) модели, методы и средства комплексной технологии формирования стенограмм речи и протоколов совещаний на тюркских языках; 2) модели, методы и средства машинного перевода текстов стенограмм тюркских языков на казахский язык; 3) наборы данных для обучения нейронных моделей; 4) модели, методы и средства формирования из текста стенограмм кратких протоколов совещаний на казахском языке. Достигнутые результаты проекта на 2025 год: Разработаны параллельные корпуса для пяти пар государственных тюркских языков, каждый корпус объемом 780 тысяч параллельных предложений. Разработан мультиязычный параллельный корпус шести государственных тюркских языков объемом 3 885 542 параллельных предложений. Разработаны и исследованы нейронные модели машинного перевода текстов стенограмм пяти тюркских языков на казахский язык на разработанных параллельных корпусах. Получены авторские свидетельства на разработанные параллельные корпуса и программу тонкой настройки нейронной модели NLLB 200 1.3B машинного перевода пяти тюркских языков на казахский язык. Разработаны модели, методы и средства формирования из текста стенограмм кратких протоколов совещаний на казахском языке абстрактивным подходом. Разработан синтетический казахский корпус суммаризации объемом 116 тысяч записей. Разработана программа парсинга датасета с казахских новостных сайтов. Данной программой собран датасет суммаризации с сайта baq.kz объемом 74726 записей. Выполнено дообучение нейронных моделей Gemma-3-4B-it, Sherkala (Llama-3.1-Sherkala-8B-Chat), mT5-base на разработанных корпусах суммаризации.

Имена и фамилии членов исследовательской группы с их идентификаторами (Scopus Author ID, Researcher ID, ORCID, при наличии) и ссылками на соответствующие профили	1. Тукеев Уалшер Scopus ID: 55701639900 https://www.scopus.com/authid/detail.uri?authorId=55701639900 Orcid 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
Использование	При оформлении протоколов совещаний, проводимых на тюркских языках.
Приоритет в сравнении с реальными аналогами	Реальные аналоги на конец 2025 года не обнаружены.
Список публикаций с ссылками на них	1. 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, 73. https://doi.org/10.3390/computers15020073 **(Q1, CS=8.1, 84%, 2025)** 2. Tukeyev, U.; Rysbek, B. Morphology-Aware Segmentation and Tokenization for Turkic Languages: A CSE-Guided Framework (The Kazakh Case). Information 2026, 17, 128. https://doi.org/10.3390/info17020128 **(Q2, CS=6.5, 74%, 2024)** 3. Zhansaya Segizbayeva, Marek Miłosz Computational Model Of Kazakh Vowel–Consonant Harmony. Journal of Problems in Computer Science and Information Technologies. KazNU. (Принята к публикации). Авторские свидетельства. 4. Авторское Свидетельство № 62307 от «22» сентября 2025 года. “Казахско–азербайджанский параллельный корпус текстов «KazAzerParC»”. ШОРМАКОВА АСЕМ НОЯБРЕВНА, Тукеев Уалшер, Рахманберді Назым Алмазқызы. 5. Авторское Свидетельство № 62995 от «14» октября 2025 года. Казахско–кыргызский параллельный корпус текстов «KazKyrParC». АБДУАЛИ БАЛЖАН АБДИРАСИЛҚЫЗЫ, Тукеев Уалшер. 6. Авторское Свидетельство № 62585 от «1» октября 2025 года. Казахско–турецкий параллельный корпус текстов «KazTrParC». РАХИМОВА ДИАНА РАМАЗАНОВНА, Тукеев Уалшер. 7. Авторское Свидетельство № 62997 от «14» октября 2025 года. Казахско–туркменский параллельный корпус текстов «KazTurkmParC» ӘМІРОВА ДИНА ТАРИХАТҚЫЗЫ, Тукеев Уалшер. 8. Авторское Свидетельство № 62996 от «14» октября 2025 года. Казахско–узбекский параллельный корпус «KazUzParC». КӘРІБАЕВА АЙДАНА СЕЙІЛҒАЗЫҚЫЗЫ, Тукеев Уалшер. Авторское Свидетельство № 62915 от «13» октября 2025 года. Программа тонкой настройки (fine-tuning) нейронной модели NLLB-200-1.3B на параллельных корпусах азер-каз, кыргыз-каз, турик-каз, туркмен-каз, узбек-каз по отдельности. РАХМАНБЕРДІ НАЗЫМ АЛМАЗҚЫЗЫ, Тукеев Уалшер, Алиев Рашид Шамильевич.

Информация о патентах	