Краткая информация о проекте

Наименование	Исследование автоматической генерации параллельных корпусов речи тюркских языков и их использование для нейронных моделей
Актуальность	Тюркоязычные исследования нейронного машинного перевода речи практически отсутствуют в силу трудностей создания параллельных корпусов речи. Создание корпуса речи затруднено в силу отсутствия фазы TTT (Text-To-Text) для большинства тюркских языков. Создание параллельного корпуса речи для тюркоязычных языков является актуальной задачей с точки зрения развития автоматических систем перевода, лингвистических исследований, информационных технологий и сохранения культурного наследия.
Цель	Целью проекта является исследование технологии для формирования параллельных корпусов речи и использованием их для обучения нейронного машинного перевода речи тюркских языков. В качестве объектов исследования берутся три пары языков: казахский-татарский, казахский-турецкий, казахский-узбекский
Задачи	1. разработка и исследование моделей и алгоритмов каскадной схемы машинного перевода речи тюркских языков с использованием модели морфологии 2. формирование параллельных корпусов речи с использованием каскадной технологии для тюркских языков; 3. разработка и обучение технологии нейронной модели машинного перевода речи тюркских языков на сформированных параллельных корпусах речи; 4. оценка разработанных технологий формирования корпуса речи и машинного перевода речи тюркских языков.
Ожидаемые и достигнутые результаты	Ожидаемые результаты проекта: Будет разработана каскадная схема машинного перевода речи с казахского языка на тюркские языки, сформированы параллельные аудио корпуса с казахского на турецкий, татарский и узбекские языка с применением их в обучении машинном перевода речи, также будут получены и оценены результаты каскадных и нейронных технологий машинного перевода речи с использованием метрик качества. За весь период реализации проекта будут подготовлены и опубликованы статьи в рейтинговых журналах согласно требованиям конкурсной документации: - не менее 3 (трех) статей и (или) обзоров в рецензируемых научных изданиях по научному направлению проекта, индексируемых в Science Citation Index Expanded и входящих в 1 (первый), 2 (второй) и (или) 3 (третий) квартиль по импакт-фактору в базе Web of Science и (или) имеющих процентиль по CiteScore в базе Scopus не менее 60 (шестидесяти);

	- не менее 1 (одной) статьи или обзора в рецензируемом зарубежном или отечественном издании, рекомендованном КОКНВО; - 1 монография отечественного издания.
Имена и фамилии членов исследовательской группы с их идентификаторами (Scopus Author ID, Researcher ID, ORCID, при наличии) и ссылками на соответствующие профили	1) 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 2) Тукеев Уалшер Ануарбекович Scopus Author ID – 55701639900 https://www.scopus.com/authid/detail.uri?authorId=55701639900 ResearcherID Web of Science https://www.webofscience.com/wos/author/record/2024064 - 2 ORCID – 0000-0001-9878-981X https://orcid.org/0000-0001-9878-981X 3) 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 4) Шормакова Асем Ноябрьевна Scopus Author ID - 55786942400 https://www.scopus.com/authid/detail.uri?authorId=55786942400 ResearcherID Web of Science – KCK-9845-2024 https://www.webofscience.com/wos/author/record/KCK-9845-2024 ORCID – 0000-0002-1637-4643 https://orcid.org/0000-0002-1637-4643 5) Карюкин Владислав Игорьевич Scopus Author ID – 57218952479; https://www.scopus.com/authid/detail.uri?authorId=57218952479 ORCID – 0000-0002-8768-0349; https://orcid.org/my-orcid?orcid=0000-0002-8768-0349 ResearcherID Web of Science – AAW-8974-2020; https://www.webofscience.com/wos/author/record/AAW-8974-2020 6) Әмірова Дина Тарихатқызы Scopus Author ID – 57196009653 https://www.scopus.com/authid/detail.uri?authorId=57196009653-1 ResearcherID Web of Science https://www.webofscience.com/wos/author/record/39381626-1 ORCID - 0000-0002-0728-905X https://orcid.org/0000-0002-0728-905X?lang=en 7) Абдуали Балжан Абдрасилқызы 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 8) 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 9) Мысов Олег ResearcherID Web of Science - PRQ-5683-2026 https://www.webofscience.com/wos/author/record/PRQ-5683-2026 ORCID – 0009-0006-9609-4073 https://orcid.org/0009-0006-9609-4073
Использование	Параллельные корпусы речи являются ценным ресурсом для разработки и обучения языковых моделей. Следует отметить, что эти модели могут быть использованы в таких задачах, как автоматическое распознавание речи, синтез речи и др. Соответственно, результаты технологии синхронного машинного перевода могут использоваться для систем понимания естественных языков, при создании чат-бот системы
Приоритет в сравнении с реальными аналогами	Нет аналогов
Список публикаций с ссылками на них	1) Karibayeva, Aidana, Vladislav Karyukin, Balzhan Abduali, and Dina Amirova. 2025. "Speech Recognition and Synthesis Models and Platforms for the Kazakh Language" *Information* 16, no. 10: 879. https://doi.org/10.3390/info16100879 (Q2, 74%) (Scopus - Document Details) 2) Karibayeva, Aidana, Vladislav Karyukin, Ualsher Tukeyev, Balzhan Abduali, Dina Amirova, Diana Rakhimova, Rashid Aliyev, and Assem Shormakova. 2025. "The Development and Experimental Evaluation of a Multilingual Speech Corpus for Low-Resource Turkic Languages" *Applied Sciences* 15, no. 24: 12880. https://doi.org/10.3390/app152412880 (Q2, 69%) (Scopus - Document Details) 3) D. Rakhimova, E. Adalı and **A. Karibayeva**, "Aligning Sentences for Kazakh-Turkish Parallel Corpora," *2024 9th International Conference on Computer Science and Engineering (UBMK)*, Antalya, Turkiye, 2024, pp. 175-180 https://doi.org/10.1109/UBMK63289.2024.10773567 (Scopus - Document Details) **4) Karibayeva Aidana**, Rakhimova Diana, Askapova, Madina. Parallel Speech Corpus of Kazakh and Turkish Languages: Methodology of Creation and Application Prospects. In book: Advances in Computational Collective Intelligence (pp.442-453) November 2025. https://doi.org/10.1007/978-3-032-10202-7_30 (Scopus - Document Details)

	5) Кәрібаева А.С., Аскапова М.К. Параллельді қазақ-түрік сөйлеу корпусын қалыптастырудың әдісі мен моделін құру. Материалы международной конференции студентов и молодых учёных «Фараби әлемі» Алматы, Казахстан, 3-5 апреля 2025 года, 453-455 стр. 6) А. Кәрібаева, Б. Абдуали, Д. Әмірова. Нейронды машиналық аудрамала кері аударма технологиясын талдау. МАТЕРИАЛЫ X Международной научно-практической конференции «Информатика и прикладная математика», Республика Казахстан, г.Алматы) ISBN 978-601-228-472-0, стр. 379-385 7) Aidana Karibayeva, Vladislav Karyukin, Diana Rakhimova, Dina Amirova, Rashid Aliyev, Adina Karybayeva. Collecting, Processing, and Evaluating the Performance of Kazakh Automatic Speech Recognition. Asian Conference on Intelligent Information and Database Systems. Springer Nature Singapore, pp. 260-271 (Scopus, процентиль 39). 8) Карибаева А.С. АНАЛИЗ ОСОБЕННОСТЕЙ КАЗАХСКО-УЗБЕКСКОЙ ЯЗЫКОВОЙ ПАРЫ ДЛЯ РАСПОЗНАВАНИЯ РЕЧИ. Сборник научных трудов и материалов Международного форума «Русистика и тюркология: литературные традиции, культурный трансфер и перевод», посвященный 95-летию Бухарского государственного университета, Бухара 2025, -С. 409-413. 9) Karibayeva A., Karyukin V., Aliyev R., Rakhimova D. Study of the speech recognition systems for Kazakh language. Proceedings of the XII International conference on computer processing of Turkic languages “TURKLANG 2024”. -P. 295-298.
Информация о патентах	не предусмотрено