

АННОТАЦИЯ
Абдуали Балжан Абдирасилкызы
за диссертацию, подготовленную для получения
степени доктора философии (PhD)
по образовательной программе
8D07109 – «Автоматизация и Internet of Things»
на тему «Разработка и исследование модели и средств
автоматизированной системы формирования стенограмм совещаний
для кыргызско-казахской языковой пары»

Актуальность темы исследования. В настоящее время стремительное развитие цифровых технологий повышает значимость областей обработки естественного языка, автоматического распознавания речи и машинного перевода. Несмотря на это, такие тюркские языки, как кыргызский и казахский, относятся к числу низкоресурсных языков. Недостаток размеченных аудиоданных, параллельных текстов и качественных корпусов для этих языков затрудняет эффективную разработку современных интеллектуальных систем.

Несмотря на лингвистическую близость кыргызско-казахской языковой пары, системы обработки речи, автоматического перевода и суммаризации текста в данном направлении изучены недостаточно. Агглютинативная природа тюркских языков, морфологическая сложность и большое разнообразие словоформ значительно усложняют процесс построения моделей.

Кроме того, в современном обществе возрастает потребность в автоматическом формировании стенограмм встреч и совещаний. Такие системы включают несколько сложных этапов: преобразование речи в текст, обработку текста, выделение ключевой информации и краткое суммирование. Особенно в многоязычной среде подобные решения позволяют ускорить обмен информацией и сократить затраты человеческих ресурсов.

В связи с этим разработка и исследование моделей и инструментов для автоматического формирования стенограмм совещаний для кыргызско-казахской языковой пары является актуальной научной и практической задачей.

В настоящее время технологии искусственного интеллекта широко применяются в области обработки человеческой речи, анализа текста и автоматизированного документирования информации. В частности, особое значение приобретают направления автоматического распознавания речи, машинного перевода и автоматизации стенограмм совещаний. Эти направления играют важную роль в развитии образования, государственного управления, бизнеса, медиа, международной коммуникации и цифровых сервисов. Особенно возрастает спрос на системы, работающие в многоязычной среде, а исследование эффективных моделей для сравнительно низкоресурсных языков, таких как казахский, становится актуальной задачей.

Системы автоматического распознавания речи позволяют преобразовывать устную речь человека в текст. Такие системы применяются при цифровизации аудиоархивов, создании субтитров, в вспомогательных технологиях, call-center системах и автоматической регистрации совещаний. В последние годы большой интерес вызывают модели Whisper, MMS и Turkic ASR, ориентированные на тюркские языки. Их основными особенностями являются многоязычность, адаптация к низкоресурсным языкам и повышение качества распознавания речи в реальных условиях.

Сфера машинного перевода также развивается очень интенсивно. Если ранее системы перевода были ограничены определёнными языковыми парами, то современные нейронные модели позволяют выполнять перевод между десятками и сотнями языков. В связи с этим особое внимание привлекают современные модели mT5, Gemma и NLLB. Эти модели способны не только выполнять перевод, но и решать задачи понимания, обработки и генерации многоязычных текстов. Поэтому сравнительное рассмотрение их архитектуры, принципов работы и областей применения представляет научный и практический интерес.

Кроме того, для оценки качества результатов распознавания речи и машинного перевода используются специальные метрики. Среди них наиболее распространёнными являются BLEU, chrF, WER и TER. Эти метрики позволяют сравнивать результат работы системы с эталонным текстом, количественно оценивать число ошибок, степень сходства и качество перевода или распознавания. Правильное понимание метрик обеспечивает объективность исследования и помогает чётко показать преимущества и ограничения моделей.

Наряду с этим в последние годы актуальным направлением исследований стала автоматизация стенограмм совещаний. Такие системы не только преобразуют речь совещания в текст, но и автоматически выделяют ключевые идеи, принятые решения, поставленные задачи и выводы. Это предоставляет значительные возможности для экономии времени в организациях, упрощения документооборота и систематизации информации. Однако для казахского языка исследования в данном направлении всё ещё ограничены, поэтому важно обобщить существующие работы и провести сравнительный анализ.

Целью данной работы является рассмотрение моделей Whisper, Turkic ASR и MMS для распознавания речи, а также моделей mT5, Gemma и NLLB для машинного перевода, описание их математических моделей, обзор научных исследований, а также объяснение метрик BLEU, chrF, WER и TER. Кроме того, предполагается анализ современных исследований по автоматизации стенограмм совещаний и определение перспектив развития данного направления для казахского языка.

Цель диссертационной работы

Разработка модели интеллектуальной системы автоматического формирования стенограмм совещаний для кыргызско-казахской языковой пары и исследование её эффективности.

Задачи исследования

Для достижения поставленной цели определены следующие задачи:

- исследование модели автоматического распознавания речи (ASR) для преобразования кыргызской речи в текст стенограмм совещаний;
- разработка метода Morphology-Aware ASR Post-Processing на основе CSE;
- исследование нейронной модели машинного перевода с кыргызского языка на казахский язык;
- разработка интегрированной системы ASR → MT для автоматизированного формирования стенограмм совещаний;
- проведение экспериментального исследования качества разработанных моделей.

Объект исследования

Процесс обработки кыргызской речи в казахский текст.

Предмет исследования

Методы и модели распознавания речи и машинного перевода в многоязычной среде.

Научная новизна полученных результатов

- разработан объемный параллельный корпус для кыргызско-казахской языковой пары, содержащий 782 293 параллельных предложений;
- предложен метод Morphology-Aware Post-Processing на основе CSE технологии для коррекции результатов распознавания кыргызской речи;
- разработана система автоматического формирования стенограмм совещаний для кыргызско-казахской языковой пары.

Положения, выносимые на защиту

- разработка объемного параллельного корпуса из 782 293 параллельных предложений для кыргызско-казахской языковой пары;
- метод Morphology Aware Post Processing на основе CSE технологии для исправления ошибок результатов распознавания кыргызской речи;
- разработка системы автоматического формирования стенограмм совещаний для кыргызско-казахской языковой пары.

Теоретическая и практическая значимость работы

- позволяет автоматизировать формирование стенограмм совещаний в государственных органах, образовательных учреждениях и бизнес-структурах;
- вносит вклад в развитие технологий обработки речи для кыргызского и казахского языков;
- предлагает универсальное решение, применимое при создании многоязычных интеллектуальных систем;
- служит основой для разработки эффективных программных средств для языков с ограниченными ресурсами.

Связь темы с научно-исследовательскими программами

Диссертационная работа выполнена в рамках грантового проекта AP23487816 «Исследование нейронных моделей формирования стенограммы речи и протоколов совещаний на тюркских языках» на 2024–2026 годы.

Объём и структура диссертационной работы

Диссертация состоит из введения, трёх глав, заключения и списка литературы.

В первой главе рассматриваются теоретические основы систем автоматического распознавания речи (ASR) и машинного перевода для кыргызско-казахской языковой пары. Проведён анализ современных научных исследований в области ASR-технологий для кыргызского и казахского языков, методов нейронного машинного перевода, а также систем автоматического формирования стенограмм совещаний.

Во второй главе представлена методология разработки системы автоматического формирования стенограмм для кыргызско-казахской языковой пары. Рассматриваются клиент-серверная архитектура системы, методы сбора и обработки данных, этапы исследования моделей распознавания речи, метод Morphology Aware Post Processing, разработка моделей машинного перевода, а также модели, используемые для формирования стенограмм.

В третьей главе представлены экспериментальные результаты и их анализ. Проведена оценка качества моделей ASR, проанализирована эффективность метода Morphology Aware Post Processing. Также представлены результаты перевода с кыргызского языка на казахский и проведено сравнение предложенной системы с существующими методами. Рассмотрены особенности практического применения системы, её ограничения и направления дальнейшего совершенствования. Обоснована эффективность предложенного подхода для низкоресурсной кыргызско-казахской языковой пары и описаны перспективы дальнейшего развития системы автоматического формирования стенограмм.

Публикации

По теме диссертации опубликовано 6 научных работ, из них 3 статьи — в рецензируемых журналах, входящих в международные базы цитирования («Web of Science» и/или «Scopus»), 1 работа — в материалах международных конференций.

Также по результатам исследования зарегистрирован объект авторского права — «Казахско–кыргызский параллельный корпус текстов KazKyrParC» — внесённый в Государственный реестр объектов, охраняемых авторским правом Республики Казахстан. Свидетельство №62995 от 14 октября 2025 года

- U. Tukeyev, A. Shormakova, A. Karibayeva, D. Rakhimova, B. Abduali, D. Amirova, N. Rakhmanberdi, R. Aliyev. An Integrated Approach to Adapting Open-Source AI Models for Machine Translation of Low-Resource Turkic Languages 2026 Computers 2026, MDPI, 15(2), 73. (Процентиль 84, Q1) <https://doi.org/10.3390/computers15020073>
- A. Karibayeva, V. Karyukin, B. Abduali, D. Amirova, Speech Recognition and Synthesis Models and Platforms for the Kazakh Language 2025 Information 2025, MDPI, 16(10), 879; (Процентиль 74, Q2) <https://doi.org/10.3390/info16100879>
- A. Karibayeva, V. Karyukin, U. Tukeyev, B. Abduali, D. Amirova, D. Rakhimova, R. Aliyev, A. Shormakova. The Development and Experimental Evaluation of a Multilingual Speech Corpus for Low-Resource Turkic Languages 2025 Applied Science. MDPI, 2025, 15(24), 12880; (Процентиль 79, Q2) <https://doi.org/10.3390/app152412880>
- Abduali B., Milosz, M., Tukeyev, U., Karibayeva, A. Methodology for creating datasets of parallel sentences in low-resource languages by using AI. International Journal of Innovative Research and Scientific Studies, 8(9), 13–23. 2025 <https://doi.org/10.53894/ijirss.v8i9.10605>
- А.Кәрібаева, Б. Абдуали, Д.Амирова. Анализ технологии обратного перевода в англо-казахском нейронном машинном переводе, 31.10-1.11.2024, IX Международная научно-практическая конференция «Информатика и прикладная математика»
- Abduali B., Tukeyev U., Zhumanov Zh., Israilova N. Study of Kyrgyz-Kazakh Neural Machine Translation 2025 17th Asian Conference on Intelligent Information and Database Systems, proceedings conference paper (Scopus) Doi: 10.1007/978-981-96-5881-7_21.
- Авторское свидетельство «Казахско-Кыргызский параллельный корпус текстов «KazKyrParC»» № №62995. 14.10.2025