МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РЕСПУБЛИКИ ТАДЖИКИСТАН
ПОЛИТЕХНИЧЕСКИЙ ИНСТИТУТ
ТАДЖИКСКОГО ТЕХНИЧЕСКОГО УНИВЕРСИТЕТА
ИМЕНИ АКАДЕМИКА М.С. ОСИМИ В ГОРОДЕ ХУДЖАНДЕ

УДК: 81.33 + 004.42

На правах рукописи

ХУДОЙБЕРДИЕВ ХУРШЕД АТОХОНОВИЧ

ПРОЕКТИРОВАНИЕ И РЕАЛИЗАЦИЯ АВТОМАТИЧЕСКИХ СИСТЕМ ОБРАБОТКИ ИНФОРМАЦИИ НА ТАДЖИКСКОМ ЯЗЫКЕ

АВТОРЕФЕРАТ

диссертации на соискание ученой степени доктора технических наук
по специальности 05.13.11 – Математическое и программное обеспечение
вычислительных машин, комплексов и компьютерных сетей

ДУШАНБЕ - 2024

Работа выполнена на кафедре программирования и информационных систем Политехнического института Таджикского технического университета имени академика М.С. Осими в городе Худжанде

Научный консультант: Усманов Зафар Джураевич,
доктор физико-математических наук, профессор,
Академик НАНТ

Официальные оппоненты: Илолов Мамадшо Илолович,
доктор физико-математических наук, профессор,
академик НАНТ, заведующий отделом математического моделирования динамических процессов центра инновационного развития науки и новых технологий

Пруцков Александр Викторович,
доктор технических наук, доцент, профессор кафедры «Вычислительная и прикладная математика», Федеральное государственное бюджетное образовательное учреждение высшего образования «Рязанский государственный радиотехнический университет»

Бекназарова Саида Сафибуллаевна,
доктор технических наук, профессор кафедры «Телевизионные и медиа технологии», Ташкентский университет информационных технологий имени Мухаммада ал-Хоразмий

Ведущая организация: Национальный Университет Таджикистана

Защита состоится «13» сентября 2024 года в 14.00 часов на заседании разового диссертационного совета 6D.KOA-049 при Таджикском техническом университете имени академика М.С. Осими по адресу: 734042, г. Душанбе, проспект академиков Раджабовых, 10А

С диссертацией можно ознакомиться в научной библиотеке Таджикского национального университета и на официальном сайте университета https://web.ttu.tj/tj/elonho/77

Автореферат разослан «____» _____ 2024 года

Отзывы на автореферат в двух экземплярах, подписанные и заверенные печатью учреждения, просим направлять по адресу: 734042, г. Душанбе, проспект академиков Раджабовых, 10А, тел.: (+992 37) 227-37-81,
e-mail: sultonzoda.sh@mail.ru

Учёный секретарь диссертационного совета,
кандидат технических наук, доцент Султонзода Ш.С.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Одной из актуальных проблем в области компьютерной лингвистики является разработка системы автоматической проверки правописания и ее редактирования на основе правил определенного языка, пакетов автоматического синтеза и распознавания речи, модуля голосового управления оконечным устройством, а также систем автоматического машинного перевода.

Системы автоматической обработки текста на естественном языке работают посредством комплекса программ и компьютерных приложений, работа которых базирована на математических моделях. Разработка системы автоматической проверки орфографии и ее редактирования на базе правил определенного языка, пакетов синтеза и определения речи, модулей голосового управления конечных автоматических устройств, также систем автоматического машинного перевода считаются важными задачами в области компьютерной лингвистики.

Разработкой современных вопросов математического моделирования компьютерной лингвистики и проектирования систем обработки естественного языка занимались такие зарубежные ученые, как Indurkhya N., Damerau F.J., Grishman R., Hutchins W.J., Hausser R.R., Cohen M., Massaro D., Liberman A.M., Black A.W., Taylor P.A., Johnson M., Nirenburg S., Somers H.L., Wilks Y., Koehn P., Mercer R.L., Schroeder M., Zen H. и другие.

В работах российских ученых Е.И. Большакова, Е.С. Клишинского, Д.В. Ланде, А.А. Носкова, О.В. Песковой, Е.В. Ягуновой, Г.Г. Белоногова, А.В. Палагина подробно рассмотрены вопросы, связанные с автоматической обработкой текстовой информации. В исследованиях этих ученых предлагается совершенно новая возможность развития перспективных систем, связанных с автоматической обработкой текстов.

Создание методологии, методов и базовых моделей разработки систем автоматической обработки текстовой информации имеет давнюю историю. В трудах таких ученых, как Д.Ш.Сулейманов, В.А.Фомичев, А.В.Анисимов, Т.В.Батура, Ф.А.Мурзин, О.Ф.Кривнова, С.В.Лесников, А.А.Марченко, Р.К.Потапова, С.Б.Потемкин, Г.Е.Кедрова, Н.Н.Сажок, А.И.Солонина, В.Н.Сорокин, Л.А.Чистович представлены способы разработки основных принципов, композиционная структура, технология разработки практических лингвистических моделей, которые в дальнейшем нашли свое применение в информационных системах обработки текстов на естественном языке.

Согласно данным, значительное количество пользователей компьютерных средств используют более совершенные системы обработки информации на естественном языке и программные продукты, в том числе электронные словари WordNet, MS Office, ABBYY, Open Office, PROMPT и OXFORD, системы перевода YANDEX и GOOGLE, работающие как в режиме онлайн, так и в режиме офлайн. Некоторые из перечисленных систем обладают возможностью создания многоязычного словаря, отражающего все возможные толкования и толкования слов определенного языка путем установления отношений между ними.

Широкое использование информационно-коммуникативных технологий в

Таджикистане вызвал большой интерес у исследователей в области математики, информационных технологий и лингвистики. Ученые под руководством академика НАНТ З.Д. Усманова обратились к совершенно новой отрасли информационно-коммуникативных технологий – компьютерной лингвистике. Проблема разработки нового направления - компьютерной лингвистики - поставила перед исследователями необходимость решения ряда важных задач. В частности, это задачи, связанные с моделированием простого двусоставного предложения (С.А. Зарипов), разработкой национальных драйверов таджикской графики и решением проблемы стандартизации печатной продукции (О.М. Солиев), преобразованием графических систем линий (Л.А. Гращенко), автоматическим морфологическим анализом (Г.М. Довудов), распознаванием автора таджикских текстов (А.А. Косимов и К.С. Бахтеев), системой автоматической обработки текста на шугнанском языке (А.Г. Гуломсафдаров).

Одной из фундаментальных задач, стоящих перед каждой страной, является четкое осознание своего места в продолжающемся процессе глобализации. Народу страны, учитывая сущность поведения современных государств мира, предстоит сделать выбор: удовлетворяться скромной ролью потребителя продуктов современного культурного, научно-технического развития других народов или предпринять активные действия, чтобы донести до всего мирового сообщества свои национальные ценности и мировоззрение. Особенно это актуально для стран, находящихся на стадии развития в условиях современного технологического процесса.

Актуальность научного исследования подтверждена Государственной стратегией «Информационно-коммуникационные технологии для развития Республики Таджикистан», Государственной программой развития государственного языка на 2020-2030 годы, Указом Президента Республики Таджикистан об объявлении 2020-2040 годы «Двадцатилетием изучения и развития естественных, точных и математических наук в сфере науки и образования», Стратегией изучения и развития математических, точных и естественных наук в сфере образования и науки на период до 2030 года, Целевой государственной программой развития математических, точных и естественных наук на 2021-2025 годы.

Цель исследования – разработка моделей, методов и алгоритмов, позволяющих создавать информационные системы автоматической обработки информации на таджикском языке для их дальнейшего использования в человеко-машинных системах управления в естественно-языковом диалоге.

Задачи исследования. Для достижения поставленной цели в рамках диссертационной работы были поставлены следующие задачи:

- разработка методологии и теоретической концепции автоматической обработки текстовой информации на таджикском языке как объекта научного исследования для определения понятий и теоретических терминов в компьютерной лингвистике;
- разработка методов поиска текстовой информации для анализа экспериментальных данных и их применения в научно-практических исследованиях, электронных словарях и компьютерных тезаурусах на таджикском языке;

- разработка модели предоставления текстовой информации и комплекса алгоритмов реализации автоматического синтеза речи на таджикском языке;

- разработка методов извлечения, представления и обработки данных с целью формирования отдельных элементов текста для реализации автоматической проверки правописания текста на таджикском языке;

- разработка моделей, методов и алгоритмов предварительной обработки данных для решения задачи автоматического перевода текста с таджикского языка на русский язык;

- разработка программного комплекса для реализации всех методов, моделей и алгоритмов обработки информации на таджикском языке;

- проведение экспериментального исследования эффективности систем автоматической обработки информации.

Объектом исследования является компьютерное моделирование вычислительных процессов и проектирования программных обеспечений для системы автоматической обработки информации на таджикском языке.

Предмет исследования – методы, модели и алгоритмы обработки информации на таджикском языке для проектирования и реализации электронных словарей, синтеза речи, автоматической проверки орфографии и компьютерного перевода.

Область исследований – разработка моделей, обоснование и тестирование эффективных численных методов с использованием ЭВМ; применение эффективных численных методов и алгоритмов в виде наборов проблемно-ориентированных программ для проведения вычислительных экспериментов; многоплановое исследование научно-технических проблем с использованием современных технологий математического моделирования и вычислительного тестирования.

Достоверность и обоснованность результатов, полученных в диссертационной работе, обоснована предложенными математическими моделями элементов текстовой информации с целью последующей обработки. В свою очередь достоверность спроектированных автоматических систем и компьютерных модулей подтверждена корректным выбором исходных данных и репрезентативной выборки текстовой информации в формировании постановки задач разработки математических и компьютерных средств и их реализации в информационных системах для автоматической обработки информации на таджикском языке. В диссертационной работе использованы результаты, полученные ранее другими авторами и отмеченные ссылками.

Методы исследования. Для решения задач, стоящих перед исследованием, были использованы методы систематического анализа, математической статистики, основы представления и обработки наборов данных, а также теория алгоритмов, математическая и компьютерная лингвистика, синтез данных, компьютерное моделирование автоматических информационных систем, технологии программирования и обработки данных.

Научная новизна исследования. В результате научно-исследовательской работы и разработки автоматических систем предложен ряд методических подходов к исследованию, анализу и автоматической обработке текстовой информации на таджикском языке:

- предложены новые научно-технические положения, математические модели, методы и структуры данных, которые в целом составляют теоретическую основу системного анализа и исследования текстовой информации;

- впервые разработаны методы и алгоритмы практического, структурного и объектно-ориентированного проектирования систем автоматической обработки данных;

- предложены новые методы создания программных средств автоматического синтеза речи на таджикском языке, система автоматической проверки орфографии TajSpell в программном пакете Microsoft Office; программные модули автоматического перевода текста с таджикского языка на русский и английские языки в виде интернет-приложения, доступного по адресу tarjumon.tajlingvo.tj.

- на основе разработанных методов, моделей и структур данных предложены новые алгоритмы машинного перевода, сформированы компьютерные параллельные корпусы Tajik-Russian-Parallel Corpus и Tajik-English-Parallel Corpus в виде веб-приложений, а также программные модули автоматического перевода текста с таджикского языка на русский и английский языки;

- разработаны новые модели, методы синтеза речи и компьютерные программы Computer Tajik Text Narrator, Tajik Text-to-Speech, повышающие эффективность практического использования ИКТ для решения актуальных лингвистических задач и речевых технологий в таджикском языке.

Все полученные результаты реализованы в программном комплексе TajLINGVO, который позволяет:

- существенно сократить время изучения таджикского языка как для пользователей в Республике Таджикистан, так и за рубежом;

- повысить уровень обоснованности принимаемых решений по компьютерной лингвистике и задачам таджикского языка;

- обеспечить формирование и использование корректного контента на таджикском языке в сети Интернет.

Теоретическая значимость исследования заключается в том, что в нем представлены примеры, методы и алгоритмы обработки элементов текста и звукового сигнала на естественном языке, способствующие изучению таджикского языка.

На основе в процессе проведения исследований и полученных данных написаны учебные книги под грифом Министерство образования и науки Республики Таджикистан по дисциплинам «Проектирование информационных систем», «Базы данных», «Практикум по программированию», «Задачи для изучения программирования» используемые при обучении бакалавров по направлению программное обеспечение информационных технологий.

Практическая значимость исследования. За последние годы были апробированы, усовершенствованы и внедрены автоматические системы и новые приложения в программном комплексе TajLINGVO. Практическое значение и значимость основных положений исследования подтверждает опыт создания программных средств для реализации электронных словарей,

электронного тезауруса, автоматического синтеза речи, проверки орфографии, автоматического перевода. Основные результаты исследований прошли опытную эксплуатацию в Худжандском научном центре НАНТ, в Управлении по инвестициям и управлению государственным имуществом Согдийской области, внедрены в учебном процессе ГОУ Худжандского государственного университета имени академика Б.Гафурова, в кафедре таджикского языка Таджикского государственного университета права, бизнеса и политики, в Политехническом институте Таджикского технического университета имени академика М.С. Осими в городе Худжанде, а также комплекс программы TajSpell внедрен в процессе документации в ЗАО «Душанбе Сити Банк». Полученные результаты и накопленный опыт разработки автоматических систем не только существенно сокращают время изучения таджикского языка для отечественных пользователей компьютерной техники при решении задач синтеза, правописания и перевода речи, но и обеспечивают иностранным пользователям методическую основу для изучения таджикского языка.

Модели, алгоритмы и программное обеспечение, разработанные в рамках диссертационного исследования, позволяют использовать таджикский контент для исследования и повседневного практического использования.

Положения, выносимые на защиту.

1. Представлена концепция автоматической обработки текстовой информации на таджикском языке как объект научного исследования и программные средства для систематического анализа, на основе которых определены понятия и теоретические термины.

2. Представлен и экспериментально проверен научно-практический подход к разработке электронных словарей и компьютерных тезаурусов, в рамках которого сформированы примеры решения поисковых задач и методы применения этих примеров в процессе реализации компьютерных словарей.

3. Впервые предложен подход автоматического синтеза речи на таджикском языке, основанный на использовании метода конкатенации слогов. При этом получены всевозможные структуры слогов и слоговых структур слов таджикского языка. Разработано программное обеспечение автоматического синтеза речи на основе собственных алгоритмов и базы данных «слог-звук». В результате синтеза речи формируется целая звуковая дорожка на основе предложенной текстовой информации в формате цифрового звукового файла.

4. Разработаны новые методы извлечения, представления и обработки данных, составляющих отдельные элементы текста, а также предложен новый способ решения проблемы автоматического правописания текста на таджикском языке.

5. Впервые изучен вопрос автоматического перевода текста с таджикского языка на русский, разработаны модели, методы и алгоритмы, которые дают возможность эффективно решать практические задачи.

6. Исследовано совместное использование системного анализа, структурированного подхода к обработке данных, объектно-ориентированного программирования и компьютерной лингвистики для разработки систем автоматической обработки текстовой информации на таджикском языке.

7. Для реализации всех представленных автоматических систем

разработан комплекс компьютерных программ TajLINGVO, проведена его экспериментальная апробация на территории Республики Таджикистан.

Все результаты и положения диссертации, представленные на защиту, получены автором или при его непосредственном участии, являются новыми и полностью доступны в открытой печати. Программный комплекс TajLINGVO зарегистрирован в Национальном патентном центре Министерства экономического развития и торговли Республики Таджикистан, а разработанные проекты автора не имеют аналогов. Их перечень доступен на сайте www.tajlingvo.tj.

Соответствие диссертации паспорту научной специальности. Диссертация выполнена по специальности 05.13.11 - «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей». В исследовании имеются совершенно уникальные результаты, относящиеся к таким направлениям, как математическое моделирование, численные методы и программные комплексы, соответствующие пунктам 1 - модели, методы и алгоритмы проектирования и анализа программ и программных систем, их эквивалентных преобразований, верификации и тестирования; 3 - модели, методы, алгоритмы, языки и программные инструменты для организации взаимодействия программ и программных систем; 4 - системы управления базами данных и знаний; 5 - программные системы символьных вычислений; 7 - человеко-машинные интерфейсы; модели, методы, алгоритмы и программные средства машинной графики, визуализации, обработки изображений, систем виртуальной реальности, мультимедийного общения паспорта специальности.

Личный вклад соискателя состоит в том, что диссертационная работа выполнена им самостоятельно, в диссертации постановка задачи их реализация, методы, модели и алгоритмы обработки информации, описанные на таджикском языке и представленные на защиту, подготовлены самостоятельно или под его непосредственном руководстве.

Уровень достоверности результатов подтверждена соответствующими актами о практическом использовании и внедрении разработанных информационных систем, документами о выдаче государственного регистрационного номера интеллектуальной продукции и информационных ресурсов в Национальном патентно-информационном центре Минэкономразвития и торговли Республики Таджикистан. Достоверность результатов подтверждается также признанием заслуг автора в данной области науки со стороны различных организаций и учреждений республики. В частности, это премия имени академика С.У. Умарова в области физико-математических, химических, геологических и технических наук Национальной академии наук РТ, 2015 г.; Госпремия для учёных и преподавателей естественных, точных и математических дисциплин, 2021 г.; диплом третьей степени республиканского конкурса «Наука – цвет процветания», номинация инновация и нововведение, 2021 г.; Почетная грамота и медаль «100 НОВЫХ ЛИЦ» стран Содружества Независимых Государств, 2022 г.

Апробация и внедрение результатов диссертации. Основные результаты диссертации докладывались на научных семинарах ХПИТТУ им.

академика М.С. Осими, а также на республиканских, международных конференциях и семинарах: международная конференция «Современные вопросы математики», посвященная 50-летию Института математики имени А. Джураева НАНТ, (26-27 мая 2023г.), г. Душанбе; всероссийская научно-практическая конференция с участием международных представителей по теме «Обмен информацией в междисциплинарных исследованиях II», (14 апреля 2023г.), Академия права и управления ФСИН России, г. Рязань, Российская Федерация; международная научно-практическая конференция «Новые достижения в области естественных наук и информационных технологий», Российско-таджикский славянский университет, (30 мая 2023г.), г. Душанбе; республиканская научно-практическая конференция, посвященная международному дню родного языка на тему «Родной язык – источник самопознания и национальной духовности», Комитет языка и терминологии при Правительстве Республики Таджикистан, (16 февраля 2023), г. Душанбе; республиканская научно-практическая конференция на тему «Применение информационно-коммуникационных технологий в индустриализации страны», Таджикский технический университет имени академика М.С. Осими (29 октября 2022 г.), г. Душанбе; республиканская конференция «Практические информационные системы: проблемы моделирования, внедрения в развивающихся странах», ХПИТТУ имени академика М.С. Осими, (2012г., 2017г., 2022г.), г. Худжанд; международная научно-практическая конференция «Наука и технологии», (26 сентября 2022 г.), г. Алматы, Республика Казахстан; республиканская научно-практическая конференция «Актуальные проблемы языкознания и лингводидактики в современных условиях», филиал МГУ имени М.В. Ломоносова в городе Душанбе, (29 октября 2022 г.), г. Душанбе; научно-практическая республиканская конференция «Актуальные вопросы перевода и лингвистики в современности», институт языков Таджикистана имени Сотима Улугзаде, (2019 г.), г. Душанбе; ежегодный научно-практический семинар «Новые информационные технологии в автоматических системах», Институт прикладной математики им. М.В. Келдыша РАН, (2013-2019 гг.), г. Москва, Российская Федерация; научно-практическая конференция преподавателей, молодых исследователей, посвященная 30-летию Государственной Независимости Республики Таджикистан, ХПИТТУ имени академика М.С. Осими, (2019г.), г. Худжанд; региональная научно-практическая конференция, посвященная 90-летию Темурхана Максудова, Филиал Технологического университета Таджикистана в городе Исфаре, (2018г.), г. Исфара; научно-практическая конференция «Применение информационно-коммуникационных технологий для инновационного развития Республики Таджикистан», Технологический университет Таджикистана, (2017г.), г. Душанбе; республиканская научно-практическая конференция на тему «Качество образования в высших учебных заведениях Республики Таджикистан», посвященная 25-летию Независимости Республики Таджикистан, ХПИТТУ имени академика М.С. Осими, (20 сентября 2016 г.), г.; третья международная научно-техническая конференция «Открытые семантические технологии проектирования интеллектуальных систем», Белорусский государственный университет информатики и радиоэлектроники – OSTIS-2013, (21-23 февраля

2013 г.), г. Минск, Республика Беларусь.

Публикации по теме диссертации. По материалам диссертационного исследования опубликовано 68 работ, в том числе 25 (11 без соавторства) из которых опубликованы в журналах, рекомендованных ВАК при Президенте РТ и ВАК РФ, 27 статей в международных сборниках статей и журналов, 7 учебных пособий под грифом Министерства образования и науки Республики Таджикистан. В патентно-информационном центре при Министерстве экономического развития и торговли Республики Таджикистан получено 18 свидетельств о государственной регистрации информационных ресурсов и интеллектуальной продукции.

Структура и объем диссертации. Диссертационное исследование состоит из 328 компьютерных страниц, введения, 6 глав, 19 таблиц, 15 рисунков, библиографии с 322 названиями и 2 приложений.

Признательность. Автор выражает искреннюю благодарность своему научному руководителю (консультанту), академику НАНТ, доктору физико-математических наук, профессору Усманову Зафару Джураевичу за полезные советы и добрые наставления при подготовке данной научной работы.

Ключевые слова: таджикский язык, синтез речи, элементы текста, частота встречаемости, электронный словарь, компьютерный тезаурус, компьютерная лингвистика, автоматическая обработка текста, автоматическая проверка орфографии, транслитерация, машинный перевод, статистический метод, классификация, кластеризация, математическая статистика, теория вероятности, численные методы, математическое моделирование, проектирование информационных системы, база данных, компьютерное моделирование, технология программирования.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснованы актуальность диссертационной работы, ее цель и основные задачи, показана научная и практическая значимость исследования. Кратко изложено содержание диссертации.

В первой главе «Компьютерная лингвистика таджикского языка» рассматриваются вопросы, связанные с определением главных особенностей проектирования и проблемами реализации информационных систем обработки информации на таджикском языке.

В **разделе 1.1** представлены общие сведения о широком использовании средств компьютерных технологий на предприятиях и в учреждениях Таджикистана, о проблемах использования таджикского языка в процессе делопроизводства на базе современных возможностей ИКТ и возникновении интереса ученых к этой отрасли науки, о создании научной школы компьютерной и математической лингвистики академиком НАНТ, профессором З.Д. Усмановым, о результатах работы молодых талантливых таджикистанских исследователей этой школы.

В **разделе 1.2** приведено описание результатов исследования по компьютерной лингвистике таджикского языка в Республике Таджикистане и совместные достижения таджикских ученых в области математики,

информационных технологий и лингвистики.

В **разделе 1.3** определена математическая модель, на базе которой спроектированы информационные системы автоматической обработки данных на таджикском языке.

Проектируемая модель системы TajLINGVO состоит из набора взаимосвязанных информационных технологий, процессов, алгоритмов, набора текстовых элементов, интерфейсов и набора результатов, необходимых для формирования цифрового изображения. Их можно описать следующим образом:

$$TajLINGVO = \{T, P, A, TE, I, R\} \quad (1)$$

где,

T – совокупность информационных технологий;

P – совокупность процессов в TajLINGVO, P_i, i=1…n;

A – набор алгоритмов Aj, j=1...m для реализации процессов {Pi};

TE – совокупность элементов текстовой информации, которые передаются на обработку с помощью алгоритмов {Aj} в процессах {Pi};

I – пользовательские интерфейсы для ввода, обработки и удаления данных;

R – результаты для передачи на обработку в процессах {Pi}.

При разработке логической структуры информационных систем в ее основу кладется определенная методология программного обеспечения. Этому способствуют современные методы и инструменты, которые дают возможность разработчикам моделировать системы с начала до конца. К такому инструменту, например, можно отнести Structured Analysis and Design Technique (SADT) – технология структурированного анализа и проектирования, инженерная методология разработки и идентификации систем в форме возрастающей стратификации подсистем.

Структура системы TajLINGVO, предложенная по методологии SADT, состоит из четырех подсистем и представляет собой набор информационных ресурсов, алгоритмов и программных средств, управляющих процессами АОТ и пользовательскими интерфейсами. Подсистемы совместно реализуют набор алгоритмов автоматической обработки предоставленных исходных данных. Результаты обработки формируют набор текстовых элементов по семантическим структурам, которые записываются в источник данных и вставляются в пользовательский интерфейс.

Подсистема «*Обеспечение информационным ресурсом*» обеспечивает формирование лингвистического ресурса текстов на основе репрезентативной выборки с учетом данных лингвистических и текстовых структур. Подсистема состоит из следующих компонентов: источников текстовой информации, различных источников данных, например, электронных словарей, заданной структуры текстовых элементов, являющихся результатом реализации определенного процесса АОТ.

Подсистема «*Алгоритмы и программные средства*» представляет собой набор алгоритмов, примененных в виде программных модулей, задач и процедур обработки структуры текстовых элементов. Программные инструменты позволяют пользователю управлять процессом АОТ.

Подсистема «*Управление процессами АОТ*» представляет собой предварительную подготовку результатов обработки входных данных. Существуют также процедуры мониторинга и проверки результатов для принятия решения пользователем. В случае получения результатов разных значений предлагается возможность повторной обработки данных.

Подсистема «*Пользовательские интерфейсы*» предоставляет возможность поиска, представления и выбора данных, записи результатов в источник данных. Также для всестороннего просмотра результатов пользователю предлагается возможность получать графические версии отчетов в виде таблиц, точек, диаграмм и гистограмм.

Для разработки модели компьютерной системы TajLINGVO необходимо на основе полученной логической структуры создание модели системы, модели информационных процессов P и программных средств, реализующих набор A-алгоритмов. Полученные данные, в зависимости от достоверности результатов, могут быть переданы на автоматическую обработку текстовых элементов, такую как разработка компьютерных синонимов, проверка орфографии, синтез речи и машинный перевод.

Структура современных информационных систем на естественном языке состоит из большого количества текстовых элементов и образует концептуальную модель базы знаний. Чтобы достичь структуры, необходимо опираться как на традиционную модель естественного языка, так и на современные методы структурированных текстовых моделей. Далее приведена математическая модель информационной структуры:

$$FM = \{LC, SW, SS, DS, GS, CS\} \quad (2)$$

где,

LC – источник текстовой информации для формирования языковых ресурсов;

SW – набор составленных слов, образованных от LT;

SS – совокупность семантических структур, описывающих SW;

DS – набор лингвистических структур, SS образовались в SW;

GS – комплекс грамматических явлений, основанных на грамматических правилах естественного языка;

CS – набор структур кода для представления DS согласно GS.

Процессы поиска, обработки, анализа и понимания текстовых элементов реализуют последовательность преобразования текстовой информации WS→CS. Предлагается схема относительно доступного анализа информационной модели системы TajLINGVO, в которой процессы обработки текстовой информации реализуются с помощью программного обеспечения, рисунок 1.

Проанализируем другие функции, используемые в информационной модели системы TajLINGVO:

P1 – создание репрезентативных примеров на основе текстовых документов (классические и современные произведения);

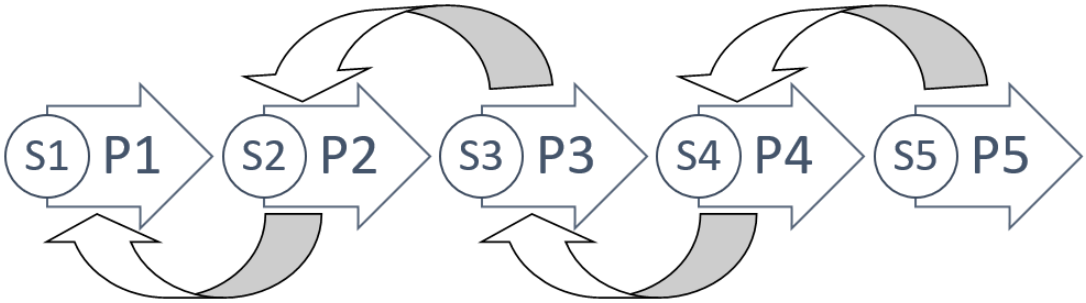


Рисунок 1. Структура информационной модели TajLINGVO

P2 – предварительная обработка текстовых документов для автоматического языкового анализа; предлагается вернуться к процессу P1 в результате определения проблемы омонимов;

P3 – процесс выбора набора элементов текстовой информации путем определения их структуры и записи в текстовую информацию. В случае нахождения нескольких значений семантической структуры текстового элемента можно вернуться к операции P2;

P4 – процесс формирования структуры элементов текста на основе правил орфографии языка; в результате определения несоответствия определяемых структур правилам естественного языка возможен возврат к процессу P3;

P5 – процесс обработки и управления данными; в результате определения неопределенного цифрового изображения текста можно вернуться к процессу P4;

S1 – источники текстовых документов;

S2 – хранилище текста;

S3 – смысловая структура текстовых элементов в соответствии с правилами грамматики естественного языка;

S4 – набор информационных структур после обработки текста;

S5 – это источник данных и цифровое отображение текстового информационного элемента для создания базы знаний.

Во второй главе «Методология компьютерного анализа и синтез естественного языка» дается описание основных методов обработки информации на естественном языке, а также решения задач математического моделирования и методов анализа и синтеза информации на таджикском языке.

В **разделе 2.1** рассмотрены вопросы обработки естественного языка, определены основные задачи анализа текста, системы языковых средств, дана разработка структуры речевых служб, исследованы лингвистические средства и теория информации; проанализированы лингвистические правила в структурной части текста в зависимости от языка, жанра и объема текстовой информации. Обосновано практическое использование методов анализа текста, таких как графемный, лексический, морфологический, синтаксический и семантический. Созданы классификации средств и методов автоматической обработки информации в формате текстовых данных.

В **разделе 2.2** в целях изучения проблем обработки текстовой информации на таджикском языке исследованы математические методы З.Д. Усманова, а

также созданы общие способы построения и кодирования текстовых элементов, таких, как слоговая структура слов, кодирование слов и предложений. На базе математических методов обработки информации исследованы статистические закономерности некоторых элементов текста: слогов, слов, анаграмм, предложений. С целью реализации синтеза речи на таджикском языке сформированы слоговые структуры слов. В данной части работы определены математические методы кодирования текстовых элементов для решения задач автоматической проверки правописания, машинного перевода текста и голосового синтеза на таджикском языке.

В **разделе 2.3** на примере математических моделей изучены методы проверки правописания в текстовых данных, в основе которых лежат два типа орфографических ошибок: когнитивные и типографские. Когнитивные ошибки – это ошибки, которые возникают, когда неизвестно правильное написание слова. При этом неправильное произношение написанного слова такое же или похожее на произношение правильного слова, например, «*кумак*» вместо «*кӯмак*». Опечатки, связанные с когнитивными ошибками, составляют около 80%. Анализируя природу получения опечаток, можно выделить четыре относительно распространенные группы. Например, для слова «*истиқлол*» возможны следующие варианты: вставка одной лишней буквы: «*исстиклол*» (ошибка x1), пропуск одной буквы: «*итиклол*» (ошибка x2), замена одной буквы другой: «*истиклол*» (ошибка x3), смещение двух соседних букв: «*итсиқлол»* (ошибка x4).

Для выполнения задачи по исправлению первых трех типов ошибок в слове при вводе текста широко используется расстояние Левенштейна. С помощью этого метода определяем математическую формулу для расчета расстояния между двумя строками: w1 – правильно написанное слово длины N и w2 – слово в неправильной форме длины M, с наименьшим количеством операций вставки (x1), удаление (x2), замена (x3) одной буквы. Тогда расстояние редактирования, то есть расстояние Левенштейна D(w1,w2), можно рассчитать по следующей формуле D(w1,w2) = D(N,M), где:

$$D(i,j) = f(x) = \begin{cases} 0, \ i = 0, j = 0 \\ i, \ j = 0, i > 0 \\ j, \ i = 0, j > 0 \\ min \begin{Bmatrix} D(i,j-1) + 1, \\ D(i-1,j) + 1, \\ D(i-1,j-1) + m(w1[i], w2[j]) \end{Bmatrix}, \ J > 0, i > 0 \end{cases} \quad (3)$$

где, шаг по i – код вероятности написания буквы с ошибкой из слова (x2), шаг по j – вставка одной буквы в слово (x1), шаг по обоим индексам символ замены буквы в слове на другую неправильную букву (x3).

В **разделе 2.4** приводятся алгоритмы и методы машинного перевода на основе методологий автоматического перевода, описываются разработанные модели бинарного перевода, стратегическое проектирование межъязыкового подхода, статистического подхода, машинного обучения и нейронных сетей.

Алгоритм машинного перевода, основанный на правилах. Первые подходы машинного перевода основывались на лингвистических правилах, которые использовались для анализа исходного предложения и создания промежуточного представления, построенного на целевом языке. Такие методы подходят для перевода между языками из близкородственных языковых семей на основе словаря.

Статистический метод машинного перевода не использует традиционные правила языка. Он в основном использует две возможные модели: модель перевода и языковую модель (рис. 2.).

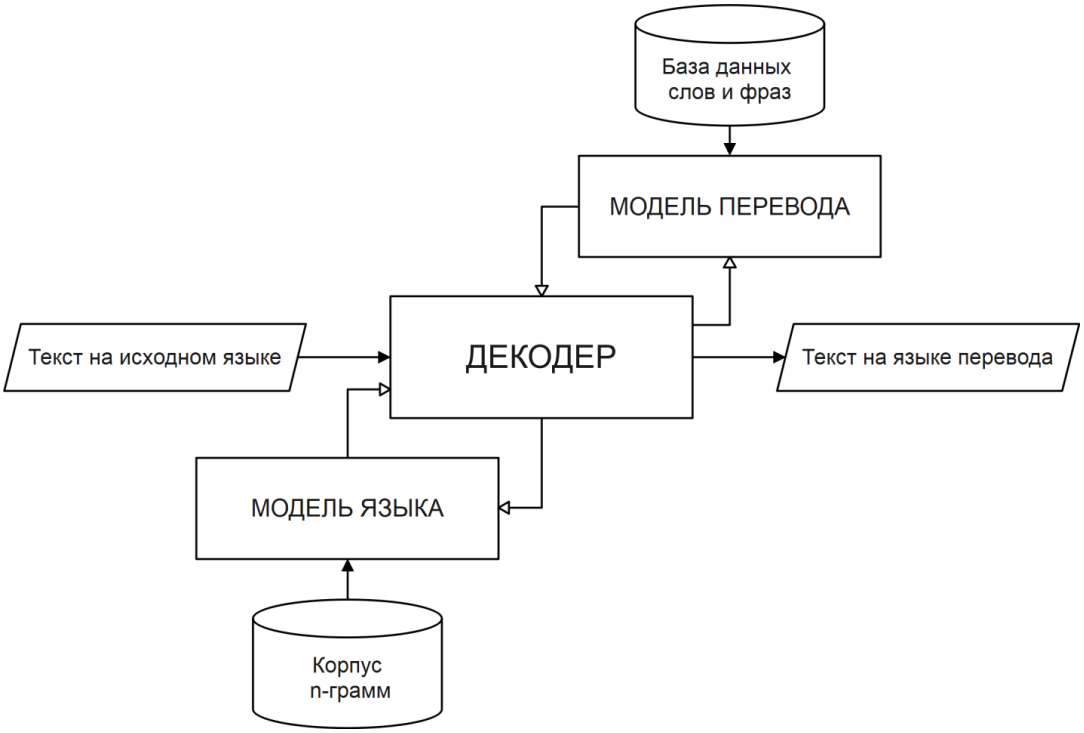


Рисунок 2. Алгоритм машинного статистического перевода

Рассмотрим математическую формулу, определяющую максимальную условную вероятность P(t|s) перевода исходного текста t по отношению к целевому языку s. Обозначая s = s1, . . . , cj, . . . , элементы sls в исходном тексте длиной ls и результатом перевода t = t1, . . . , ti, . . . , tltc с длиной lt, максимальная вероятность обучения пропорциональному переводу может быть получена с помощью следующей статистической математической модели машинного перевода, как показано в формуле:

$$t_{\text{бехт}} = argmax_t\, P(t|s) = argmax_t\, P(s|t) \times P(t) \qquad (4)$$

где, P(s|t) – модель перевода, а P(t) – языковая модель.

По формуле необходимо рассчитать вероятность обратной передачи P(s|t). В случае увеличения составляющей языковой модели мы получаем гарантию перевода с учетом всех грамматических правил языка. Процесс поиска этого наилучшего перевода называется декодированием, и он выполняется

компонентом, называемым декодером.

Согласно нашей модели, возникает вероятность обратного перевода p(s|t). Разработан комплекс методов его расчета на основе двуязычного корпуса. В качестве элементов корпуса можно использовать только *слова* или *словосочетания* на двух параллельных языках.

Модель перевода, опирающаяся на лексику, обеспечивает основу большинства современных методов статистического машинного перевода. В этой модели оценка выравнивания выполняется с использованием распределения вероятностей лексического перевода P(ti|sai), которое определяется путем расчета выравнивания соответствующих пар слов в двуязычном обучающем корпусе. Математическим путем, используя формулу разложения P(t,a|s), получаем следующее уравнение:

$$P(t, a|s) = \prod_{i=1}^{l_t} P(t_i|s_{ai}) P(a_i|a_{i-1}, i, l_t, l_s) \quad (5)$$

где, a – вектор позиций выравнивания, ai = j для слова ti в t.

Модели, основанные на словосочетаниях, используются как относительно длинные элементы перевода. Если переводимый текст состоит из более чем одного слова, называемого словосочетанием, модель перевода охватывает больше информации содержания текста, что приводит к лучшему выбору слов из разных вариантов перевода. При этом предложенное к переводу словосочетание не имеет лингвистической обработки, а соответствующий анализ не производится на основе правил языка: морфологии, синтаксиса и семантики.

Если исходный текст s разбить на I-количество фраз, то модель перевода P(s|t) рассчитывается следующим образом:

$$P(s|t) = \prod_{i=1}^{l} \emptyset(s_i|t_i) d(a_i - b_{i-1} - 1) \quad (6)$$

Языковое моделирование является важным компонентом многих задач обработки естественного языка. В алгоритме статистического машинного перевода лингвистическая модель отвечает за создание перевода с характеристиками логарифмически-линейной модели. Языковая модель изучается на корпусе одного языка, чтобы иметь возможность оценивать вероятность последовательностей слов. Более подходящим методом формирования лингвистической модели является n-грамма.

Лингвистические модели n-грамм. Условно обозначим позиционирование вектора пропорционального перевода a как P(w1,...,wm), который состоит из последовательности слов w1,...,wm. Вероятность совпадения рассчитывается с использованием правил соединения как произведение условной вероятности каждого слова wi, как показано в следующей формуле.

$$P(w_1, \dots, w_m) = \prod_{i=1}^{m} P(w_i|w_1, \dots, w_{i-1}) \quad (7)$$

Затем, используя цепь Маркова [8-А], появление новых переводов предыдущих слов можно приблизить и ограничить до n - 1, как показано в следующей формуле:

$$P(w_1, \dots, w_m) \approx \prod_{i=1}^{m} P\left(w_i \middle| w_{i-(n-1)}, \dots, w_{i-1}\right) \quad (8)$$

В результате мы получаем n-граммную модель порядка n, которая оценивает условную вероятность слова с учетом предыдущих n – 1 слов. Если значение n = 1, n-грамма называется униграммой, если n = 2, n-грамма называется диграммой, а если n = 3, n-грамма называется триграммой. Условная вероятность n-грамм рассчитывается с использованием оценки максимального правдоподобия путем суммирования числа частот следующим образом:

$$P\left(w_i \middle| w_{i-(n-1)}, \dots, w_{i-1}\right) = \frac{count(w_{i-(n-1)},\dots,w_{i-1},w_i))}{count(w_{i-(n-1)},\dots,w_{i-1})} \quad (9)$$

В большинстве случаев при оценке модели n-грамм в машинном переводе относительная длина n-граммных фраз равна трем, то есть для изучения модели используется триграмма.

Результаты исследования основаны на разработке и внедрении системы автоматического перевода текстов на таджикский язык на основе модели бинарного машинного перевода.

Достижению этой цели способствует решение следующих основных задач:

- разработка моделей, методов и математических алгоритмов на основе методологии бинарного машинного перевода таджикского языка;
- создание логичной и реальной параллельной структуры ресурсов, в первую очередь «русско-таджикскую» и «англо-таджикскую» для информационного обеспечения системы машинного перевода;
- определение эффективных алгоритмов поиска, выделения и сортировки текстовых элементов на таджикском языке и путей их реализация в программных модулях для параллельной обработки ресурсов.

В **разделе 2.5** приведены результаты исследования методов и алгоритмов в системах автоматического синтеза речи. На базе природы формирования человеческого голоса и свойств текста на таджикском языке с абстрактным лингвистическим анализом получен цифровой портрет с учетом структур элементов текста и методов кодирования речи.

Алгоритм метода конкатенативного синтеза речи. Последовательность речевых элементов поступает в секцию обработки сигналов, которая выбирает соответствующую звуковую реализацию элементов из базы данных элементов естественной речи и объединяет их в непрерывный речевой сигнал (рис. 3).

Предварительный анализ текста. Для выделения самых мелких частей речи используется механизм получения списка текстовых элементов. Относительно важными элементами декомпозиции текста являются слова и слоги. Для анализа текста используются два основных метода – статистический и словарный. Для моделей, основанных на использовании словаря, должен быть доступен предопределенный словарь. При этом отмечен вариант алгоритма с

наибольшей согласованностью в зависимости от направления обработки текста. Второй вариант словарного алгоритма – это алгоритм, который находит разделение с наименьшим количеством слов.

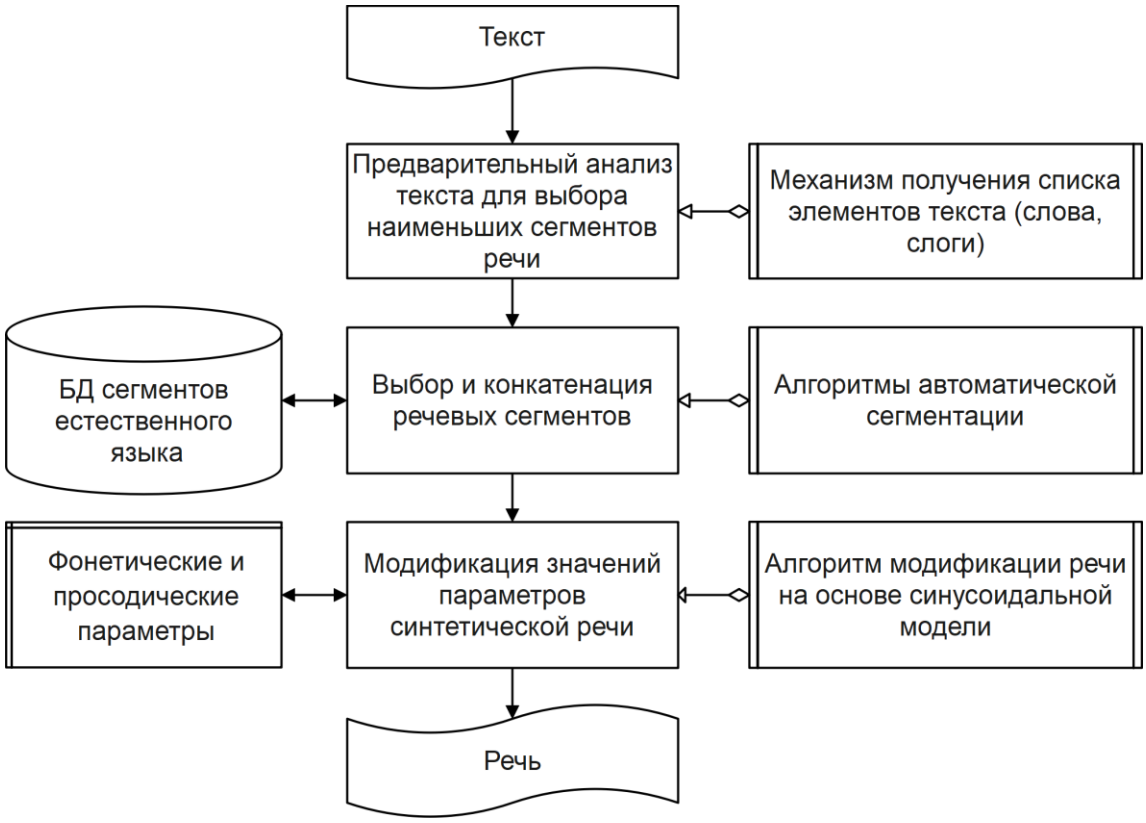


Рисунок 3. Алгоритм синтеза речи на основе конкатенации элементов

Для моделей на основе словаря предоставляется список слов, каждому из которых сопоставлена оценка вероятности того, что это настоящее слово. Пусть $W = \left\{\{w_i, g(w_i)\}_{i=1,\dots,n}\right\}$ будет таким списком, в котором есть кандидат на одно слово, а также функции его качества. Наибольший алгоритм прямого сопоставления текста T для генерации текущего лучшего слова несколько раз с T=t* для каждого этапа можно определить следующим образом:

$\{w^*, t^*\} = argmax_{wt=T} g(w)$ где, ставится условие {w, g(w)} ε W.

Алгоритм декомпозиции кратчайшего пути использует предположение, что правильное расщепление должно либо максимизировать длину всех слов, либо минимизировать общее количество слов. Для предложения S из m символов {c1,c2,…cm} – это наилучшее разбитое на части предложение S* из n* слов.

$$S^* = argmin_{w_1 \dots w_i \dots w_n = T}\ (n) \qquad (10)$$

Эта задача балансировки трансформируется в задачу поиска кратчайшего пути для ориентированного нефазированного графа.

Выбор и соединение частей речи. Для реализации этого этапа необходимо сформировать базу данных элементов естественного языка. Вышеуказанные части производящие звуки речи, такие как слова, слоги или фонемы в данной форме, вместе образуют единую звуковую часть. Помимо высокой

эффективности, автоматическая процедура обеспечивает согласованность размещения границ компонента в пределах ее значений на речевом сигнале.

Алгоритм автоматического расщепления позволяет использовать известные модели континуума. При расчете вероятности Pj того, что состояние компонента qj соответствует наблюдениям в момент времени ρ от t − τ + 1 до t, соответствует формуле:

$$P_{j_{p+1}}(m,\tau) = \sum_{l\in L_m} P_{j_p}(l,\tau) b_{j_l}\left(O_{p+1}\right) \quad t-\tau+1 \le p < t \qquad (11)$$

где,

t - текущая позиция в данном списке,

τ – длина потенциальной составляющей,

ρ – индекс времени, используемый во внутренней текстовой рекурсии. bji(Oρ+1) – вероятность того, что наблюдение O в момент времени ρ + 1 образуется l-м распределением j-компонентной модели. Другими словами, Pjp+1(m, τ) – это вероятность того, что векторы наблюдения Ot-τ+1,...,Ot генерируются из распределения 1,...,M, т.е. базы данных компонентов.

Для решения задачи синтеза речи был изучен относительно сбалансированный механизм, состоящий из комплекса этапов: предварительный анализ текста; выбор и соединение компонентов речи естественного языка из базы данных на основе автоматического алгоритма декомпозиции; изменение значений фонетического и просодического измерений синтетической речи с использованием синусоидальной модели синтеза речи. Таким образом, результаты исследования могут быть непосредственно использованы в проектировании и реализации механизма синтеза речи в таджикском языке, который подробно описан в шестой главе диссертационной работы.

В третьей главе «Объектно-ориентированное моделирование систем обработки текста естественного языка» исследуется компьютерное моделирование разработки систем автоматической обработки информации на естественном языке с учетом объектно-ориентированного подхода. На основе набора диаграмм языка UML разработан типовой проект информационной системы обработки текстовых данных на таджикском языке с учетом модели действия, взаимодействия, структуры и реальности.

В **разделе 3.1** описаны современные средства моделирования процессов. Методологической основой моделирования информационных систем является определение и анализ общей взаимосвязи взаимосвязанных объектов, а также достижение общих целей всеми рабочими группами. Интегрированная архитектура АОТ реализована из набора взаимосвязанных информационных технологий, процессов, алгоритмов, набора методов обработки текста, инструментов, интерфейсов и набора процессов. Предлагаемая модель представляет собой цифровое представление таджикского языка.

В современных условиях для моделирования программного обеспечения и информационных систем используются стандартные методы и языки функционального моделирования, такие как IDEF, DFD, UML.

В **разделах 3.2-3.5** предложены способы моделирования поведения,

взаимодействия и концептуальная модель системы автоматической обработки информации, что связано с разработкой диаграммы языка UML. Определены основные действия, выполняемые информационной системой: комплексное формирование элементов текста, управление процессами обработки информации, автоматическая обработка текстовых данных, синтез речи, обработка тезауруса, проверка орфографии и машинный перевод.

На основе диаграммы классов системы TajLINGVO, проведен статистическая структура выявления общих объектов информационной системы и их логические связи. Кроме того, были исследованы возможные случаи появления текстового элемента в процессе обработки (рис. 4).

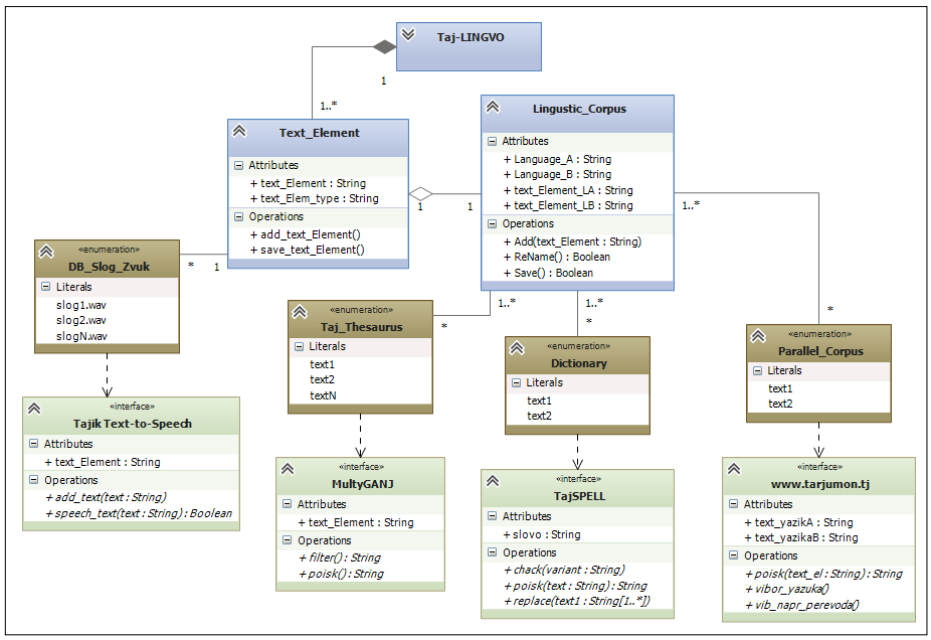


Рисунок 4. Диаграмма классов системы TajLINGVO

В **разделе 3.5** определена физическая модель информационной системы с целью обобщения прикладных возможностей программного обеспечения и технических средств. Установлены логические связи между статистической структурой информационной системы, программными компонентами и узлами технических средств при реализации информационной системы.

В целом разработана компьютерная модель типовой информационной системы обработки текстовой информации на таджикском языке с учетом модели поведения, взаимодействия, статистической структуры и физических средств.

В четвертой главе «Проектирование, разработка и внедрение автоматической проверки правописания таджикского языка» рассматриваются вопросы проектирования, обработки и реализации системы автоматической проверки орфографии текста на таджикском языке.

В **разделах 4.1-4.3** представлены результаты исследования и реализации электронных словарей, компьютерного тезауруса таджикского языка, системы автоматической конвертации шрифтов в тексте на стандартные шрифты таджикского языка.

На основе моделей и математических методов разработаны следующие

алгоритмы для решения задачи выявления и исправления орфографических ошибок в текстовой информации на таджикском языке:

- алгоритм транслитерации текста в стандартный алфавит;
- алгоритм обнаружения орфографических ошибок;
- алгоритм исправления ошибок;
- алгоритм проверки орфографии.

В **разделе 4.4** представлен алгоритм проверки орфографии таджикского языка с возможностью обнаружения орфографических ошибок и их исправления. Процедура проверки основана на трех условиях: в слове отсутствует одна буква, две соседние буквы изменили свое место в слове, в слове есть лишняя буква.

Если одно из трех указанных условий найдено и исправлено (условие W=S[I]), то сохраняем слово в списке возможно правильных слов C[J]. Сортируем список результатов в порядке возрастания частоты их появления.

Убираем первые семь элементов из списка правильных слов. Если до конца списка S[I] процедура проверки не находит совпадения со словом W, то определяется «совпадающее слово не найдено».

Согласно предложенному алгоритму слово с ошибкой сравнивается с каждым словом словаря. Основная методология, адаптированная в алгоритме, заключается в том, что ранее сформированный список слов S[1...i] можно преобразовать в итоговый список слов C[1..j]. Наконец, процедура возвращает значение желаемого слова, которое пользователь может выбрать. Процедура проверки реализуется хеш-функцией, обрабатывающей словарную базу данных, то есть хеш-таблицу.

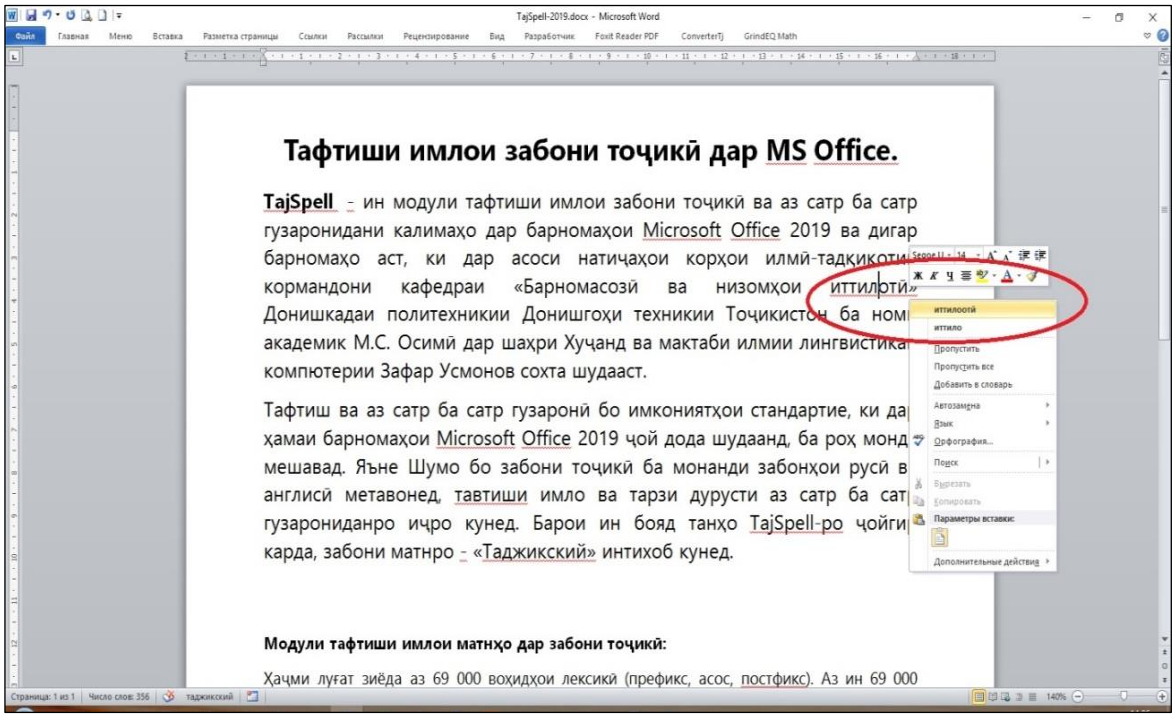


Рисунок 5. Проверка орфографии TajSpell в MS Word

В разделе 4.5 на основе полученных результатов приводится описание модуля TajSpell с возможностью исправления таджикского текста, проверкой

орфографии, переходом от строки к строке и тезаурусом таджикского языка. Таким образом можно проверять таджикские тексты в пакете прикладных программ MS Office, в стандартной кодировке UNICODE (рис. 5).

Модуль TajSpell в приложениях Microsoft Office полностью поддерживает буквы таджикского языка и реализован на основе обмена с модулем проверки орфографии в кодировке UNICODE.

В пятой главе «Проектирование, разработка и внедрение таджикского автоматического переводчика» дается описание проектирования, разработки и внедрения таджикского автоматического переводчика.

В **разделе 5.1** приведены результаты исследования проблем перевода текстов с разных языков на таджикский и наоборот, а также выявлены проблемы художественного перевода и его зависимость от машинного перевода на базе технологии Google.

В **разделах 5.2-5.4** рассматривается проблема разработки автоматической системы, алгоритмы автоматической транслитерации букв, реализация статистического метода перевода и логическая структура машинного перевода на примере таджикского языка.

В **разделе 5.5** приводится описание информационной системы машинного перевода текста с таджикского языка на русский и английский языки соответственно. Решение данной задачи было основано на базе двух источников данных – параллельного корпуса Taj-Rus-Corp и Taj-Eng-Corp.

На основе статистической системы машинного перевода и системы машинного перевода на основе правил была разработана модель переводчика таджикского языка. Для обеспечения машинного перевода текста на таджикский язык разработана информационная система в виде Web-приложения (рис. 6).

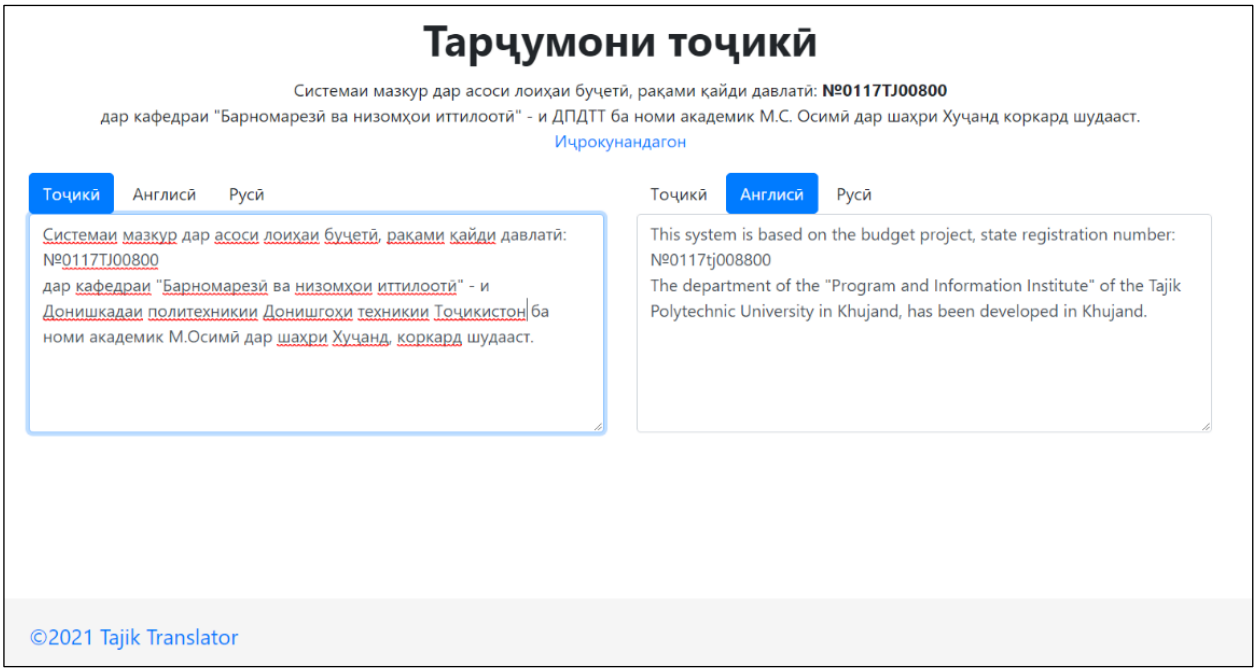


Рисунок 6. Web-приложение таджикского переводчика - www.tarjumon.tj

На первом этапе общее количество элементов запаса определяется следующим образом:

- таджикско-русский – 42 000, в том числе более 27 000 слов;
- русско-таджикский – 68 000, в том числе словарный запас более 54 000;
- англо-таджикский – 12 000, в том числе словарный запас более 5 000;
- таджикско-английский – 24000, в том числе более 11000 слов.

Проект доступен в Интернете по адресу www.tarjumon.tj для онлайн-перевода текстовой информации с таджикского языка на русский, английский языки и в обратном порядке.

В шестой главе «Проектирование, разработка и внедрение компьютерного синтеза таджикской речи по тексту» рассматривается постановка задачи математического моделирования и компьютерной реализации синтеза таджикской речи на основе предложенного текста.

В **разделе 6.1** осуществлена задача анализа текстовых данных на основе различных слоговых структур. Из результатов, приведенных в таблицы 1 видно, что объем (количество букв) 1 и 14 – это минимум и максимум структуры слова. Слово с числом более 14 букв в обработанном тексте не обнаружено, хотя такие слова встречаются и в таджикском языке.

Таблица 1. Статистика таджикских слов по количеству букв

Длина слова	1	2	3	4	5	6	7
Встречаемость в %	0,87	16,14	10,94	11,32	16,95	13,95	12,81

Длина слова	8	9	10	11	12	13	14
Встречаемость в %	8,88	4,98	2,92	1,00	0,57	0,10	0,02

По статистической закономерности текстовых данных в таджикском языке при обработке текстовых данных выявлено всего 274 различные структуры слов (элемент, гласная - 1, согласная - 0) в объеме 1724472 слов.

Установлено, что 8 единиц покрывают 50%, а 23 элемента покрывает 75% таджикских текстов. Также выявлено, что 51 элемент охватывает 90%, а 76 частей охватывают 95% таджикских текстов (табл. 2).

Таблица 2. Частота встречаемости слов в форме слогового состава (до 50%)

№	$W_{0,1}^{*}$	%	№	$W_{0,1}^{*}$	%	№	$W_{0,1}^{*}$	%
1	01	11,006	9	010010	3,684	17	1010	1,192
2	010	8,849	10	0101010	3,258	18	01001010	1,142
3	01010	6,781	11	0100	2,799	19	010100	1,087
4	01001	5,486	12	01010101	1,735	20	01001011	1,053
5	10	5,096	13	01011	1,711	21	100	0,986
6	0101	5,066	14	1001	1,280	22	10101	0,960
7	010101	4,773	15	010011	1,226	23	10010	0,957
8	0100101	3,787	16	0101001	1,218			

На основе соответствующего распределения 274 единиц выявлено всего 9 различных сочетаний слогов в таджикском языке, 6 из которых соответствуют правилам таджикского языка: «1», «10», «01», «010», «100», «0100» (табл. 3)

Таблица 3. Частота встречаемости (в %) слоговых структур

Слоги	1	10	01	100	010	0100	001	0010	00100
Встречаемость	8.10	5.74	56.56	0.78	25.75	2.95	0,05	0,06	0,01

Согласно композиционной структуре разработан алгоритм пословного членения с учетом 6 слоговых шаблонов. Компьютерная программа на основе разработанного алгоритма была использована для проведения статистического исследования различных слогов таджикского языка. На 3800 страницах случайной выборки было выявлено 3259 различных производных слогов.

Для обеспечения прозрачности полученных результатов были исследованы статистические закономерности слогового состава таджикского языка в структуре слов, а также используемых слов.

В **разделах 6.2-6.4** на основе математических моделей и специальных методов программирования разработаны следующие алгоритмы:

1. Алгоритм произношения слов.
2. Алгоритм произношения цифр и символов.
3. Алгоритмы безударного и ударного произношения текста.
4. Алгоритм произношения морфемы слова.
5. Алгоритм произношения таджикского текста, содержащего русские слова.

На базе полученных результатов разработаны автоматическая система с возможностью синтеза речи на таджикском языке «Tajik Text-to-Speech», автоматический «диктор» в ОС Windows «Tajik Text Narrator», а также онлайн пользователей модуль www.tajlingvo.tj\talaffuz.

В **разделе 6.5** исследуются проблемы распознавания речи на таджикском языке на основе сравнительного анализа разработанной системы синтеза речи. Сравнительный анализ системы распознавания устной речи определил, что для достижения цели распознавания речи на таджикском языке используются возможности алгоритма динамического обмена измерениями времени, алгоритмов распознавания слогов таджикской речи в пространственно-временных изменениях. На основе этих показателей предложен алгоритм распознавания речи на таджикском языке на базе анализа слоговой структуры слов, который будет использоваться в дальнейших исследованиях.

Научные результаты, полученные в рамках низкоуровневой автоматической обработки синтеза речи на таджикском языке, в будущем будут использованы как основа для решения проблемы распознавания речи на таджикском языке, в связи с чем были проанализированы основные проблемы решения задачи автоматического распознавания речи таджикского языка.

ВЫВОДЫ

1. На основе проведенного анализа достижений в сфере компьютерной лингвистики, результатов научных исследований в зарубежных странах и в Республике Таджикистан, собственных экспериментов и теоретических исследований **сформулированы** задачи исследования, заключающиеся в проектировании, разработке и реализации автоматизированных информационных систем обработки информации на таджикском языке [1-А]-[3-А], [8-А], [26-А], [29-А], [32-А].

2. Для решения проектирования информационных систем обработки информации на таджикском языке в условиях глобализации таджикского языка и факторов использования государственного языка в делопроизводстве **предложен** объектно-ориентированный подход. Сущностью объектно-ориентированного подхода является анализ элементов текста и речи как объекта управления; моделирование процессов поведения и взаимодействия элементов текста; статическая и концептуальная модель системы обработки информации; формирование физической модели системы методов обработки информации [4-А], [28-А], [39-А], [61-А], [65-А].

3. **Разработаны** новые математические модели, методы и алгоритмы обработки информации, на основе которых реализованы новые средства формирования базы данных и программирования для анализа текстовых данных на таджикском языке [6-А], [16-А], [22-А], [38-А], [65-А], [68-А].

4. На основе методологии теоретически обоснована и практически **исследована** проблема проектирования, разработки и реализации прикладных программных обеспечений для решения задач автоматической проверки правописания, машинного перевода и синтеза речи на таджикском языке [21-А], [60-А].

5. **Предложена** объектно-ориентированная методология разработки автоматизированных информационных систем, состоящая из совокупности моделей, методов, алгоритмов и процедур, которые **реализованы** в задачах моделирования процессов обработки информации на естественном языке [1-А], [6-А], [35-А].

6. В результате анализа методологических основ проектирования автоматических информационных систем обработки информации **обоснованы** методы компьютерного моделирования процессов и статистического анализа элементов текста, а так же алгоритмы и программные средства автоматизации процессов их реализации [7-А], [17-А], [24-А], [30-А].

7. В работе **сформулированы** основы метода эффективного сбора, анализа и обработки текстовой информации на таджикском языке. **Представлена** многоуровневая модель процессов получения цифрового портрета текста, на основе которого **выделены** основные характеристики и **составлена** классификация его элементов текста [14-А], [15-А], [18-А], [25-А], [33-А], [53-А], [58-А], [59-А].

8. Для проектирования, разработки и реализации задачи автоматической проверки правописания текста на таджикском языке **разработаны** механизмы, процедуры и алгоритмы обработки текстовых данных. **Реализован** комплекс автоматических компьютерных систем, включающий в себя электронные

словари, компьютерный тезаурус, конвертация нестандартных шрифтов на стандартную кодировку Unicode, модуль TajSpell с возможностью исправления орфографии, расстановка переноса слов, тезаурус таджикского языка в пакете программ MS Office [12-А], [13-А], [19-А], [37-А], [54-А], [62-А], [64-А], [66-А].

9. Для решения задачи разработки таджикского автоматического переводчика **обоснованы** математические модели логических структур артефактов, методы машинного перевода и алгоритмы их реализации. **Сформирована** система транслитерации текстов с латиницы и кириллицы на таджикскую кириллицу. Для информационного обеспечения системы машинного перевода **сформированы** параллельные таджикско-русский и таджикско-английский корпуса. На основе технологии Google **разработан** комплекс программ двустороннего автоматического перевода текста в виде Web-приложения с возможностью онлайн-перевода текстовой информации с таджикского языка на русский и английский [4-А], [5-А], [11-А], [27-А], [32-А], [34-А], [36-А], [52-А], [55-А], [56-А].

10. Впервые **спроектирована** система автоматического синтеза речи на таджикском языке, основанная на методе конкатенации слогов. **Предложены** математические модели слоговых структур слов таджикского языка, на их основе **получено** многообразие слогов и **сформирована** база слог-звук. **Разработан** ряд алгоритмов озвучивания текста на таджикском языке с учетом слогов, морфем, чисел, знаков препинания и слов русизмами. Полученные результаты **реализованы** в прикладных программах озвучивания текста на таджикском языке Tajik Text-to-Speech и Computer Tajik Text Narrator [9-А], [20-А], [23-А], [40-А], [57-А], [67-А].

11. Полученные результаты были **представлены** на научно-исследовательских конференциях на уровне республики и за рубежом, где получили высокую оценку. **Внедрение** результатов работы в государственных учреждениях и высших учебных заведениях позволило решить задачи эффективного использования таджикского языка в процессе делопроизводства, а также может **способствовать** развитию науки математического моделирования, проектирования информационных систем и компьютерной лингвистики [41-А]-[50-А].

12. Результаты диссертационной работы могут послужить **фундаментальной основой** для изучения особенностей таджикского языка как для граждан Республики Таджикистан, так и всем желающим за его пределами. Все достигнутые результаты и разработанные проекты находятся в свободном доступе в сети интернет по адресу www.tajlingvo.tj [51-А].

РЕКОМЕНДАЦИИ ПО ПРАКТИЧЕСКОМУ ИСПОЛЬЗОВАНИЮ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ

Результаты, полученные в диссертации, являются решением актуальных и приоритетных проблем подготовки математических и компьютерных моделей изучения языка и автоматических методов обработки текстовых данных в вопросах проверки орфографии в тексте, машинного перевода текста, синтеза и распознавания речи на таджикском языке. Данный комплекс вопросов имеет большое значение в повышении качества изучения таджикского языка с

использованием возможностей информационных технологий и ускорения процесса оформления документов в Республике Таджикистан и за рубежом.

Итоги диссертационного исследования также могут быть использованы в учебном процессе, в научно-исследовательских институтах и высших профессиональных учреждениях при чтении специальных курсов в области компьютерной лингвистики и информационных технологий. Кроме того, они могут быть широко использованы при написании курсовых и дипломных работ студентами, диссертаций аспирантами, соискателями ученых степеней в области математики, информационных технологий и компьютерной лингвистики. Системы автоматической обработки текстов, разработанные на таджикском языке, рекомендуются к использованию на таджикском языке в документационной деятельности в организациях и на предприятиях внутри страны и за рубежом.

.

СПИСОК ОСНОВНЫХ ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

Монографии

[1-А] **Худойбердиев, Х.А.** Низомҳои худкори коркарди маълумот бо забони тоҷикӣ. [Матн] / З.Д. Усманов **Х.А. Худойбердиев** – Хуҷанд, ДДХБСТ, 2022. – 186 с. (на таджикском языке)

[2-А] **Худойбердиев, Х.А.** Комплекси барномаҳо барои талаффузи овози тоҷикӣ аз рӯйи матн. [Матн] / Усмонов З.Д., **Х.А. Худойбердиев** – Душанбе. Адиб, 2014. –158 с. (на таджикском языке)

[3-А] **Худойбердиев, Х.А.** Опыт компьютерного синтеза таджикской речи по тексту. [Матн] / З.Д. Усманов, **Х.А. Худойбердиев** – Душанбе, Ирфон, 2010, –145 с.

Статьи, опубликованные в изданиях из перечня ведущих рецензируемых журналов, рекомендованных ВАК при Президенте Республики Таджикистан, ВАК Российской Федерации

[4-А] **Худойбердиев, Х.А.** Оид ба низоми тарҷумони омории мошинӣ барои забони тоҷикӣ [Матн] / **Х.А.Худойбердиев** // Паёми донишгоҳи технологии Тоҷикистон. – 2023. № 3 (55). –С. 140-146.

[5-А] **Худойбердиев, Х.А.** Разработка и реализация системы машинного перевода на основе правил с русского на таджикский язык [Текст] / **Х.А.Худойбердиев** // Политехнический Вестник ТТУ имени академика М.С. Осими (Серия: Интеллект. Инновации. Инвестиции.). – 2023. –№2(62). –С. 33-36.

[6-А] **Худойбердиев, Х.А.** Моделирование системы автоматической обработки текста на таджикском языке [Текст] / **Х.А.Худойбердиев** // International Journal of Open Information Technologies. –2023.– Т.11, № 3.– С.27-33.

[7-А] **Худойбердиев, Х.А.** Цифровой портрет таджикского языка на основе статистических закономерностей кириллического алфавита [Текст] / **Х.А.Худойбердиев**, Ш.Н. Ашурова // Политехнический Вестник ТТУ имени академика М.С. Осими (Серия: Интеллект. Инновации. Инвестиции.). – 2022. – №4(60). – С. 29-32.

[8-А] **Худойбердиев, Х.А.** Вклад Усманова Зафара Джураевича в компьютерную лингвистику таджикского языка [Текст] / **Х.А.Худойбердиев** //

Вестник Технологического университета Таджикистана. – 2022. № 4-1 (51). –С. 140-146.

[9-А] **Худойбердиев, Х.А.** Амсиласозии раванди шинохти нутқ дар заминаи нутқи забони тоҷикӣ [Матн] / Б.Х.Ашурзода, **Х.А.Худойбердиев** // Паёми Политехникӣ. ДДТ ба номи М.Осимӣ. (Бахши Интеллект, Инноватсия, Инвеститсия.) – 2022. – № 2 (58). – С. 39-42.

[10-А] **Худойбердиев, Х.А.** Масъалаҳои тарҳрезӣ ва коркарди луғатҳои электронӣ дар коркарди низомҳои худкори тарҷумон бо забони тоҷикӣ [Матн] / **Х.А.Худойбердиев //** Паёми Политехникӣ. ДДТ ба номи М.Осимӣ. (Бахши Интеллект, Инноватсия, Инвеститсия.) – 2022. – № 1 (57). – С. 41-47.

[11-А] **Худойбердиев, Х.А.** О проблемах художественного перевода и его взаимосвязь с машинным переводом на примере таджикского языка [Текст] / **Х.А.Худойбердиев** // Вестник технологического университета Таджикистана. – 2021. – № 4 (47). – С. 163-168.

[12-А] **Худойбердиев, Х.А.** Об алгоритме проверки орфографии на примере таджикского языка [Текст] / **Х.А.Худойбердиев //** Политехнический Вестник ТТУ имени академика М.С. Осими (Серия: Интеллект. Инновации. Инвестиции.). – 2021. – № 3 (31). – С. 48-53.

[13-А] **Худойбердиев, Х.А.** Система автоматической проверки орфографии таджикского языка – TajSpell [Текст] / О.М.Солиев, **Х.А.Худойбердиев**, Г.М.Довудов //. Вестник технологического университета Таджикистана. – 2021. – № 3 (46). – С. 188-193.

[14-А] **Худойбердиев, Х.А.** О распознавании автора текста на узбекском языке с помощью символьных униграмм [Текст] / **Х.А.Худойбердиев**, А.А.Косимов, П.Э.Зульфикарова // Проблемы вычислительной и прикладной математики. Научно-инновационный центр информационно-коммуникационных технологий Ташкентского университета информационных технологий имени М. аль-Хоразми. – 2020. – № 6 (30). – С. 49-55.

[15-А] **Худойбердиев**, **Х.А.** Оид ба монандкунии матн дар асоси басомади ҳиҷоҳо [Текст] / **Х.А.Худойбердиев**, А.А.Қосимов, Х.А.Тошхӯҷаев // Политехнический вестник. серия: интеллект. инновации. инвестиции. – 2020. – 2 (50). – С. 52-56.

[16-А] **Худойбердиев, Х.А.** О распознавании автора текста на основе частотности слогов [Текст] / **Х.А.Худойбердиев**, А.А.Косимов // Доклады Академии наук Республики Таджикистан. – 2019. – Т.62, № 11-12. – С. 641-645.

[17-А] **Худойбердиев, Х.А.** О статистических закономерностях слогового состава таджикского языка [Текст] / **Х.А. Худойбердиев //** Вестник Таджикского технического Университета, – 2015. – № 3 (31). – С. 48-53.

[18-А] **Худойбердиев, Х.А**. О соотношении словоформ и словоупотреблений в русском переводе произведения А.Фирдоуси «Шахнаме» [Текст] / **Х.А.Худойбердиев**, А.А.Косимов // Доклады Академии наук Республики Таджикистан. – 2015. – Т.58, № 9. – С. 786-792.

[19-А] **Худойбердиев, Х.А**. Об автоматическом конвертировании таджикского текста к стандартной графике [Текст] / **Х.А.Худойбердиев** // Доклады Академии наук Республики Таджикистан, – 2014. – Т.57, № 3. – С. 210-214.

[20-А] **Худойбердиев, Х.А.** О синтезе таджикской речи с русизмами [Текст] / З.Д.Усманов, **Х.А.Худойбердиев** // Доклады Академии наук Республики Таджикистан. – 2009. – Т.52, – № 5. – С. 358-361.

[21-А] **Худойбердиев, Х.А.** О синтезаторе таджикской речи по тексту [Текст] / З.Д.Усманов, **Х.А.Худойбердиев** // Доклады Академии наук Республики Таджикистан. – 2009. –Т.52, № 4. – С. 267-271.

[22-А] **Худойбердиев, Х.А.** Об автоматическом разложении слов на слоги. [Текст] / **Х.А.Худойбердиев** // Доклады Академии наук Республики Таджикистан. – 2007. – Т.50, № 5. – С. 417-419.

[23-А] **Худойбердиев, Х.А.** Алгоритм безударного озвучивания таджикского текста. [Текст] / З.Д.Усманов, **Х.А.Худойбердиев** // Доклады Академии наук Республики Таджикистан. – 2007. – Т.50, № 4. – С. 302-305.

[24-А] **Худойбердиев, Х.А.** О многообразии слогов таджикского языка. [Текст] / **Х.А.Худойбердиев** // Известия Академии наук Республики Таджикистан. – 2007. – №2 (127). – С. 31-34.

[25-А] **Худойбердиев, Х.А.** О слоговой структуре слов таджикского языка [Текст] / З.Д.Усманов, **Х.А.Худойбердиев** // Доклады Академии наук Республики Таджикистан. – 2006. – Т. 49, № 6. – С. 489-492.

Статьи в других журналах

[26-А] **Худойбердиев, Х.А.** Рушди илми лингвистикаи компютерӣ дар Ҷумҳурии Тоҷикистон [Матн] / О.М. Солиев, **Х.А. Худойбердиев**, Г.М. Довудов, Ш.Н. Ашӯрова // Паёми ДПДТТ ба номи академик М.С.Осимӣ. – 2022. – № 2 (23). – С. 17-24.

[27-А] **Худойбердиев, Х.А.** Проектирование и программная реализация автоматической транслитерации в цифровой библиотеке [Текст] / **Х.А. Худойбердиев**, М.П. Музаффаров, Ф.Э. Мирзозода // Вестник ПИТТУ имени академика М.С.Осими. – 2022. – № 1 (22). – С. 7-15.

[28-А] **Худойбердиев, Х.А.** Перспективы развития информационного пространства и цифровизации в Таджикистане: обзор основных тенденций [Текст] / Х.Т. Максудов, **Х.А. Худойбердиев**, Ш.Х. Максудов // Вестник ПИТТУ имени академика М.С. Осими. – 2021. – № 4 (21). – С. 7-18.

[29-А] **Khurshed A. Khudoyberdiev.** The Algorithms of Tajik Speech Synthesis by Syllable. Polytechnic institute of Tajik technical university named after academician M.S. Osimi, - Polytechnic institute of Tajik technical university named after academician M.S. Osimi, Khujand. Tajikistan. International Forum “IT-Technologies for Engineering Education: New Trends and Implementing Experience” (ITEE-2019). Anthropological Dimension of Digital Technologies in Engineering Education ITM Web of Conferences 35, 07003 (2020).

[30-А] **Худойбердиев, Х.А.** Сравнительный анализ систем распознавания звука Sphinx и Mozilla Deepspeech [Текст] / **Х.А. Худойбердиев**, Р.М. Воситов // Вестник ПИТТУ имени академика М.С.Осими. – 2021. – № 1 (18). – С. 7-13.

[31-А] **Худойбердиев, Х.А.** Муаммоҳои тарҷумаи бадеӣ ва вобастагии он бо тарҷумаи мошинӣ дар Тоҷикистон [Матн] / З.А. Раҳмонов, **Х.А. Худойбердиев** // Паёми ДПДТТ ба номи академик М.С. Осимӣ. – 2020. – № 2 (7). – С. 7-11

[32-А] **Худойбердиев, Х.А.** Разработка параллельного корпуса таджикского и русского языков [Текст] / **Худойбердиев, Х.А.**, О.М. Солиев, П.А. Солиев //

Новые информационные технологии в автоматизированных системах. – 2019. – № 22. – С. 179-181.

[33-А] **Худойбердиев, Х.А.** Информационная система и каталогизации кодексов республики Таджикистан [Текст] / **Х.А. Худойбердиев**, И.А. Джалолов // Вестник ПИТТУ имени академика М.С.Осими. – 2019. – № 3 (12). – С. 9-18.

[34-А] **Худойбердиев, Х.А**. Захираи мувозии забони точикӣ-русӣ: коркард ва тавсифи он [Матн] / **Х.А. Худойбердиев**, А.А. Назаров // Паёми ДПДТТ ба номи академик М.С.Осимӣ. – 2019. – № 1(10). – С. 7-12.

[35-А] **Худойбердиев, Х.А.** Сегментация речевого сигнала на базе слоговых структур таджикского языка [Текст] / **Х.А. Худойбердиев** // Новые информационные технологии в автоматизированных системах. – 2018. – № 21. – С. 181-182.

[36-А] **Худойбердиев, Х.А.** Сохтори мантиқӣ ва таҳлили артефактҳои тарҷумаи мошинӣ [Матн] / **Х.А. Худойбердиев**, З.А. Раҳмонов // Паёми ДПДТТ ба номи академик М.С. Осимӣ. – 2018. – № 2 (7). – С. 7-11.

[37-А] **Худойбердиев, Х.А.** Лингвистический тезаурус таджикского языка [Текст] / **Х.А. Худойбердиев**, О.М. Солиев // Новые информационные технологии в автоматизированных системах. – 2017. – № 20. – С. 103-105.

[38-А] **Худойбердиев, Х.А.** Модель анализа и сегментации речевого сигнала для послогового распознавания таджикской речи [Текст] / **Х.А. Худойбердиев** // Вестник Технологического университета Таджикистана. – 2017. – № 4 (31). – С. 85-87.

[39-А] **Худойбердиев, Х.А.** О множестве анаграмм в произведениях К.Худжанди [Текст] / **Х.А. Худойбердиев**, А.А. Косимов // Вестник ПИТТУ имени академика М.С. Осими. – 2017. – №2 (3). – С. 14-22.

[40-А] **Худойбердиев, Х.А.** О синтезаторе таджикской речи по тексту [Текст] / **Х.А. Худойбердиев** // Новые информационные технологии в автоматизированных системах. – 2013. – № 16 – С. 273-276.

Выступления и тезисы в конференциях

[41-А] **Худойбердиев, Х.А.** О некоторых способах математического моделирования синтеза и распознавания речи [Текст] / **Х.А. Худойбердиев** // Материалы международной конференции «Современные проблемы математики», посвящённой 50-летию Института математики им. А. Джураева Национальной академии наук Таджикистана. – Душанбе, Института математики им. А. Джураева НАНТ, 2023. – С. 253-255.

[42-А] **Худойбердиев, Х.А.** Формирование электронного словаря для системы автоматического перевода текста с таджикского языка на русский [Текст] / **Х.А. Худойбердиев**, А.А. Назаров, Ш.Н. Ашурова // Всероссийская научно-практическая конференция с международным участием «Информационный обмен в междисциплинарных исследованиях II». – Рязань, 2023. – С. 227-231.

[43-А] **Худойбердиев, Х.А.** Низомҳои худкор барои коркарди матн бо забони тоҷикӣ [Матн] / **Х.А. Худойбердиев //** Международная научно-практическая конференция «Новые достижения в области естественных наук и информационных технологий». – Душанбе, РТСУ, 2023. – С. 194-196.

[44-А] **Худойбердиев, Х.А.** Тарҳрезии низомҳои худкор барои коркарди

матн бо забони тоҷикӣ [Матн] / **Х.А. Худойбердиев //** Конференсияи илмӣ-амалии ҷумҳуриявӣ бахшида ба рӯзи байналмилалии забони модарӣ таҳти унвони “Забони модарӣ – сарчашмаи худшиносӣ ва маънавиёти миллӣ”. – Душанбе, Кумитаи забон ва истилоҳоти назди Ҳукумати ҶТ, 2023.

[45-А] **Худойбердиев, Х.А.** Баланд бардоштани сифати корҳои хаттӣ бо истифодаи барномаи зидди асардуздӣ (Antiplagiat_TJ) [Матн] / **Х.А. Худойбердиев**, А.А. Косимов, М.Х. Файзуллозода, Х.М. Муродов, Ё.О. Зулфов // Конференсияи ҷумҳуриявӣ илмию амалӣ дар мавзӯйи «Тадбиқи технологияҳои иттилоотӣ ва коммуниткатсионӣ дар саноаткунонии кишвар», бахшида ба ҳадафи чоруми стратегии миллӣ. – Душанбе, Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ, 2022.

[46-А] **Худойбердиев, Х.А.** Современные тенденции в компьютерной лингвистике таджикского языка [Текст] / **Х.А. Худойбердиев** // Республиканская научно-практическая конференция «Актуальные проблемы лингвистики и лингводидактики в современных условиях». – Душанбе, Филиал Московского государственного университета имени М.В. Ломоносова в городе Душанбе, 2022. – С. 279-284.

[47-А] **Худойбердиев, Х.А.** О проблеме автоматической транслитерации текста на таджикском языке [Текст] / **Х.А. Худойбердиев** // IV Международная научно-практическая конференция «Наука и технологии». – Алматы, Казахстан, 2022. – С. 101-106.

[48-А] **Худойбердиев, Х.А.** Таҳлили масъалаҳои асосии пешбарии тарҷумаи мошинӣ дар мисоли забони тоҷикӣ [Матн]. / **Х.А. Худойбердиев** // Конференсия ҷумҳуриявии илмӣ-амалӣ Масъалаҳои мубрами тарҷума ва забоншиносӣ дар замони муосир”. – Душанбе, Донишкадаи давлатии забонҳои тоҷикистон ба номи Сотим Улуғзода, 2019.

[49-А] **Худойбердиев, Х.А.** Методҳо ва алгоритмҳо барои шинохти овоз [Матн] / Н.С. Маҳмудов, **Х.А. Худойбердиев**, Ғ.Ҳ. Сафаров // Конференсияи илмӣ-амалии омӯзгорон, муҳаққиқони ҷавон бахшида ба 30-солагии Истиқлолияти давлатии Ҷумҳурии Тоҷикистон. – Хуҷанд, ДПДТТХ ба номи академик М.С.Осими, 2019.

[50-А] **Худойбердиев, Х.А.** Алгоритмы послогового распознавания таджикской речи в амплитудно-временном пространстве [Текст] / **Х.А. Худойбердиев**//Научно-практическая конференция «Применение информационно-коммуникационных технологий для инновационного развития Республики Таджикистан». – Душанбе, ТУТ, 2017.

Авторские свидетельства и государственная регистрация информационных ресурсов

[51-А] **Худойбердиев, Х.А.** Web-приложение “Автоматические системы обработки информации на таджикском языке – www.tajlingvo.tj" [SOFT] / Х.А. Худойбердиев // – 28.04.2022. – № 4202200496.

[52-А] **Худойбердиев, Х.А.** Web-приложение таджикский переводчик (tarjumon.tj) [SOFT] / **Х.А.Худойбердиев**, О.М.Солиев, П.А.Солиев, Г.М.Довудов, А.А.Назаров // – 03.12.2021/ –№ 4202100482.

[53-А] **Худойбердиев, Х.А.** Web-сайт “Электронный каталог кодексов Республики Таджикистан” [SOFT] / **Х.А. Худойбердиев**, И.А. Джалолов // –

25.02.2021. –№ 4202100470.

[54-А] **Худойбердиев, Х.А.** Автоматическая система TajSpell-2.0. для проверки орфографии таджикского языка в офисном пакете приложений MS Office 2010-2019 [SOFT] / З.Д. Усманов, О.М. Солиев, **Х.А. Худойбердиев**, Г.М. Довудов // – 30.07.2020. – № 4202000456.

[55-А] **Худойбердиев, Х.А.** Web-приложение Tajik-Russian-Parallel Corpus [SOFT] / **Х.А. Худойбердиев**, О.М. Солиев, Г.М. Довудов, А.А. Косимов // – 30.04.2019. – № 4201900402.

[56-А] **Худойбердиев, Х.А.** Web-приложение Tajik-English-Parallel Corpus [SOFT] / **Х.А. Худойбердиев**, О.М. Солиев, А.А. Назаров, П.А. Солиев // – 30.04.2019. – № 4201900401.

[57-А] **Худойбердиев, Х.А.** Компьютерный Диктор таджикского текста Computer Tajik Text Narrator [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, А.А. Худойбердиев // –10.06.2018. – № 4201800386.

[58-А] **Худойбердиев, Х.А.** База данных «Единица измерений текстов произведений таджикских современных поэтов и писателей» [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, А.А. Косимов // – 16.05.2018. –№ 4201800381.

[59-А] **Худойбердиев, Х.А.** База данных «Единица измерений текстов произведений таджикских классических поэтов и писателей» [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, А.А. Косимов // – 16.05.2018. – № 4201800380.

[60-А] **Худойбердиев, Х.А.** Web-приложение проверки уникальности текста на таджикском языке Taj_Text_Plagiat [SOFT] / З.Д. Усманов, О.М. Солиев, **Х.А. Худойбердиев**, П.А. Солиев // – 16.05.2018. – № 4201800378.

[61-А] **Худойбердиев, Х.А.** База данных αβ-кодирования для распознавания анаграмм [SOFT] / З.Д. Усманов, О.М. Солиев, **Х.А. Худойбердиев**, Г.М. Довудов, А.А. Косимов // – 16.05.2018. – № 4201800377.

[62-А] **Худойбердиев, Х.А.** Таджикский языковой пакет для тезауруса в Microsoft Office [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, О.М. Солиев, Г.М. Довудов // – 04.10.2012. – № 4201200237.

[63-А] **Худойбердиев, Х.А.** Таджикский языковой пакет для расстановки переносов в Microsoft Office [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, О.М. Солиев, Г.М. Довудов // – 04.10.2012. – № 4201200236.

[64-А] **Худойбердиев, Х.А.** Таджикский языковой пакет для проверки орфографии в Microsoft Office [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, О.М. Солиев, Г.М. Довудов // – 04.10.2012. – № 4201200235.

[65-А] **Худойбердиев, Х.А.** Компьютерный мультиязыковый словарь MultiGanj. [SOFT] / З.Д. Усманов, С. Холматова, **Х.А. Худойбердиев**, О.М. Солиев // – 12.11.2008. – № 077TJ.

[66-А] **Худойбердиев, Х.А.** Компьютерный русско-таджикский словарь [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, О.М. Солиев // – 29.01.2008. – № 054TJ.

[67-А] **Худойбердиев, Х.А.** Компьютерное озвучивание таджикского текста Tajik Text-to-Speech [SOFT] / **Х.А. Худойбердиев** // – 04.09.2007. – № 041TJ.

[68-А] **Худойбердиев, Х.А.** Таджикский текстовый редактор Tajik Word (TW) [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, О.М. Солиев // – 05.07.2007. – № 030TJ.

ВАЗОРАТИ МАОРИФ ВА ИЛМИ ҶУМҲУРИИ ТОҶИКИСТОН
ДОНИШКАДАИ ПОЛИТЕХНИКИИ
ДОНИШГОҲИ ТЕХНИКИИ ТОҶИКИСТОН
БА НОМИ АКАДЕМИК М.С. ОСИМӢ ДАР ШАҲРИ ХУҶАНД

УДК: 81.33 + 004.42

Бо ҳуқуқи муаллиф

ХУДОЙБЕРДИЕВ ХУРШЕД АТОХОНОВИЧ

БАЛОИҲАГИРӢ ВА АМАЛИГАРДОНИИ НИЗОМҲОИ ХУДКОРИ КОРКАРДИ МАЪЛУМОТ БО ЗАБОНИ ТОҶИКӢ

АВТОРЕФРАТИ
диссертатсия барои дарёфти дараҷаи илмии доктори илмҳои техникӣ
аз рӯйи ихтисоси 05.13.11 - Таъминоти математикӣ ва барномавии мошинҳои ҳисоббарор, муҷтамаъҳо ва шабакаҳои компютерӣ

ДУШАНБЕ – 2024

Диссертатсия дар кафедраи барномарезӣ ва низомҳои иттилоотии
Донишкадаи политехникии Донишгоҳи техникии Тоҷикистон
ба номи академик М.С.Осимӣ дар шаҳри Хуҷанд иҷро шудааст

Мушовири илмӣ: Усмонов Зафар Ҷӯраевич,
доктори илмҳои физика ва математика, профессор,
Академики АМИТ

Муқарризони расмӣ: Илолов Мамадшо Илолович,
доктори илмҳои физика ва математика, профессор,
академики АМИТ, мудири шуъбаи амсиласозии
математикии ҷараёнҳои динамикии Маркази рушди
инноватсионии илм ва технологияҳои нав

Прутсков Александр Викторович,
доктори илмҳои техникӣ, дотсент, Муассисаи
давлатии буҷетии федералии таҳсилоти олии
касбии “Донишгоҳи давлатии радиотехникии
Рязан”, профессори кафедраи «Математикаи амалӣ
ва ҳисобӣ»

Бекназарова Саида Сафибуллаевна
доктори илмҳои техникӣ, профессори кафедраи
«Технологияҳои телевизионӣ ва медиа» Донишгоҳи
технологияҳои иттилоотии Тошкент ба номи
Мухаммад ал-Хоразмӣ

Муассисаи муқарриз: Донишгоҳи миллии Тоҷикистон

Ҳимояи диссертатсия “13” сентябри соли 2024, соати 14:00 дар ҷаласаи Шурои диссертатсионии якдафъаинаи 6D.KOA-049 назди Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ, бо нишонии 734042, ш. Душанбе, хиёбони академикҳо Раҷабовҳо, 10А баргузор мегардад.

Бо диссертатсия ва автореферати он дар китобхонаи илмии Донишгоҳи техникии Тоҷикистон ба номи академик М.С.Осимӣ ва сомонаи расии донишгоҳ https://web.ttu.tj/tj/elonho/77 шиносоӣ пайдо кунед.

Автореферат санаи «____» _____ соли 2024 ирсол шудааст.

Хоҳишмандем такризҳоро нисбати автореферат дар ду нусха бо муҳри муассиса ба суроғаи зерин ирсол намоед: 734042, ш. Душанбе, хиёбони академикҳо Раҷабовҳо, 10А, тел: (+992 37) 227-37-81,
e-mail: sultonzoda.sh@mail.ru

Котиби илмии Шурои диссертатсионии
якдафъаина, номзади илмҳои техникӣ, дотсент Султонзода Ш.М.

ТАВСИФИ УМУМИИ КОР

Мубрамияти мавзӯи тадқиқот. Яке аз ҷумлаи масоили ҳалталаб дар соҳаи забоншиносии компютерӣ - ин таҳияи низоми худкори санҷиш ва таҳрири имло дар заминаи қоидаҳои забони мушаххас, бастаҳои худкори синтез ва шинохти нутқ, модули идоракунии овоз барои дастгоҳи ниҳоӣ, инчунин низомҳои тарҷумаи худкори мошинӣ ба ҳисоб меравад.

Низомҳои коркарди худкори матн бо забони табиӣ тавассути муҷтамеи барномаҳо ва замимаҳои компютерӣ амал мекунанд, ки кори онҳо ба амсилаҳои математикӣ асос ёфтааст. Таҳияи низоми худкори санҷиш ва таҳрири имло дар асоси қоидаҳои забони муайян, бастаҳои синтез ва таърифи нутқ, модулҳои идоракунии овоз барои дастгоҳҳои худкори ниҳоӣ, низомҳои тарҷумаи худкори мошинӣ вазифаҳои муҳими соҳаи забоншиносии компютерӣ маҳсуб мешаванд.

Масоили муосири амсиласозии математикии забоншиносии компютерӣ ва тарҳрезии низомҳои коркарди забонҳои табиӣ дар осору тадқиқоти муҳаққиқони хориҷӣ, ба монанди Indurkhya N., Damerau F.J., Grishman R., Hutchins W.J., Hausser R.R., Cohen M., Massaro D., Liberman A.M., Black A.W., Taylor P.A., Johnson M., Nirenburg S., Somers H.L., Wilks Y., Koehn P., Mercer R.L., Schroeder M., Zen H. ва диг. баррасӣ шудааст.

Дар осори олимони рус Е.И. Болшаков, Е. Клишинский, Д.В. Ланде, А.А. Носков, О.В.Пескова, Е.В. Ягунова, Г.Г. Белоногов, А.В. Палагин масоили коркарди худкори иттилооти матнӣ муфассал таҳқиқ шудааст. Дар тадқиқоти олимони зикршуда имкониятҳои комилан ҷадид барои рушди низомҳои дорои дурнамо, ки бо коркарди худкори матн робита доранд, пешниҳод гардиданд.

Коркарди методология, усулҳо ва амсилаҳои заминавии таҳияи низомҳои коркарди худкори иттилооти матнӣ таърихи дурудароз дорад. Дар осори илмии як зумра олимон ба монанди Д.Ш.Сулейманов, В.А.Фомичев, А.В.Анисимов, Т.В.Батура, Ф.А.Мурзин, О.Ф.Кривнова, С.В.Лесников, А.А.Марченко, Р.К.Потапова, С.Б.Потемкин, Г.Е.Кедрова, Н.Е. Сажок, А.Н. Солонина, В.Н.Сорокин, Л.А.Чистович тарзҳои таҳияи принсипҳои асосӣ, сохтори таркибӣ, технологияи таҳияи амсилаҳои амалии лингвистӣ, ки минбаъд дар низомҳои иттилоотии коркарди матнҳо ба забони табиӣ истифода шуданд, пешниҳод шудаанд.

Тибқи маълумот, шумораи зиёди корбарони компютер аз низомҳои пешрафтаи коркарди иттилооти забони табиӣ ва маҳсули нармафзор, аз ҷумла луғатҳои электронии WoқdNet, MS Office, ABBYY, Open Office, РҚОМРТ ва ОХFОҚD, низомҳои тарҷумавии YANDEX ва GOOGLE-ро истифода мебаранд, ки ҳам дар шакли бархат (онлайн) ва ҳам бидуни тамос (офлайн) амал мекунанд. Баъзе аз низомҳои номбаршуда имконияти эҷоди луғати бисёрзабонаро доранд, ки тамоми тафсирҳои имконпазир ва тафсирҳои эҳтимолии вожаҳоро дар забони мушаххас бо роҳи муқаррар кардани робитаҳои байни онҳо инъикос мекунанд.

Истифодаи фарохи технологияҳои иттилоотӣ-иртиботӣ дар Тоҷикистон таваҷҷӯҳи олимони соҳаҳои риёзӣ, технологияҳои иттилоотӣ ва забоншиносиро ба вуҷуд овардааст. Олимон таҳти роҳбарии академики АМИТ З.Д. Усмонов ба таҳқиқи як самти комилан нави технологияҳои иттилоотию иртиботӣ - забоншиносии компютерӣ оғоз карданд. Мушкилоти инкишоф додани самти нав - забоншиносии компютерӣ пешорӯи олимон ҳаллу фасли як қатор мушкилоти

ҳалталаби муҳимро гузошта шудааст. Аз ҷумла, вазифаҳои ба амсиласозии ҷумлаи дутаркиба (С.А. Зарипов), таҳияи драйверҳои миллии графикаи тоҷикӣ ва ҳалли масъалаи меъёркунии маҳсулоти чопӣ (О.М. Солиев), табдил додани низомҳои графикии хат (Л.А. Грашенко), таҳлили морфологии худкор (Г.М. Довудов), шинохти муаллифи матнҳои тоҷикӣ (А.А. Косимов ва К.С. Бахтеев), низоми коркарди худкори матн ба забони шуғнонӣ (А.Г. Ғуломсафдаров) вобастабуда ба зумраи ин гуна вазифаҳо дохил мешаванд.

Яке аз вазифаҳои бунёдии ҳар як кишвар ба таври возеҳу равшан дарк кардани ҷойгоҳи он дар раванди ҷаҳонишавӣ мебошад. Мардуми кишвар моҳияти рафтори давлатҳои муосири ҷаҳонро ба инобат гирифта, бояд интихоб кунад: аз нақши хоксоронаи истеъмолкунандаи маҳсули рушди муосири фарҳангӣ ва илмӣ-техникии давлатҳову миллатҳои дигар қаноатманд бошад ё барои ба тамоми ҷомеаи ҷаҳонӣ расонидани арзишҳои миллӣ ва ҷаҳонбинии худ чорабиниҳои фаъол андешад. Ин махсусан ба кишварҳое, ки дар марҳилаи рушд дар шароити раванди муосири технологӣ қарор доранд, рабт дорад.

Мубрамияти тадқиқоти илмӣ бо Стратегияи давлатии «Технологияҳои иттилоотӣ-коммуникатсионӣ баҳри рушди Ҷумҳурии Тоҷикистон», Барномаи давлатии рушди забони давлатӣ барои солҳои 2020-2030, Фармони Президенти Ҷумҳурии Тоҷикистон оиди эълони солҳои 2020-2040 “Бистсолаи омӯзиш ва рушди илмҳои табиатшиносӣ, дақиқ ва риёзӣ дар соҳаи илм ва маориф”, Стратегияи омӯзиш ва рушди илмҳои риёзӣ, дақиқ ва табиатшиносӣ дар соҳаи маориф ва илм то соли 2030, Барномаи мақсадноки давлатии рушди илмҳои риёзӣ, дақиқ ва табиатшиносӣ барои солҳои 2021-2025 барои давраи то соли 2030 тасдиқи худро меёбад.

Ҳадафи тадқиқот – таҳияи амсилаҳо, усулҳо ва алгоритмҳое мебошад, ки барои эҷоду таҳияи низомҳои иттилоотии коркарди худкори иттилоот бо забони тоҷикӣ барои истифодаи минбаъдаи онҳо дар низомҳои идоракунии инсону мошин дар муколамаи табиии забонӣ имкон медиҳанд.

Вазифаҳои тадқиқот. Барои ноил шудан ба мақсади зикршуда дар доираи рисолаи диссертатсионӣ вазифаҳои зерин гузошта шудаанд:

- таҳияи методология ва консепсияи назариявии коркарди худкори иттилооти матнӣ бо забони тоҷикӣ ҳамчун объекти тадқиқоти илмӣ барои муайян кардани мафҳумҳо ва истилоҳоти назариявӣ дар забоншиносии компютерӣ;

- таҳияи усулҳои ҷустуҷӯи иттилооти матнӣ барои таҳлили маълумоти озмоишӣ ва истифодаи он дар тадқиқоти илмӣ-амалӣ, луғатҳои электронӣ ва тезаурусҳои компютерӣ бо забони тоҷикӣ;

- таҳияи амсилаи пешниҳоди иттилооти матнӣ ва муҷтамеи алгоритмҳои амалисозии синтези худкори нутқ бо забони тоҷикӣ;

-таҳияи усулҳои истихроҷ, пешниҳод ва коркарди маълумот бо мақсади ташаккули унсурҳои инфиродии матн барои амалигардонии тафтиши худкори имлои матн дар забони тоҷикӣ;

- таҳияи амсилаҳо, усулҳо ва алгоритмҳои коркарди пешакии додаҳо барои ҳалли масъалаи тарҷумаи худкори матн аз забони тоҷикӣ ба русӣ;

- таҳияи муҷтамеи барномавӣ барои татбиқи ҳамаи усулҳо, амсилаҳо ва алгоритмҳои коркарди иттилоот бо забони тоҷикӣ;

- гузаронидани тадқиқоти озмоишии самаранокии низомҳои коркарди худкори иттилоот.

Объекти тадқиқот амсиласозии компютерии равандҳои ҳисоббарорӣ ва тарҳрезии нармафзор барои низоми коркарди худкори иттилоот бо забони тоҷикӣ мебошад.

Мавзӯи тадқиқот – усулҳо, амсилаҳо ва алгоритмҳои коркарди иттилоот бо забони тоҷикӣ барои тарҳрезӣ ва татбиқи луғатҳои электронӣ, синтези нутқ, тафтиши худкори имло ва тарҷумаи компютерӣ мебошад.

Соҳаи тадқиқот – таҳияи амсилаҳо, асосноккунӣ ва санҷиши усулҳои самараноки ададӣ бо истифода аз МЭҲ; истифода кардани усулҳои самарабахши ададӣ ва алгоритмҳо дар шакли муҷтамеи барномаҳои ба масъалаҳои ҳалталаб самтгузоришуда бо мақсади гузарондани озмоишҳои ҳисоббарорӣ; тадқиқоти бисёрҷонибаи мушкилоти ҳалталаби илмӣ-техникӣ бо истифода аз технологияи муосири амсиласозии математикӣ ва санҷиши ҳисоббарорӣ.

Эътимоднокии натиҷаҳо ва хулосаҳои пешниҳодшуда дар кори диссертатсия бо пешниҳоди амсилаҳои математикии элементҳои маълумоти матнӣ бо мақсади коркарди минбаъдаи онҳо асоснок карда мешаванд. Дар навбати худ, самаранокии низомҳои худкор ва модулҳои компютерии тарҳрезишуда бо интихоби дурусти додаҳои ибтидоӣ ва интихоби сарчашмаҳои маълумоти матнӣ дар вазифагузории ташаккул ва таҳияи воситаҳои математикӣ ва компютерӣ, татбиқи онҳо дар раванди коркарди худкори маълумот бо забони тоҷикӣ тасдиқ карда мешавад. Дар диссертатсия натиҷаҳои қаблан гирифтаи дигар олимон истифода мешаванд, ки бо истинодҳо қайд карда шудаанд.

Усулҳои тадқиқот. Барои ҳалли вазифаҳои дар назди таҳқиқот гузошташуда усулҳои таҳлили низомманд, омори математикӣ, асосҳои пешниҳод ва коркарди муҷтамеи додаҳо, инчунин назарияи алгоритмҳо, лингвистикаи математикӣ ва компютерӣ, синтези маълумот, амсиласозии компютерии низомҳои худкори иттилоотӣ, технологияҳои барномасозӣ ва коркарди маълумот истифода шуданд.

Навгонии илмии тадқиқот. Дар натиҷаи кори илмӣ-тадқиқотӣ ва таҳияи низомҳои худкор як қатор равишҳои методии таҳқиқ, таҳлил ва коркарди худкори иттилооти матнӣ бо забони тоҷикӣ пешниҳод шудааст:

- муқаррароти нави илмӣ-техникӣ, амсилаҳои математикӣ, усулҳо ва сохторҳои маълумот пешниҳод карда шудаанд, ки дар маҷмӯъ заминаи назариявии таҳлили низомманд ва таҳқиқи иттилооти матниро ташкил медиҳанд;

- усулҳо ва алгоритмҳои лоиҳакашии амалӣ, сохторӣ ва ба объект нигаронидашудаи низомҳои коркарди худкори иттилоот бори аввал таҳия карда шуданд;

- усулҳои нави эҷоди нармафзор барои синтези худкори нутқ бо забони тоҷикӣ, низоми худкори тафтиши имло TajSpell дар бастаи нармафзори Microsoft Office пешниҳод шудааст; модулҳои нармафзор барои тарҷумаи худкори матн аз забони тоҷикӣ ба забонҳои русӣ ва англисӣ дар шакли замимаи интернетӣ дар суроғаи tarjumon.tajlingvo.tj. дастрас аст;

- дар асоси усулҳо, амсилаҳо ва сохторҳои маълумоти таҳияшуда, алгоритмҳои нави тарҷумаи мошинӣ, корпуси мувозии компютерӣ Tajik-Russian-

Parallel Corpus и Tajik-English-Parallel Corpus дар шакли веб-замимаҳо, инчунин модулҳои нармафзор барои тарҷумаи худкори матн аз забони тоҷикӣ ба забонҳои русӣ ва англисӣ пешниҳод карда шуд;

- амсилаҳои нав, усулҳои синтези нутқ ва барномаҳои компютерӣ Computer Tajik Text Narrator, Tajik Text-to-Speech, ки самаранокии истифодаи амалии ТИК-ро барои ҳаллу фасли масъалаҳои муосири забоншиносӣ ва технологияҳои нутқ дар забони тоҷикӣ афзун мегардонад.

Ҳамаи натиҷаҳои бадастомада дар муҷтамеи нармафзори TajLINGVO татбиқ карда мешаванд, ки он имкон медиҳад:

- мӯҳлати омӯзиши забони тоҷикӣ ҳам барои истифодабарандагони Ҷумҳурии Тоҷикистон ва ҳам корбарон дар хориҷа ба таври назаррас кам гардад;

- сатҳи асоснокии тасмимҳои оиди забоншиносии компютерӣ ва мушкилоти ҳалталаби забони тоҷикӣ қабулшуда баланд гардад;

- ташаккул ва истифодаи мундариҷаи саҳеҳу дурусти забони тоҷикиро дар интернет таъмин намояд.

Аҳамияти назариявии тадқиқот дар он аст, ки дар он намунаҳо, усулҳо ва алгоритмҳои коркарди унсурҳои матн ва сигналҳои садоӣ бо забони табиӣ оварда шудаанд, ки ба омӯзиши забони тоҷикӣ мусоидат менамоянд.

Дар асоси дар ҷараёни гузаронидани тадқиқот ва додаҳои ба даст оварда, китобҳои дарсӣ бо Қарори Вазорати маориф ва илми Ҷумҳурии Тоҷикистон барои фанҳои «Балоиҳагирии низомҳои иттилоотӣ», «Манбаи додаҳо», «Амалияи барномарезӣ», «Масъалаҳо барои омӯзиши барномасозӣ» ба чоп дода шудаанд, ки дар раванди таълими бакалаврони самти таъмини барномавии технологияи иттилоотӣ истифода карда мешаванд.

Арзиши амалии тадқиқот. Дар давраи солҳои ахир низомҳои худкор ва замимаҳои нав дар муҷтамеи нармафзори TajLINGVO озмуда, такмил ва татбиқ карда шуданд. Аҳамият ва арзиши амалии муқаррароти асосии тадқиқотро таҷрибаи эҷоди нармафзор барои татбиқи луғатҳои электронӣ, тезауруси электронӣ, синтези худкори нутқ, тафтиши имло ва тарҷумаи худкор тасдиқ мекунад. Натиҷаҳои асосии тадқиқот дар Маркази илми шаҳри Хуҷанди АМИТ, дар Идораи сармоягузорӣ ва идораи амволи давлатии вилояти Суғд дар мавриди истифода қарор дода шудаанд, татбиқи амалӣ дар МДТ Донишгоҳи давлатии Хуҷанд ба номи академик Б.Ғафуров, дар кафедраи забони тоҷикии Донишгоҳи ҳуқуқ, биснез ва сиёсати Тоҷикистон, дар донишкадаи политехникии донишгоҳи техникии Тоҷикистон ба номи академик М.С.Осимӣ дар раванди таълим васеъ истифода карда мешаванд, инчунин маҷмӯи барномаҳои TajSpell дар раванди ҳуҷҷатгузорӣ дар ҶСП «Душанбе Сити Банк» ворид карда шуданд. Натиҷаҳои бадастомада ва таҷрибаи андӯхташудаи таҳияи низомҳои худкор на танҳо мӯҳлати омӯзиши забони тоҷикиро барои корбарони компютерҳо дар Тоҷикистон дар ҳалли масъалаҳои синтез, имло ва тарҷумаи нутқ ба таври назаррас кам мекунад, балки ба корбарони хориҷӣ заминаи методии омӯхтани забони тоҷикиро фароҳам меорад.

Амсилаҳо, алгоритмҳо ва нармафзорҳое, ки дар доираи тадқиқоти диссертатсионӣ таҳия шудаанд, имкон медиҳанд, ки мундариҷаи тоҷикӣ барои тадқиқот ва истифодаи амалии ҳамарӯза истифода гардад.

Муқаррароте, ки барои дифоъ пешниҳод карда мешавад:

1. Мафҳуми коркарди худкори иттилооти матнӣ бо забони тоҷикӣ ҳамчун объекти тадқиқоти илмӣ ва воситаҳои нармафзор барои таҳлили низомманд пешниҳод гардида, дар асоси онҳо мафҳумҳо ва истилоҳоти назариявӣ муайян карда шуданд.

2. Равиши илмӣ-амалии таҳияи луғатҳои электронӣ ва тезаурусҳои компютерӣ пешниҳод гардида, санҷида шуданд, ки дар доираи он намунаҳои ҳаллу фасли масъалаҳои ҷустуҷӯ ва усулҳои истифодаи намунаҳои зикршуда дар раванди татбиқи луғатҳои компютерӣ ташаккул дода шуданд.

3. Бори аввал равиши синтези худкори нутқ бо забони тоҷикӣ пешниҳод шудааст, ки дар заминаи усули пайвандкунии ҳиҷоҳо асос ёфтааст. Ҳамзамон дар забони тоҷикӣ тамоми сохторҳои имконпазири ҳиҷо ва таркиби ҳиҷоии вожаҳо ба даст оварда шуданд. Нармафзори синтези худкори нутқ дар асоси алгоритмҳои худӣ ва пойгоҳи додаҳои “ҳиҷо-садо” таҳия шудааст. Дар натиҷаи синтези нутқ дар асоси иттилооти матнии пешниҳодшуда дар шакли файли овозии рақамӣ, мавҷи садоии пурраи овозӣ ташкил карда мешавад.

4. Усулҳои нави истихроҷ, пешниҳод ва коркарди додаҳое, ки унсурҳои алоҳидаи матнро ташкил медиҳанд, таҳия шуда, тарзи нави ҳалли масъалаи имлои худкори матн дар забони тоҷикӣ пешниҳод шудааст.

5. Бори аввал масъалаи аз забони тоҷикӣ ба забони русӣ ба таври худкор тарҷума кардани матн мавриди таҳқиқ қарор гашта, амсилаҳо, усулҳо ва алгоритмҳое таҳия карда шуданд, ки барои ҳалли самараноки масъалаҳои амалӣ имкон медиҳанд.

6. Истифодаи муштараки таҳлили низомманд, равиши сохторӣ нисбати коркарди додаҳо, барномасозӣ ба объект нигаронидашуда ва забоншиносии компютерӣ барои таҳияи низомҳои коркарди худкори иттилооти матнӣ ба забони тоҷикӣ таҳқиқ карда шуд.

7. Барои татбиқи ҳамаи низомҳои худкори пешниҳодшуда маҷмӯи барномаҳои компютерии TajLINGVO тартиб дода шуда, дар ҳудуди Ҷумҳурии Тоҷикистон озмоиши таҷрибавии он гузаронида шуд.

Ҳама натиҷаҳо ва муқаррароти рисолаи ба ҳимоя пешниҳодшуда аз ҷониби муаллиф ё бо иштироки бевоситаи ӯ ба даст оварда шуда, моҳиятан нав буда, дар матбуоти кушод пурра дастрасанд. Муҷтамеи нармафзори TajLINGVO дар Маркази миллии патентии Вазорати рушди иқтисод ва савдои Ҷумҳурии Тоҷикистон ба қайд гирифта шудааст ва лоиҳаҳои таҳиякардаи муаллифӣ ҳамто надоранд. Рӯйхати онҳо дар сомонаи www.tajlingvo.tj дастрас аст.

Мутобиқати рисола ба шиносномаи ихтисоси илмӣ. Рисола аз рӯи ихтисоси 05.13.11 – «Таъминоти нармафзори математикӣ ва барномавӣ барои компютерҳо, муҷтамеъ ва шабакаҳои компютерӣ» анҷом дода шудааст. Тадқиқот натиҷаҳои комилан беназирро фаро гирифтааст, ки ба чунин соҳаҳо, чун амсиласозии математикӣ, усулҳои ададӣ ва муҷтамеи нармафзор мувофиқ ба банди 1 - амсилаҳо, усулҳо ва алгоритмҳои тарҳрезӣ ва таҳлили барномаҳо ва муҷтамеи нармафзор, табдилдиҳии муодилаи онҳо, тасдиқ ва озмоиш; 3 - амсилаҳо, усулҳо, алгоритмҳо, забонҳо ва васоити нармафзор барои ташкили ҳамкории нармафзор ва низомҳои нармафзор; 4 - низомҳои идоракунии додаҳо ва донишҳо; 5 - низомҳои нармафзор барои ҳисоббарории рамзӣ; 7 - интерфейсҳои инсонӣ ва мошинӣ; амсилаҳо, усулҳо, алгоритмҳо ва нармафзор

барои графикаи компютерӣ, визуализатсия, коркарди тасвирҳо, низомҳои воқеияти хаёлӣ, муошират бисёрраасонаии шиносномаи ихтисос мувофиқат мекунад.

Саҳми шахсии унвонҷӯ аз он иборат аст, ки гузориши масъала, амалӣ кардани онҳо, усулҳо, амсилаҳо ва алгоритмҳои коркарди иттилоот, ки ба забони тоҷикӣ тавсиф шудаанд ва дар рисола ба ҳимоя пешниҳод шудаанд, аз ҷониби ӯ мустақилона ва бо роҳбарии бевоситаи ӯ анҷом дода шудаанд.

Сатҳи эътимоднокии натиҷаҳо бо санадҳои дахлдор оид ба татбиқи амалӣ ва ба кор истифода додани низомҳои иттилоотӣ, ҳуҷҷатҳо дар бораи додани рақами бақайдгирии давлатии маҳсулоти зеҳнӣ ва захираҳои иттилоотӣ дар Маркази миллии патентии Вазорати рушди иқтисод ва савдои Ҷумҳурии Тоҷикистон тасдиқи худро ёфтаанд. Эътимоднокии натиҷаҳоро инчунин эътирофи хизматҳои муаллиф дар ин соҳаи илм аз ҷониби ташкилоту муассисаҳои гуногуни ҷумҳурӣ собит месозад. Аз ҷумла, мукофоти ба номи академик С. Умаров дар соҳаи илмҳои физикаю математика, кимиё, геология ва техникии Академияи миллии илмҳои Ҷумҳурии Тоҷикистон, 2015; Ҷоизаи давлатӣ барои олимон ва омӯзгорони фанҳои табиӣ, дақиқ ва математика, 2021; дипломи дараҷаи сеюми озмуни ҷумҳуриявии «Илм - ҷилои шукуфоӣ», номинатсияи инноватсия ва навгонӣ, 2021; Ифтихорнома ва медали «100 Чеҳраи нав»-и кишварҳои Иттиҳоди Давлатҳои Мустақил, 2022.

Тасдиқ ва татбиқи натиҷаҳои диссертатсия. Натиҷаҳои асосии рисола дар семинарҳои илмии ДПДТТХ ба номи академик М.С. Осимӣ, инчунин дар конфронсу семинарҳои ҷумҳуриявӣ ва байналмилалӣ: конфронси байналмилалӣ «Масъалаҳои муосири математика», бахшида ба 50-солагии Донишкадаи математикаи ба номи А.Ҷӯраев АМИТ, (26-27 майи соли 2023), Душанбе; Конфронси илмӣ-амалии умумирусиягӣ бо иштироки намояндагони байналмилалӣ дар мавзӯи «Мубодилаи иттилоот дар тадқиқотҳои байнисоҳавӣ II», (14 апрели 2023 с.), Академияи ҳуқуқ ва идоракунии Хадамоти федеролии иҷрои ҷазои Россия, ш. Рязан, Федератсияи Россия; конфронси байналмилалии илмӣ-амалии «Дастовардҳои нав дар соҳаи илмҳои табиатшиносӣ ва технологияҳои иттилоотӣ», Донишгоҳи славянии Россияву Тоҷикистон, (30 майи 2023), Душанбе; Конфронси ҷумҳуриявии илмӣ-амалӣ бахшида ба Рӯзи байналмилалии забони модарӣ дар мавзӯи «Забони модарӣ - сарчашмаи худшиносӣ ва маънавиёти миллӣ», Кумитаи забон ва истилоҳоти назди Ҳукумати Ҷумҳурии Тоҷикистон, (16 феврал, 2023с.), Душанбе; Конфронси ҷумҳуриявии илмӣ-амалӣ дар мавзӯи «Татбиқи технологияҳои иттилоотию иртиботӣ дар саноатикунонии кишвар», Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ (29 октябри 2022 с.), Душанбе; Конфронси ҷумҳуриявии “Низомҳои амалии иттилоотӣ: мушкилоти амсиласозӣ, воридсозӣ дар кишварҳои рӯ ба тараққӣ”, ДПХДТТ ба номи академик М.С. Осимӣ (2012 с., 2017 с., .2022 с.) ш. Хуҷанд; конфронси байналмилалии илмӣ-амалии “Илм ва технологияҳо” (26 сентябри 2022 сол), Алмаато Ҷумҳурии Қазоқистон; конфронси ҷумҳуриявии илмӣ-амалии “Мушкилоти мубрами забоншиносӣ ва лингводидактика дар шароити муосир”, филиали ДДМ ба номи М.В. Ломоносов дар шаҳри Душанбе (29 октябри 2022 с.), ш. Душанбе; Конфронси ҷумҳуриявии илмӣ-амалии «Масъалаҳои мубрами тарҷума ва забоншиносӣ дар замони

муосир», Донишкадаи забонҳои Тоҷикистон ба номи Сотим Улуғзода, (2019), Душанбе; семинари ҳарсолаи илмӣ-амалии «Технологияҳои нави иттилоотӣ дар низомҳои худкор», Донишкадаи математикаи амалии ба номи М.В. Келдиш АИР, (аз соли 2013 то соли 2019), Москва, Федератсияи Россия; Конфронси илмӣ-амалии омӯзгорон, муҳаққиқони ҷавон бахшида ба 30-солагии Истиқлолияти давлатии Ҷумҳурии Тоҷикистон, ДПХДТТ ба номи академик М.С. Осимӣ, (2019), Хуҷанд; конфронси илмӣ-амалии минтақавӣ бахшида ба 90-солагии устод Темурхон Мақсудов, Филиали ДТТ дар шаҳри Исфара, (2018), Исфара; конфронси илмӣ-амалии «Татбиқи технологияҳои иттилоотию иртиботӣ барои рушди инноватсионии Ҷумҳурии Тоҷикистон», Донишгоҳи технологии Тоҷикистон, (2017 с.), Душанбе; Конфронси ҷумҳуриявии илмӣ-амалӣ дар мавзӯи «Сифати таълим дар муассисаҳои таҳсилоти олии касбии Ҷумҳурии Тоҷикистон», бахшида ба 25-солагии Истиқлолияти давлатии Ҷумҳурии Тоҷикистон, ДПХДТТ ба номи академик М.С. Осимӣ, (20.09.2016с.), Хуҷанд; конфронси сеюми илмӣ-техникии байналмилалии «Технологияҳои семантикии кушодаи тарҳрезии низомҳои зеҳнӣ», Донишгоҳи давлатии информатика ва радиоэлектроникаи Беларус - OSTIS-2013, (21-23 феврали 2013с.), Минск, Ҷумҳурии Беларус дар шакли маърӯзаҳо пешниҳод шудаанд.

Таълифот оиди мавзӯи рисола. Дар заминаи маводи тадқиқоти диссертатсионӣ 68 таълифот ба чоп расидааст, аз ҷумла 25 (11 бе ҳаммуаллифӣ), аз ҷумла дар маҷаллаҳои тавсиянамудаи Комиссияи олии аттестатсионии назди Президенти Ҷумҳурии Тоҷикистон ва Комиссияи олии аттестатсионии Федератсияи Россия, 27 мақола дар маҷмӯаҳои байналмилалии мақолаҳо ва маҷаллаҳо, 7 китоби дарсӣ таҳти мӯҳри Вазорати маориф ва илми Ҷумҳурии Тоҷикистон. Аз ҷониби муаллиф дар Маркази патентӣ-иттилоотии назди Вазорати рушди иқтисод ва савдои Ҷумҳурии Тоҷикистон 18 шаҳодатномаи бақайдгирии давлатии захираҳои иттилоотӣ ва маҳсули зеҳнӣ гирифта шудааст.

Сохтор ва ҳаҷми рисола. Тадқиқоти диссертатсия аз 328 саҳифаи чопи компютерӣ, муқаддима, 6 боб, 19 ҷадвал, 15 расм, феҳрасти адабиёт бо 322 номгӯй ва 2 замима иборат аст.

Сипосгузорӣ. Муаллиф ба мушовири илмии худ, академики АМИТ, доктори илмҳои физика ва математика, профессор Зафар Ҷӯраевич Усмонов барои маслиҳатҳои муфид ва дастурҳои нек дар таҳияи кори илмии пешниҳодшуда миннатдориву сипоси самимӣ баён мекунад.

Вожаҳои калидӣ: забони тоҷикӣ, синтези нутқ, унсурҳои матн, басомади дучоршавӣ, луғати электронӣ, тезауруси компютерӣ, забоншиносии компютерӣ, коркарди худкори матн, санҷиши худкори имло, транслитератсия, тарҷумаи мошинӣ, усули оморӣ, тасниф, кластерсозӣ, омори математикӣ, назарияи эҳтимолият, усулҳои ададӣ, амсиласозии математикӣ, тарҳрезии низомҳои иттилоотӣ, додаи маълумотҳо, амсиласозии компютерӣ, технологияи барномасозӣ.

МУНДАРИҶАИ АСОСИИ КОР

Дар муқаддима мубрамият мавзӯи интихобшуда таъкид шуда, ҳадаф ва вазифаҳои асосии он асоснок карда шудаанд, аҳамияти илмӣ ва амалии тадқиқот нишон дода шудааст. Мундариҷаи мухтасари рисола баён карда шудааст.

Дар **боби якум «Забоншиносии компютерии забони тоҷикӣ»** масъалаҳое, ки бо муайян кардани хусусиятҳои асосии тарҳрезӣ ва мушкилоти ҳалталаби татбиқи низомҳои коркарди иттилоот ба забони тоҷикӣ алоқаманданд, баррасӣ карда мешаванд.

Дар **фасли 1.1** маълумоти умумӣ дар бораи истифодаи густурдаи васоити технологияҳои компютерӣ дар корхонаҳо ва муассисаҳои Тоҷикистон, дар бораи мушкилоти истифодаи забони тоҷикӣ дар раванди коргузорӣ дар асоси имкониятҳои муосири ТИК ва зиёдшавии таваҷҷӯҳи муҳаққиқон ба соҳаи мазкури илм, дар бораи таъсиси мактаби илмии забоншиносии компютерӣ ва математикӣ аз ҷониби академики АМИТ, профессор З.Д. Усмонов, дар бораи натиҷаҳои кори олимони ҷавони боистеъдоди ин мактаб маълумот оварда шудааст.

Дар **фасли 1.2** тавсифи натиҷаҳои таҳқиқот дар самти забоншиносии компютерии забони тоҷикӣ дар Ҷумҳурии Тоҷикистон ва дастовардҳои муштараки олимони тоҷик дар соҳаи математика, технологияҳои иттилоотӣ ва забоншиносӣ пешниҳод шудааст.

Дар **фасли 1.3** амсилаи математикӣ муайян карда шудааст, ки дар заминаи он низомҳои иттилоотии коркарди худкори додаҳо бо забони тоҷикӣ тарҳрезӣ шудаанд.

Амсилаи тарҳрезишудаи низоми TajLINGVO аз маҷмӯи технологияҳои иттилоотӣ, равандҳо, алгоритмҳо, маҷмӯи унсурҳои матнӣ, интерфейсҳо ва маҷмӯи натиҷаҳое, ки барои ташаккули тасвири рақамӣ заруранд, иборат аст. Онҳоро ба таври зерин тавсиф кардан мумкин аст:

$$TajLINGVO = \{T, P, A, TE, I, R\} \quad (1)$$

ки дар ин ҷо,

T - маҷмӯи технологияҳои иттилоотӣ;

P - маҷмӯи равандҳо дар TajLINGVO, P_i, i=1…n;

A - маҷмӯи алгоритмҳои Aj, j=1...m барои амалисозии равандҳои {Pi};

TE - маҷмӯи унсурҳои иттилооти матнӣ, ки барои коркард бо истифода аз алгоритмҳои {Aj} дар равандҳои {Pi} интиқол дода мешаванд;

I - интерфейсҳои корбар барои ворид, коркард ва соқит кардани маълумот;

R - натиҷаҳо барои интиқол ба коркард дар равандҳои {Pi}.

Дар ҷараёни таҳияи сохтори мантиқии низомҳои иттилоотӣ он ба усулҳои мушаххаси нармафзор такя мекунад. Ба ин усулҳо ва воситаҳои муосир мусоидат мекунанд, ки ба таҳиягарон имкон медиҳанд, ки низомҳоро аз аввал то ба охир амсила созанд. Ба чунин восита, масалан, Structured Analysis and Design Technique (SADT) - технологияи таҳлил ва тарҳрезии сохторшуда, методологияи муҳандисии таҳия ва муайянкунии низомҳоро дар шакли табақабандии афзуншавандаи зернизомҳо дохил кардан мумкин аст.

Сохтори низоми TajLINGVO, ки тибқи методологияи SADT пешниҳод шудааст, аз чор зернизом иборат буда, маҷмӯи захираҳои иттилоотӣ, алгоритмҳо ва нармафзор мебошад, ки равандҳои КХМ ва интерфейсҳои корбаронро идора мекунанд. Муштарак зернизомҳо муҷтамеи алгоритмҳоро барои коркарди худкори маълумоти сарчашмаҳои пешниҳодшуда амалӣ мекунанд. Натиҷаҳои коркард маҷмӯи унсурҳои матниро дар асоси сохторҳои семантикӣ ташаккул

медиҳанд, ки онҳо ба манбаи маълумот сабт шуда, ба интерфейси корбар дохил карда мешаванд.

Зернизоми *«Таъмин бо захираи иттилоотӣ»* ташаккули захираи лингвистии матнҳоро дар асоси намунаи репрезентативӣ бо назардошти маълумоти сохторҳои забонӣ ва матнӣ таъмин менамояд. Зернизом аз ҷузъҳои зерин: манбаъҳои иттилооти матнӣ, манбаъҳои гуногуни маълумот, масалан, луғатҳои электронӣ, сохтори пешниҳодшудаи унсурҳои матнӣ, ки дар натиҷаи амалисозии раванди муайяни КХМ мебошанд.

Зернизоми *“Алгоритмҳо ва васоити нармафзор”* маҷмӯи алгоритмҳое мебошад, ки дар шакли модулҳои барномавӣ, вазифаҳо ва тартиби коркарди сохтори унсурҳои матнӣ истифода мешаванд. Воситаҳои барномавӣ ба корбар имкон медиҳанд, ки раванди КХМ-ро назорат кунад.

Зернизоми *“Идоракунии равандҳои КХМ”* омодагии пешакии натиҷаҳои коркарди маълумоти воридшударо ифода мекунад. Инчунин тартиби назорат ва тафтиши натиҷаҳо барои қабули қарор аз ҷониби корбар мавҷуд аст. Агар натиҷаҳо арзишҳои гуногунро ифода кунанд, имкони коркарди дубораи маълумот пешкаш мешавад.

Зернизоми *“Интерфейси корбарӣ”* имкони ҷустуҷӯ, пешниҳод ва интихоби маълумот ва сабти натиҷаҳоро ба манбаи додаҳо пешниҳод мекунад. Инчунин, барои ҳамаҷонибаи аз назар гузаронидани натиҷаҳо ба корбар имконият дода мешавад, ки фарзияҳои графикии ҳисоботҳоро дар намуди ҷадвалҳо, нуқтаҳо, диаграммаҳо ва гистограммаҳо гиранд.

Барои таҳия намудани амсилаи низоми компютерии TajLINGVO дар заминаи сохтори мантиқии ҳосилшуда амсилаи низомманд, амсилаи равандҳои иттилоотии Р ва васоити нармафзоре, ки маҷмӯи алгоритмҳои А-ро амалӣ месозад, зарур аст. Маълумоти ҳосилшударо вобаста аз эътимоднокии натиҷаҳо ба коркарди худкори унсурҳои матнӣ, аз қабили коркарди муродифҳои компютерӣ, санҷиши имло, синтези нутқ ва тарҷумаи мошинӣ интиқол додан мумкин аст.

Сохтори низомҳои иттилоотии муосир дар забони табиӣ аз миқдори зиёди унсурҳои матнӣ иборат буда, амсилаи консептуалии пойгоҳи донишро ташкил медиҳад. Барои ноил шудан ба сохтор ҳам ба амсилаи анъанавии забони табиӣ ва ҳам усулҳои муосири амсилаҳои матнии сохторӣ такя кардан зарур меояд. Дар зер амсилаи математикии сохтори иттилоотӣ оварда шудааст:

$$FM = \{LC, SW, SS, DS, GS, CS\} \quad (2)$$

дар ин ҷо,

LC – манбаи иттилооти матнӣ барои ташаккули захираҳои забонӣ;

SW – маҷмӯи вожаҳои аз LT сохташуда;

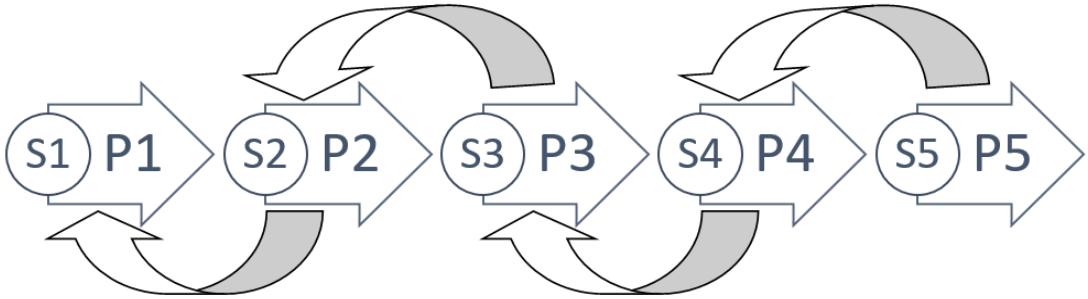
SS – маҷмӯи сохторҳои семантикӣ, ки SW-ро тавсиф мекунанд;

DS – маҷмӯи сохторҳои забонӣ, SS ба SW табдил ёфтааст;

GS – маҷмӯи ҳодисаҳои грамматикӣ, ки ба қоидаҳои грамматикии забони табиӣ асос ёфтааст;

CS - маҷмӯи сохторҳои рамз барои муаррифии DS тибқи GS.

Равандҳои ҷустуҷӯ, коркард, таҳлил ва дарки унсурҳои матнӣ пайдарпайии табдили иттилооти матниро амалӣ мекунанд WS→CS. Нақшаи таҳлили нисбатан дастраси амсилаи иттилоотии низоми TajLINGVO пешниҳод шудааст, ки дар он равандҳои коркарди иттилооти матнӣ тавассути нармафзор амалӣ карда мешаванд, расми 1.



Расми 1. Сохтори амсилаи иттилоотии TajLINGVO

Акнун функсияҳои дигареро, ки дар амсилаи иттилоотии низоми TajLINGVO истифода мешаванд, таҳлил менамоем:

P1 – эҷоди намунаҳои репрезентативӣ дар заминаи ҳуҷҷатҳои матнӣ (осори классикӣ ва муосир);

P2 – коркарди пешакии ҳуҷҷатҳои матнӣ барои таҳлили худкори забон; таклиф карда мешавад, ки дар натиҷаи муайян кардани мушкилоти омонимҳо бозгашт ба раванди P1 амалӣ шавад;

P3 - раванди интихоби маҷмӯи унсурҳои иттилооти матнӣ бо роҳи муайян кардани сохтори онҳо ва сабти он дар иттилооти матнӣ мебошад. Агар якчанд арзишҳои сохтори семантикии элементи матн пайдо шаванд, шумо метавонед ба амалиёти P2 баргардед;

P4 - раванди ташаккули сохтори унсурҳои матнӣ дар асоси қоидаҳои имлои забон; дар натиҷаи муайян кардани номувофиқатии сохторҳои муайяншуда бо қоидаҳои забони табиӣ бозгашт ба раванди P3 имконпазир аст;

P5 - раванди коркард ва идоракунии маълумот; дар натиҷаи муайян кардани тасвири рақамии номуайяни матн ба раванди P4 баргаштан мумкин аст;

S1 - сарчашмаҳои ҳуҷҷатҳои матнӣ;

S2 - махзани матн;

S3 - сохтори семантикии унсурҳои матнӣ тибқи қоидаҳои грамматикаи забони табиӣ;

S4 - маҷмӯи сохторҳои иттилоотӣ пас аз коркарди матн;

S5 - ин манбаи маълумот ва инъикоси рақамии унсури иттилооти матнӣ барои эҷоди пойгоҳи дониш аст.

Дар **боби дуюми** «Методикаи таҳлили компютерӣ ва синтези забони табиӣ» тавсифи усулҳои асосии коркарди иттилоот бо забони табиӣ, инчунин ҳаллу фасли масъалаҳои амсиласозии математикӣ ва усулҳои таҳлил ва синтези иттилоот дар забони тоҷикӣ баён шудааст.

Дар **фасли 2.1** масъалаҳои коркарди забони табиӣ баррасӣ шудаанд, вазифаҳои асосии таҳлили матн, низоми воситаҳои забонро муайян шудаанд, таҳияи сохтори хадамоти нутқ пешниҳод шудааст, васоити лингвистӣ ва

назарияи иттилоот таҳқиқ шуданд; қоидаҳои забоншиносӣ дар қисми сохтории матн вобаста ба забон, жанр ва ҳаҷми иттилооти матнӣ таҳлил карда шуданд. Истифодаи амалии усулҳои таҳлили матн, аз қабили усули графемавӣ, лексикӣ, морфологӣ, синтаксисӣ ва семантикӣ асоснок карда шудааст. Таснифи васоит ва усулҳои коркарди худкори иттилоот дар шакли додаҳои матнӣ ичро шудааст.

Дар **фасли 2.2** бо мақсади омӯзиши масъалаҳои коркарди иттилооти матнӣ бо забони тоҷикӣ усулҳои математикии З.Д. Усмонов, инчунин усулҳои умумии сохтор ва рамзгузории унсурҳои матн, аз қабили сохтори ҳиҷоии вожаҳо, рамзгузории вожаҳо ва ҷумлаҳоро таҳия шуданд. Дар заминаи усулҳои математикии коркарди иттилоот қонуниятҳои омории баъзе унсурҳои матн: ҳиҷоҳо, вожаҳо, анаграммаҳо, ҷумлаҳо таҳқиқ шуданд. Бо мақсади амалӣ гардондани синтези нутқ бо забони тоҷикӣ таркибҳои ҳиҷоии вожаҳо ташаккул дода шудаанд. Дар ин қисми кор усулҳои математикии рамзгузории унсурҳои матнӣ барои ҳалли масъалаҳои тафтиши худкори имло, тарҷумаи мошинии матн ва синтези овозӣ бо забони тоҷикӣ муайян карда шудаанд.

Дар **фасли 2.3,** бо такя ба амсилаҳои математикӣ усулҳои тафтиши имло дар додаҳои матниро омӯхта шуданд, ки дар асоси онҳо ду намуди хатогиҳои имлоӣ: когнитивӣ ва чопӣ ба вуҷуд меоянд. Хатогиҳои когнитивӣ хатоҳое мебошанд, ки ҳангоми номаълум будани имлои дурусту саҳеҳи вожа ба вуҷуд меоянд. Дар ин ҳолат талаффузи нодурусти вожаи хаттӣ ба талаффузи дурусти вожа, масалан, “*кӯмак*” бар ивази “*кумак*” якхела ё шабоҳат дорад. Хатогиҳои чопии аз сабаби хатогиҳои маърифатӣ тавлидшуда тақрибан 80%-ро ташкил медиҳанд. Ҳангоми таҳлили хусусияти хатогиҳо метавон чор гурӯҳи нисбатан маъмулро тасниф кард. Масалан, барои вожаи “*истиқлол*” чунин шакл имконпазир аст: ворид кардани як ҳарфи иловагӣ: “*исстиклол*” (хатох1), партофтани як ҳарф: “*итиклол”* (хато х2), иваз кардани як ҳарф бо дигар: “*истиклол”* (хато х3), иваз кардани ду ҳарфи ҳамшафат: “*итсиқлол”* (хато х4).

Барои ичрои супориш оиди ислоҳи се намуди аввали хатогӣ дар вожа ҳангоми ворид кардани матн усули масофаи Левенштейн васеъ истифода мешавад. Бо истифода аз ин усул формулаи математикии ҳисоб кардани масофаи байни ду сатрро муайян мекунем: w1 - вожаи дуруст навишташудаи дарозии N ва w2 - вожа дар шакли нодурусти дарозии M, бо камтарин шумораи амалиёти воридсозӣ (x1), ҳазф (x2), ивази (x3) амалиёти як ҳарф. Он гоҳ масофаи таҳрир, яъне масофаи Левенштейн D(w1,w2) бо формулаи зерин D(w1,w2)=D(N,M) ҳисоб карда мешавад, ки дар он:

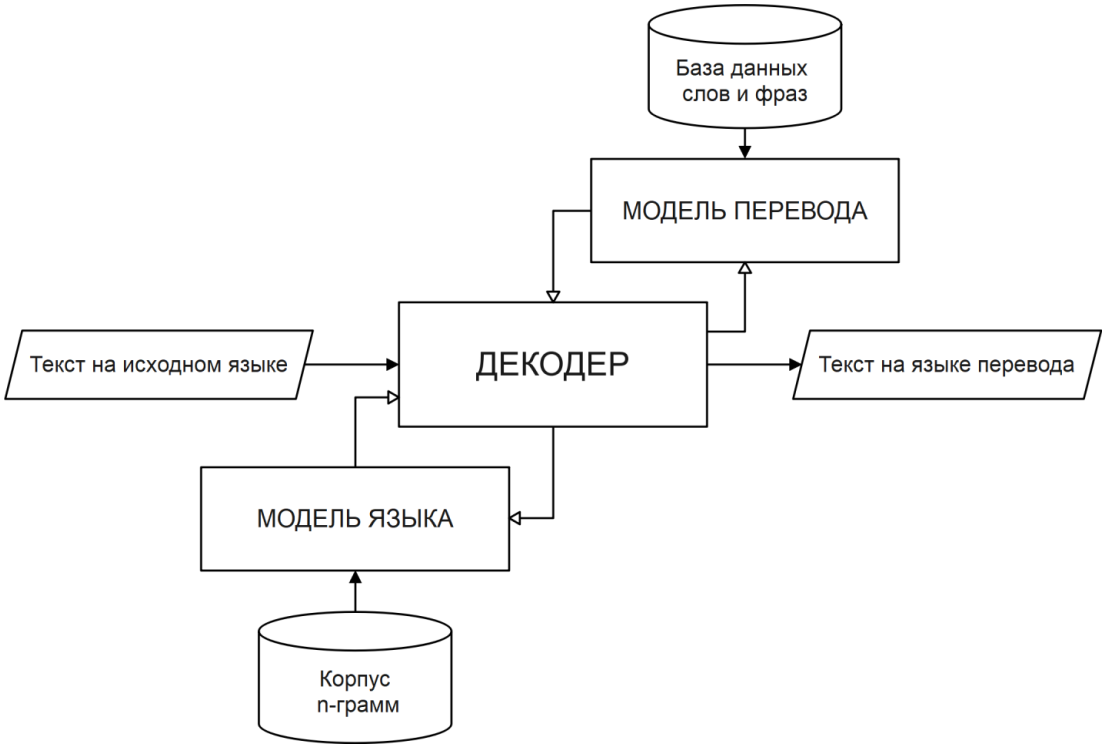
$$D(i,j) = f(x) = \begin{cases} 0, & i = 0, j = 0 \\ i, & j = 0, i > 0 \\ j, & i = 0, j > 0 \\ min\begin{Bmatrix} D(i,j-1)+1, \\ D(i-1,j)+1, \\ D(i-1,j-1)+m(w1[i],w2[j]) \end{Bmatrix}, & J > 0, i > 0 \end{cases} \quad (3)$$

ки дар он, қадам ба i - рамзи эҳтимолияти навиштани ҳарф бо хато аз вожа (x2), қадам ба j - ворид кардани як ҳарф ба вожа (x1), қадам аз рӯи ҳарду индекс рамзи иваз кардани як ҳарф дар як вожа бо дигар ҳарфи нодуруст (x3) мебошад.

Дар **фасли 2.4** алгоритмҳо ва усулҳои тарҷумаи мошинӣ дар асоси усулҳои тарҷумаи худкор муаррифӣ мешаванд, амсилаҳои тарҷумаи дуӣ (бинарӣ), тарҳсозии стратегии равиши байнизабонӣ, равиши оморӣ, таълими мошинӣ ва шабакаҳои нейрониро тавсиф шудаанд.

Алгоритми тарҷумаи мошинӣ ба қоида асосёфта. Равишҳои ибтидоии тарҷумаи мошинӣ ба қоидаҳои забоншиносӣ асос ёфтаанд, ки барои таҳлили ҷумлаи ибтидоӣ ва ба вуҷуд овардани пешниҳоди мобайнӣ дар забони мавриди ҳадаф истифода мешуданд. Усулҳои зикршуда барои тарҷумаи байни забонҳои аз оилаҳои забонҳои ба ҳам наздик тавассути луғат мувофиқанд.

Усули оморӣи тарҷумаи мошинӣ қоидаҳои анъанавии забонро истифода намебарад. Он асосан ду амсилаи имконпазир: амсилаи тарҷума ва амсилаи забониро истифода мекунад (расми 2.).



Расми 2. Алгоритми тарҷумаи мошинии оморӣ

Формулаи математикиро баррасӣ менамоем, ки эҳтимолияти шартии калонтарини P(t|s) тарҷумаи матни ибтидоӣ t-ро нисбат ба забони мавриди ҳадаф s муайян мекунад. Бо ифодаи s = s1, . . . , cj, . . . , унсурҳои sls дар матни ибтидоӣ бо дарозии ls ва натиҷаи тарҷума t = t1, . . . , ти, . . . , tltc дарозиаш lt, эҳтимолияти калонтарини таълим додани тарҷумаи мутаносиб бо истифода аз амсилаи зерини оморӣ тарҷумаи мошини математикӣ, тавре ки дар формулаи зер нишон дода шудааст, ба даст овардан мумкин аст:

$$t_{беҳт} = argmax_t\, P(t|s) = argmax_t\, P(s|t) \times P(t) \qquad (4)$$

ки дар он, P(s|t) - амсилаи тарҷума ва P(t) - амсилаи забон аст.

Тибқи формула эҳтимолияти интиқоли баръакси P(s|t)-ро ҳисоб кардан лозим аст. Дар ҳолати зиёд кардани ҷузъи амсилаи забон мо кафолати тарҷумаро бо назардошти ҳамаи қоидаҳои грамматикии забон ҳосил мекунем. Раванди ҷустуҷӯи тарҷумаи мазкури беҳтарин рамзкушоӣ номида мешавад ва онро ҷузъе иҷро мекунад, ки декодер ном дорад.

Тибқи амсилаи мо эҳтимолияти тарҷумаи баръакс p(s|t) ба вуҷуд меояд. Муҷтамеи усулҳои ҳисоб кардани он дар асоси корпуси дузабона таҳия шудааст. Ҳамчун унсурҳои корпус танҳо *вожаҳо* ё *ибораҳоро* дар ду забони мувозӣ истифода бурдан мумкин аст.

Амсилаи тарҷумае, ки ба лексика такя мекунад, барои аксари усулҳои муосири тарҷумаи мошинии оморӣ замина фароҳам меорад. Дар ин амсила баҳодиҳии баробаркунӣ бо истифода аз тақсимоти эҳтимолияти тарҷумаи лексикӣ P(ti|sai) амалӣ мешавад, ки тавассути ҳисоб кардани баробаркунии ҷуфтҳои вожаҳои мувофиқ дар корпуси таълимии дузабона муайян карда мешавад. Тариқи ҳисоби математикӣ, бо истифода аз формулаи тақзияи P(t,a|s), мо муодилаи зеринро ҳосил менамоем:

$$P(t,a|s) = \prod_{i=1}^{l_t} P(t_i|s_{ai}) P(a_i|a_{i-1}, i, l_t, l_s) \qquad (5)$$

ки дар он, *a* - вектори мавқеъҳои баробаркунӣ, *ai = j* барои вожаи *ti* дар *t*.

Амсилаҳои ба ибораҳо асосёфта ҳамчун унсурҳои нисбатан дарози тарҷума истифода мешаванд. Агар матни тарҷумашаванда аз як вожа зиёдтар бошад, ки инро ибора меноманд, амсилаи тарҷума иттилооти бештарро дар бораи мундариҷаи матн фаро мегирад, ки дар натиҷа вожаҳо аз вариантҳои мухталифи тарҷума нисбатан мувофиқ интихоб карда мешаванд. Ҳамзамон ибораи барои тарҷума пешниҳодшуда коркарди забонӣ надорад ва таҳлили дахлдор дар асоси қоидаҳои забон: морфология, синтаксис ва семантика гузаронида намешавад.

Агар матни ибтидоӣ s ба I-шумораи ибораҳо тақсим шавад, амсилаи тарҷумаи P(s|t) ба таври зерин ҳисоб карда мешавад:

$$P(s|t) = \prod_{i=1}^{I} \emptyset(s_i|t_i) d(a_i - b_{i-1} - 1) \qquad (6)$$

Амсиласозии забонӣ ҷузъи муҳими аксари вазифаҳои коркарди забони табиӣ мебошад. Дар алгоритми тарҷумаи мошинии оморӣ амсилаи забонӣ барои тавлиди тарҷума бо хусусиятҳои амсилаи логарифмӣ-хаттӣ масъул аст. Амсилаи забон дар заминаи корпуси як забон омӯхта мешавад, то ки эҳтимолияти пайдарпайии вожаҳоро арзёбӣ шавад. Усули нисбатан мувофиқ барои тавлиди амсилаи забонӣ n-грамма аст.

Амсилаҳои лингвистии n-грамма. Шартан, мо мавқеъгирии вектори мутаносиби тарҷумаи *a*-ро ҳамчун P(w1,...,wm), ки аз пайдарпайии вожаҳои w1,...,wm иборат аст, ишора мекунем. Эҳтимолияти мувофиқат бо истифода аз қоидаҳои пайванд ҳамчун ҳосилаи эҳтимолияти шартии ҳар як вожаи wi, тавре ки дар формулаи зерин нишон дода шудааст, ҳисоб карда мешавад.

$$P(w_1, \dots, w_m) = \prod_{i=1}^{m} P(w_i | w_1, \dots, w_{i-1}) \quad (7)$$

Баъдан, бо истифода аз занҷири Марков [8-А], пайдоиши тарҷумаҳои нави вожаҳои қаблиро метавон наздик ва бо n - 1 маҳдуд кард, тавре ки дар формулаи зерин нишон дода шудааст:

$$P(w_1, \dots, w_m) \approx \prod_{i=1}^{m} P\left(w_i | w_{i-(n-1)}, \dots, w_{i-1}\right) \quad (8)$$

Дар натиҷа мо амсилаи n-грамми тартиби n-ро ба даст меорем, ки он эҳтимолияти шартии вожаро бо назардошти n – 1 вожаҳои қаблӣ баҳо медиҳад. Агар қимати n=1 бошад, n-грамм униграмма номида мешавад, агар n=2 бошад, n-грамм диаграмма ва агар n=3 бошад, n-грамм триграмма номида мешавад. Эҳтимолияти шартии n-грамм бо истифода аз баҳодиҳии эҳтимолии шабоҳат тариқи ҷамъбасти шумораи басомад ба таври зерин ҳисоб карда мешавад:

$$P\left(w_i | w_{i-(n-1)}, \dots, w_{i-1}\right) = \frac{count(w_{i-(n-1)}, \dots, w_{i-1}, w_i))}{count(w_{i-(n-1)}, \dots, w_{i-1})} \quad (9)$$

Дар аксари мавридҳо ҳангоми баҳодиҳии амсилаи n-грамм дар тарҷумаи мошинӣ дарозии нисбии ибораҳои n-грамм ба се баробар аст, яъне барои омӯхтани амсила триграмма истифода мешавад.

Натиҷаҳои таҳқиқот тавассути таҳия ва воридкунии низоми тарҷумаи худкори матнҳо ба забони тоҷикӣ дар асоси амсилаи тарҷумаи мошинии дуӣ (бинарӣ) ба даст омаданд.

Барои ноил шудан ба ҳадафи зикршуда ҳалли вазифаҳои асосии зерин мусоидат мекунад:

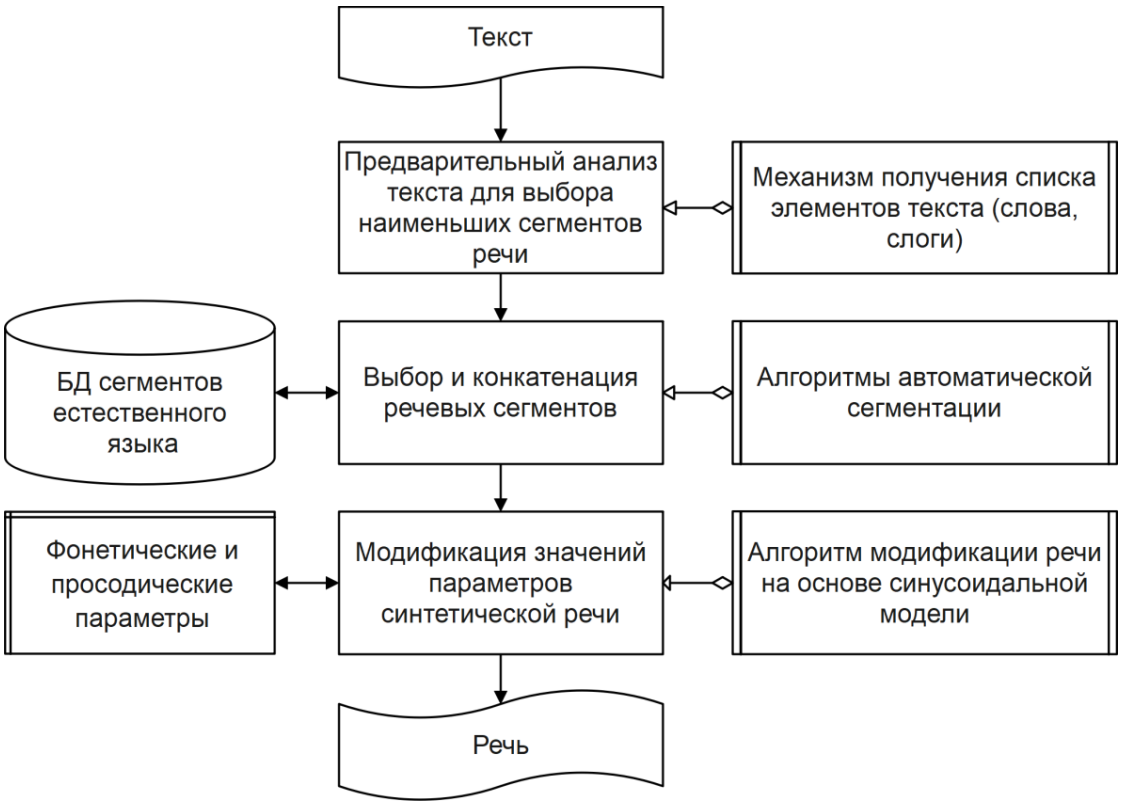
- таҳияи амсилаҳо, усулҳо ва алгоритмҳои математикӣ дар заминаи методологияи тарҷумаи дуии (бинарии) мошинии забони тоҷикӣ;

-таҳияи сохтори мувозии мантиқӣ ва воқеии захираҳо, пеш аз ҳама «русӣ-тоҷикӣ» ва «англисӣ-тоҷикӣ» барои таъмини иттилоотии низоми тарҷумаи мошинӣ;

-муайян намудани алгоритмҳои самарабахши ҷустуҷӯ, интихоб ва ба навъҳо ҷудо кардани унсурҳои матнӣ бо забони тоҷикӣ ва роҳҳои татбиқи онҳо дар модулҳои барномавӣ барои коркарди мувозии захираҳо.

Дар **фасли 2.5** натиҷаҳои таҳқиқи усулҳо ва алгоритмҳо дар низомҳои синтези худкори нутқ пешниҳод шудаанд. Дар заминаи хусусияти ташаккули овози инсон ва вежагиҳои матн бо забони тоҷикӣ тавассути таҳлили абстрактии забонӣ симои рақамӣ бо назардошти сохторҳои унсурҳои матн ва усулҳои рамзгузории нутқ ба даст оварда шудааст.

Алгоритми усули синтези конкатенативии нутқ. Пайдарҳамии унсурҳои нутқ ба қисмати коркарди сигнал ворид мешавад, ки аз пойгоҳи додаҳои унсурҳои табиии нутқ татбиқи мувофиқи овозии унсурҳоро интихоб мекунад ва онҳоро ба сигнали бефосилаи нутқ муттаҳид мекунад (расми 3).



Расми 3. Алгоритми синтези нутқ дар асоси конкатенатсияи элементҳо

Таҳлили пешакии матн. Барои ҷудо кардани қисмҳои хурдтарини нутқ механизми истихроҷи рӯйхати унсурҳои матнӣ истифода мешавад. Унсурҳои нисбатан муҳими таҷзияи матн вожаҳо ва ҳиҷоҳо мебошанд. Барои таҳлили матн ду усули асосӣ: оморӣ ва луғат истифода мешаванд. Барои амсилаҳое, ки лар асоси луғат таҳия шудаанд, луғати қаблан муайяншуда бояд дастрас бошад. Ҳамзамон варианти алгоритми дорои мувофиқати калонтарин вобаста ба самти коркарди матн қайд карда мешавад. Варианти дуюми алгоритми луғат алгоритме мебошад, ки тақсимотро бо вожаҳои камтарин пайдо мекунад.

Барои амсилаҳои ба луғат асосёфта рӯйхати вожаҳо пешниҳод карда мешавад, ки ба ҳар кадоме аз онҳо бо арзёбии эҳтимолияти он, ки ин вожаи воқеӣ аст, мувофиқат карда шудааст. Бигзор $W = \left\{\{w_i, g(w_i)\}_{i=1,\dots,n}\right\}$ рӯйхате бошад, ки номзад ба як вожа аст, инчунин вазифаҳои сифати онро дар бар мегирад. Калонтарин алгоритми мутобиқати мустақими матни T-ро барои тавлиди вожаи беҳтарини ҷорӣ чанд маротиба бо T=t* барои ҳар як марҳила метавон ба таври зерин муайян кард:

$\{w^*, t^*\} = argmax_{wt=T} g(w)$, ки дар он шарти {w,g(w)}εW гузошта мешавад.

Алгоритми таҷзияи роҳи кӯтоҳтарин он фарзияро истифода мебарад, ки таҷзияи дуруст бояд дарозии ҳама вожаҳоро ҳадди аксар афзоиш диҳад ё шумораи умумии вожаҳоро кам кунад. Барои як ҷумлаи S аз m рамз {c1,c2,…cm} - беҳтарин ҷумлаи ба қисмҳо тақсимшуда S* аз n* вожаҳо мебошад.

$$S^* = argmin_{w_1 \dots w_i \dots w_n = T}\ (n) \qquad (10)$$

Ин масъалаи мувозинат ба масъалаи ҷустуҷӯи роҳи кӯтоҳтарин барои графи бемарҳилаи самтдор табдил меёбад.

Интихоб ва пайвасти қисмҳои нутқ. Барои татбиқи марҳилаи мазкур додаи маълумоти унсурҳои забони табииро ташаккул додан зарур аст. Қисмҳои дар боло зикршуда, ки садоҳои нутқро ба вуҷуд меоранд, ба монанди вожаҳо, ҳиҷоҳо ё фонемаҳо дар шакли мазкур, якҷоя қисми садоиро ташкил медиҳанд. Бар замми самаранокии баланд амали худкор кафолат медиҳад, ки мувофиқати ҷойгиркунии сарҳадҳои ҷузъ дар доираи маънои он дар сигнали нутқ таъмин шавад.

Алгоритми таҷзияи худкор имкон медиҳад, ки амсилаҳои андозаи баробари вақт истифода шаванд. Ҳангоми ҳисоб кардани эҳтимолияти Pj, ки ҳолати ҷузъи qj ба мушоҳидаҳо дар лаҳзаи вақти ρ аз t − τ + 1 то t мувофиқат мекунад, ба формулаи зерин мутобиқ аст:

$$P_{j_{p+1}}(m,\tau) = \sum_{l \in L_m} P_{j_p}(l,\tau) b_{j_l}(O_{p+1}) \quad t - \tau + 1 \le p < t \qquad (11)$$

ки дар ин ҷо,

t - мавқеи ҷорӣ дар рӯйхати додашуда,

τ - дарозии ҷузъи эҳтимолӣ,

ρ – индекси вақт, ки дар рекурсияи матни дохилӣ истифода мешавад. bji(Oρ+1) эҳтимолияти он аст, ки мушоҳидаи O дар лаҳзаи вақти ρ + 1 тавассути тақсимоти l-уми амсилаи j-унсурдор ба вуҷуд меояд. Ба ибораи дигар, Pjp+1(m, τ) - ин эҳтимолияти он аст, ки векторҳои мушоҳидаи Ot-τ+1,...,Ot аз тақсимоти 1,..., M, яъне пойгоҳи додаҳои унсурҳо ба вуҷуд меоянд.

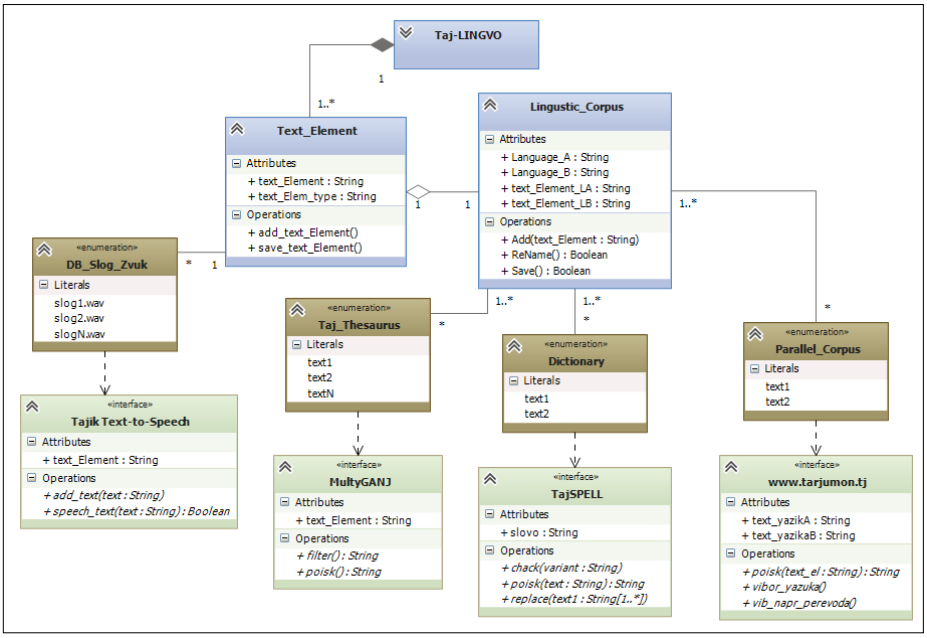
Барои ҳалли масъалаи синтези нутқ механизми нисбатан мутавозин таҳқиқ карда шуд, ки аз маҷмӯи марҳилаҳо: таҳлили пешакии матн; интихоб ва пайвасткунии ҷузъҳои нутқи табиии забон аз пойгоҳи додаҳо дар асоси алгоритми худкори таҷзия; тағйир додани қиматҳои ченакҳои фонетикӣ ва просодикии нутқи сунъӣ бо истифода аз амсилаи синусоидалии синтези нутқ иборат аст. Ҳамин тариқ, натиҷаҳои таҳқиқотро бевосита дар тарҳрезӣ ва татбиқи механизми синтези нутқ дар забони тоҷикӣ истифода кардан мумкин аст, ки дар боби шашуми рисола муфассал тавсиф дода шудааст.

Дар **боби сеюм** «Амсиласозии ба объект нигаронидашудаи низомҳои коркарди матни забони табиӣ» амсиласозии компютерии таҳияи низомҳои коркарди худкори иттилоот бо забони табиӣ бо назардошти равиши ба объект нигаронидашуда баррасӣ карда шудааст. Дар асоси маҷмӯи диаграммаҳои забони UML тарҳи намунавии низоми иттилоотии коркарди маълумоти матнӣ бо забони тоҷикӣ бо назардошти амсилаи амал, амали мутақобила, сохтор ва воқеият таҳия карда шуд.

Дар **фасли 3.1** воситаҳои муосири амсиласозии равандҳо тавсиф шудааст. Асоси методологии амсиласозии низомҳои иттилоотӣ муайян ва таҳлили намудани робитаи умумии байниҳамдигарии объектҳои ба ҳам алоқаманд, инчунин ноил шудан ба ҳадафҳои умумӣ аз ҷониби ҳамаи гурӯҳҳои корӣ

мебошад. Архитектураи ҳамгироишудаи КХМ тавассути маҷмӯи технологияҳои иттилоотӣ, равандҳо, алгоритмҳо, муҳтамеи усулҳои коркарди матн, воситаҳо, интерфейсҳо ва маҷмӯи равандҳо амалӣ карда мешавад. Амсилаи пешниҳодшуда ба сифати муаррифии рақамии забони тоҷикӣ мебошад.

Дар шароити муосир барои амсиласозии нармафзор ва низомҳои иттилоотӣ усулҳои меъёрӣ ва забонҳои амсиласозии функсионалӣ, аз қабили IDEF, DFD, UML истифода мешаванд. Дар заминаи диаграммаи синфии низоми TajLINGVO сохтори омории муайян кардани объектҳои умумии низоми иттилоотӣ ва робитаҳои мантиқии онҳо ба роҳ монда шуд. Ғайр аз ин, ҳодисаҳои эҳтимолии пайдо шудани унсури матн ҳангоми коркард мавриди таҳқиқ қарор гирифтанд (расми 4.).



Расми 4. Диаграммаи синфҳои низоми TajLINGVO

Дар **фаслҳои 3.2-3.5** усулҳои амсиласозии рафтор, таъсири мутақобила ва амсилаи консептуалии низоми коркарди худкори иттилоот пешниҳод мешаванд, ки бо таҳияи диаграммаи забони UML алоқаманд аст. Амалҳои асосие, ки низоми иттилоотӣ иҷро мекунад: ташаккули маҷмӯавии унсурҳои матн, идоракунии равандҳои коркарди иттилоот, коркарди худкори маълумоти матнӣ, синтези нутқ, коркарди тезаурус, санҷиши имло ва тарҷумаи мошинӣ муайян карда шуданд.

Дар **фасли 3.5** амсилаи физикии низоми иттилоотӣ бо мақсади ҷамъбасти имкониятҳои татбиқии нармафзор ва воситаҳои техникӣ муайян карда шудааст. Робитаҳои мантиқӣ байни сохтори омории низоми иттилоотӣ, ҷузъҳои нармафзор ва воҳидҳои воситаҳои техникӣ дар рафти татбиқи низоми иттилоотӣ муқаррар карда шудаанд.

Дар маҷмӯъ амсилаи компютерии низоми иттилоотии намунавии коркарди иттилооти матнӣ бо забони тоҷикӣ бо назардошти амсилаи рафтор, таъсири мутақобила, сохтори оморӣ ва васоити физикӣ таҳия гардид.

Дар **боби чорум** «Тарҳрезӣ, таҳия ва татбиқи санҷиши худкори имлои забони тоҷикӣ» масоили тарҳрезӣ, коркард ва татбиқи низоми худкори санҷиши имлои матни забони тоҷикӣ мавриди таҳқиқ қарор гирифт.

Дар **фаслҳои 4.1-4.3** натиҷаҳои тадқиқот ва татбиқи луғатҳои электронӣ, тезауруси компютерии забони тоҷикӣ ва низоми худкори табдил додани ҳуруф дар матн ба ҳуруфи стандартии забони тоҷикӣ пешниҳод шудаанд.

Дар асоси амсилаҳо ва усулҳои математикӣ барои ҳалли масъалаи дарёфт ва ислоҳи хатоҳои имлоӣ дар иттилооти матнӣ дар забони тоҷикӣ алгоритмҳои зерин таҳия шудаанд:

- алгоритми транслитератсияи матн ба алифбои меъёрӣ;
- алгоритми дарёфт кардани хатогиҳои имлоӣ;
- алгоритми ислоҳи хатогиҳо;
- алгоритми санҷиши имло.

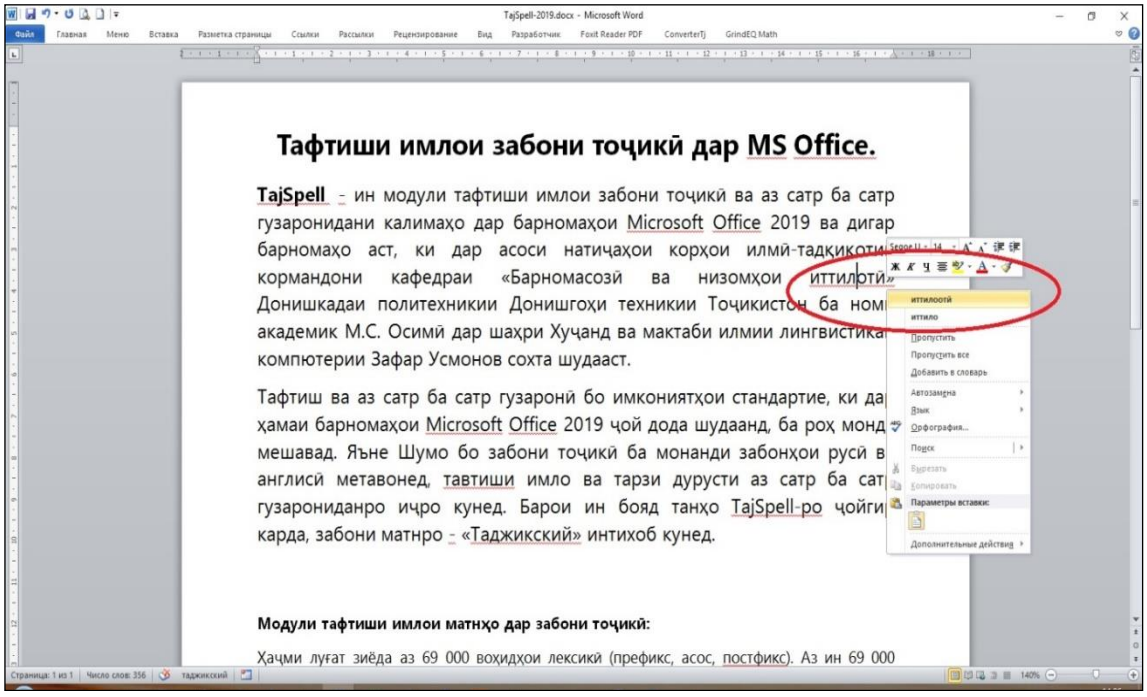
Дар **фасли 4.4** алгоритми санҷиши имлои забони тоҷикӣ бо имконияти дарёфт кардани хатоҳои имлоӣ ва ислоҳи онҳо пешниҳод шудааст. Тартиби тафтиш ба се шарт такя мекунад: дар вожа як ҳарф гум шудааст, ду ҳарфи шафат ҷои худро дар вожа иваз кардаанд ва дар вожа ҳарфи изофӣ мавҷуд аст.

Агар яке аз се шарти зикршуда дарёфт ва ислоҳ карда шавад (шарт W=S[I]), он гоҳ мо вожаро дар рӯйхати вожаҳои эҳтимол дуруст C[J] нигоҳ медорем. Рӯйхати натиҷаҳоро бо тартиби афзоиши басомади пайдоиши онҳо мураттаб мекунем.

Ҳафт унсури аввалро аз рӯйхати вожаҳои дуруст хориҷ мекунем. Агар тартиби тафтиш то охири рӯйхат S[I] бо вожаи W мувофиқ нашавад, он гоҳ муайян карда мешавад, ки “вожаи мувофиқ ёфт нашуд”.

Тибқи алгоритми пешниҳодшуда вожаи хато бо ҳар як вожаи луғат муқоиса карда мешавад. Усули асосии ба алгоритм мутобиқшуда аз он иборат аст, ки рӯйхати қаблан ташаккулдодашудаи вожаҳо S[1...i] метавонад ба рӯйхати вожаҳои ҷамъбастӣ C[1..j] табдил дода шавад. Ниҳоят, раванд маънои вожаи дилхоҳро, ки корбар метавонад онро интихоб кунад, бармегардонад. Тартиби тафтиш аз ҷониби хэш-амал иҷро мешавад, ки захираи луғатро, яъне хэш-ҷадвалро коркард мекунад.

Дар фасли 4.5. дар заминаи натиҷаҳои ҳосилшуда тавсифи модули TajSpell бо имконияти ислоҳи матни тоҷикӣ, санҷиши имло, гузариш аз сатр ба сатр ва тезауруси забони тоҷикӣ оварда шудааст. Ҳамин тариқ, матнҳои тоҷикиро дар бастаи замимаҳои MS Office, дар рамзгузории меъёри UNICODE тафтиш кардан мумкин аст (расми 5.).



Расми 5. Санҷиши имлои TajSpell дар MS Word

Модули TajSpell дар барномаҳои Microsoft Office ҳуруфи забони тоҷикиро пурра дастгирӣ мекунад ва дар асоси мубодила бо модули санҷиши имло дар рамзбандии UNICODE амалӣ мешавад.

Дар **боби панҷуми** «Тарҳрезӣ, таҳия ва ворид намудани тарҷумони худкори тоҷикӣ» тавсифи тарҳрезӣ, таҳия ва ворид намудани тарҷумони худкори тоҷикӣ оварда шудааст.

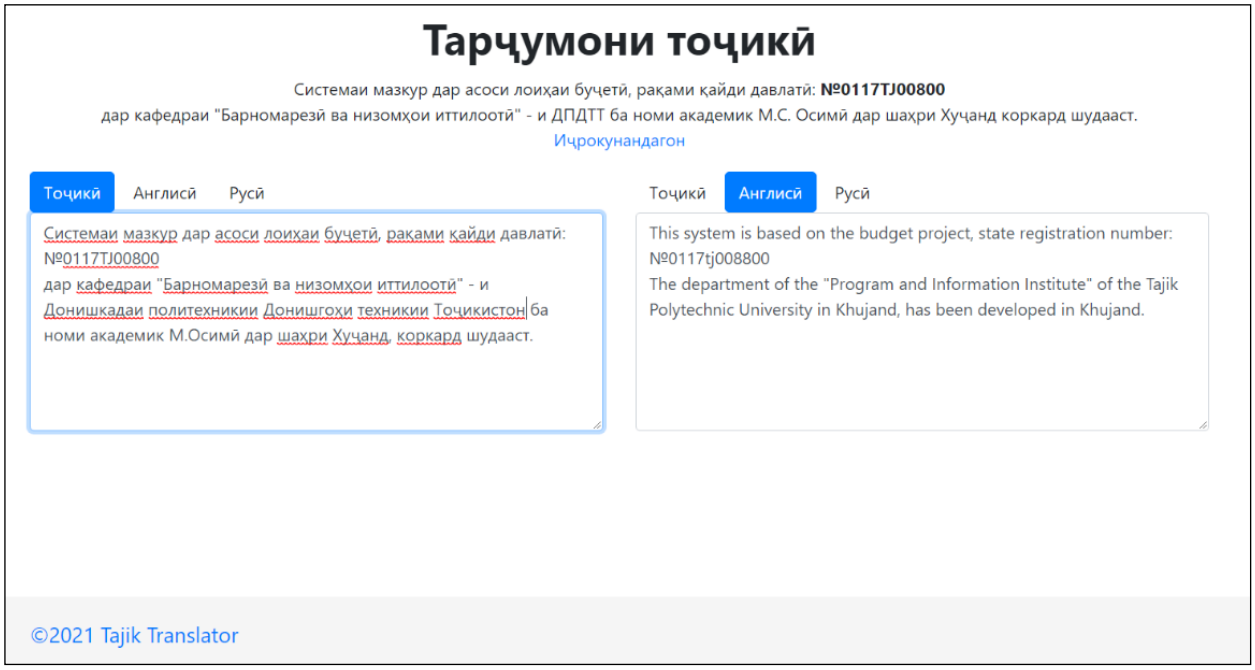
Дар **фасли 5.1** натиҷаҳои таҳқиқи мушкилоти тарҷумаи матнҳо аз забонҳои гуногун ба забони тоҷикӣ ва баръакс дар бар гирифта, инчунин мушкилоти тарҷумаи матни адабӣ ва вобастагии он аз тарҷумаи мошинӣ дар заминаи технологияи Google муайян карда шудаанд.

Дар **фаслҳои 5.2-5.4** масъалаҳои таҳияи низоми худкор, алгоритмҳои транслитератсияи худкори ҳуруф, татбиқи усули тарҷумаи оморӣ ва сохтори мантиқии тарҷумаи мошинӣ дар намунаи забони тоҷикӣ баррасӣ шуданд.

Дар **фасли 5.5** мутаносибан тавсифи низоми иттилоотӣ барои тарҷумаи мошинии матн аз забони тоҷикӣ ба русӣ ва англисӣ оварда шудааст. Ҳалли ин вазифа ба ду манбаи маълумот асос меёбад - корпуси мувозии Taj-Rus-Corp и Taj-Eng-Corp. Дар марҳилаи якум шумораи умумии унсурҳои захиравӣ ба таври зерин муайян карда мешавад:

- тоҷикӣ-русӣ – 42000, аз ҷумла зиёда аз 27 000 вожа;
- русӣ-тоҷикӣ – 68 000, аз ҷумла захираи луғавӣ беш аз 54 000;
- англисӣ-тоҷикӣ – 12 000, аз ҷумла захираи луғавӣ беш аз 5 000;
- тоҷикӣ-англисӣ – 24 000, аз ҷумла зиёда аз 11 000 вожа.

Дар асоси низоми омории тарҷумаи мошинӣ ва низоми тарҷумаи мошинии ба қоида асосёфта амсилаи тарҷумони забони тоҷикӣ таҳия карда шуд. Барои таъмин намудани тарҷумаи мошинии матн ба забони тоҷикӣ низоми иттилоотӣ дар шакли Web-барнома таҳия карда шуд (расми 6).



Расми 6. Веб-замимаи тарҷумони тоҷик - www.tarjumon.tj

Лоиҳа дар Интернет дар www.tarjumon.tj барои тарҷумаи бархати иттилооти матнӣ аз забони тоҷикӣ ба русӣ, англисӣ ва баръакс дастрас аст.

Дар **боби шашум** «Тарҳрезӣ, таҳия ва татбиқи синтези компютерии нутқи тоҷикӣ аз рӯи матн» масъалаи таҳияи вазифаи амсиласозии математикӣ ва татбиқи компютерии синтези нутқи тоҷикӣ дар асоси матни пешниҳодшуда мавриди баррасӣ қарор дода шудааст.

Дар **фасли 6.1** вазифаи таҳлили маълумоти матниро дар заминаи сохторҳои мухталифи ҳиҷоӣ амалӣ шудааст. Аз натиҷаҳои дар ҷадвали 1 овардашуда маълум мешавад, ки ҳаҷми (шумораи ҳарфҳо) 1 ва 14 ҳаҷми ҳадди ақал ва ҳадди аксари сохтори вожа мебошад.

Дар матни коркардшуда вожаи дорои зиёда аз 14 ҳарф дарёфт нашуд, ҳарчанд ин гуна вожаҳо дар забони тоҷикӣ низ мавҷуданд.

Тибқи қонунияти омории маълумотҳои матнӣ дар забони тоҷикӣ ҳангоми коркарди маълумотҳои матнӣ ҳамагӣ 274 сохтори гуногуни вожаҳо (унсур, садонок – 1, ҳамсадо – 0) дар ҳаҷми 1 724 472 вожа муайян карда шудааст.

Ҷадвали 1. Омори вожаҳои тоҷикӣ аз рӯи ҳарфҳо

Дарозии вожа	1	2	3	4	5	6	7
Душчоршавӣ, бо %	0,87	16,14	10,94	11,32	16,95	13,95	12,81

Дарозии вожа	8	9	10	11	12	13	14
Дучоршавӣ, бо %	8,88	4,98	2,92	1,00	0,57	0,10	0,02

Муқаррар карда шудааст, ки 8 воҳид 50% ва 23 унсур 75% матнҳои тоҷикиро фаро мегирад. Инчунин муайян гардид, ки 51 унсур 90% ва 76 қисм 95% матнҳои тоҷикиро фаро мегирад (ҷадв. 2).

Ҷадвали 2. Басомади дучоршавии вожаҳо дар шакли таркиби ҳиҷоӣ

№	$W^*_{0,1}$	%	№	$W^*_{0,1}$	%	№	$W^*_{0,1}$	%
1	01	11,006	9	010010	3,684	17	1010	1,192
2	010	8,849	10	0101010	3,258	18	01001010	1,142
3	01010	6,781	11	0100	2,799	19	010100	1,087
4	01001	5,486	12	01010101	1,735	20	01001011	1,053
5	10	5,096	13	01011	1,711	21	100	0,986
6	0101	5,066	14	1001	1,280	22	10101	0,960
7	010101	4,773	15	010011	1,226	23	10010	0,957
8	0100101	3,787	16	0101001	1,218			

Дар асоси тақсимоти мувофиқи 274 воҳид дар забони тоҷикӣ ҳамагӣ 9 таркиби гуногуни ҳиҷо муайян карда шудааст, ки 6-тои онҳо ба қоидаҳои забони тоҷикӣ мувофиқанд: “1”, “10”, “01”, “010”, "100", "0100" (ҷадв. 3)

Ҷадвали 3. Басомади пайдоиши таркибҳои ҳиҷо (ба ҳисоби фоиз)

Ҳиҷоҳо	1	10	01	100	010	0100	001	0010	00100
Дучоршавӣ	8.10	5.74	56.56	0.78	25.75	2.95	0,05	0,06	0,01

Аз рӯи сохти таркибӣ алгоритми таҷзияи вожа ба вожа бо назардошти 6 қолаби ҳиҷоӣ тартиб дода шудааст. Барномаи компютерӣ дар заминаи алгоритми таҳияшуда барои таҳқиқи омории ҳиҷоҳои гуногуни забони тоҷикӣ истифода шуд. Дар 3800 саҳифаи тасодуфан интихобшуда 3259 ҳиҷои гуногуни таркибӣ муайян карда шуд.

Барои таъмини шаффофияти натиҷаҳои бадастомада қонуниятҳои омории таркиби ҳиҷои забони тоҷикӣ дар сохтори вожаҳо, инчунин вожаҳои истифодашаванда таҳқиқ карда шуданд.

Дар фаслҳои **6.2-6.4** алгоритмҳои зерин дар асоси амсилаҳои математикӣ ва усулҳои махсуси барномасозӣ таҳия шудаанд:

1. Алгоритми талаффузи вожаҳо.
2. Алгоритми талаффузи ададҳо ва аломатҳо.
3. Алгоритмҳои талаффузи бе зада ва заданоки матн.
4. Алгоритми талаффузи морфемаи вожа.
5. Алгоритми талаффузи матни тоҷикии дорои вожаҳои русӣ.

Дар заминаи натиҷаҳои бадастомада низоми худкори дорои имконияти синтези нутқ ба забони тоҷикӣ “Tajik Text-to-Speech”, “ровии худкор” дар СО Windows «Tajik Text Narrator», инчунин модули корбари бархати www. .tajlingvo.tj\talaffuz таҳия карда шудаанд.

Дар **фасли 6.5** масъалаҳои шинохти нутқ дар забони тоҷикӣ дар асоси таҳлили муқоисавии низоми синтези нутқи таҳияшуда мавриди таҳқиқ қарор гирифтанд. Таҳлили муқоисавии низоми шинохти нутқи шифоҳӣ муайян намуд, ки барои ноил шудан ба ҳадафи шинохти нутқ дар забони тоҷикӣ имкониятҳои алгоритми мубодилаи динамикии ченкунии вақт, алгоритми шинохти ҳиҷоҳои нутқи забони тоҷикӣ дар тағйирёбии фазою вақт истифода мешаванд. Дар

заминаи нишондодҳои зикршуда дар асоси таҳлили сохтори ҳиҷоии вожаҳо алгоритми шинохти нутқ дар забони тоҷикӣ пешниҳод шудааст, ки он дар тадқиқотҳои минбаъда истифода мешавад.

Натиҷаҳои илмие, ки дар доираи коркарди худкори сатҳи пасти синтези нутқ дар забони тоҷикӣ ба даст оварда шудаанд дар давраи минбаъда ҳамчун пойгоҳ барои ҳалли мушкилоти шинохти нутқ дар забони тоҷикӣ истифода мешаванд ва вобаста ба ин мушкилоти асосии ҳаллу фасли вазифаи шинохти худкори нутқ бо забони тоҷикӣ таҳлил карда шуд.

ХУЛОСА

1. Дар асоси таҳлили дастовардҳои соҳаи забоншиносии компютерӣ, натиҷаҳои тадқиқотҳои илмӣ дар кишварҳои хориҷӣ ва Ҷумҳурии Тоҷикистон, таҷрибаҳо ва таҳқиқоти назариявии муаллиф **вазифаҳои тадқиқот муқаррар гардиданд**, ки аз тарҳрезӣ, таҳия ва татбиқи низомҳои худкори коркарди иттилоот ба забони тоҷикӣ [1-М]-[3-М], [8-М], [26-М], [29-М], [32-М] иборатанд.

2. Барои ҳалли масъалаи тарҳрезии низомҳои иттилоотии коркарди иттилоот бо забони тоҷикӣ дар шароити тақвияти робитаҳои байнидавлатии сиёсӣ, фарҳангӣ ва илмӣ, омилҳои истифодаи забони давлатӣ дар коргузорӣ равиши ба объект нигаронидашуда **пешниҳод шудааст**. Мундариҷаи равиши ба объект нигаронидашуда - таҳлили унсурҳои матн ва нутқ ба сифати объекти идоракунӣ; амсиласозии равандҳои рафтор ва таъсири мутақобилаи унсурҳои матн; амсилаи оморӣ ва консептуалии низоми коркарди иттилоот; ташаккули амсилаи физикии низоми усулҳои коркарди иттилоот мебошад [4-М], [28-М], [39-М], [61-М], [65-М].

3. Амсилаҳо, усулҳо ва алгоритмҳои ҷадиди коркарди иттилоот **таҳия гардида,** дар заминаи онҳо васоити нави ташкили додаи маълумот ва барномасозӣ барои таҳлили додаҳои матнӣ бо забони тоҷикӣ татбиқ шудаанд [6-М], [16-М], [22-М], [38-М], [65-М], [68-М].

4. Бо фарогирии усулҳо мушкилоти тарҳрезӣ, таҳия ва татбиқи нармафзори амалӣ барои ҳалли масъалаҳои санҷиши худкори имло, тарҷумаи мошинӣ ва синтези нутқ бо забони тоҷикӣ аз ҷиҳати назариявӣ асоснок ва аз ҷиҳати амалӣ **омӯхта шудааст** [21-М], [60-М].

5. Усулҳои ба объект нигаронидашудаи таҳияи низомҳои худкоршудаи иттилоотӣ **пешниҳод шудаанд**, ки аз муҷтамеи амсилаҳо, усулҳо, алгоритмҳо ва амалиёте иборат аст, ки дар масъалаҳои амсиласозии равандҳои коркарди иттилоот бо забони табиӣ **татбиқ карда шуданд** [1-М], [6-М], [35-М].

6. Дар натиҷаи таҳлили заминаҳои методии тарҳреҳии низомҳои коркарди худкори иттилоот усулҳои амсиласозии компютерии равандҳо ва таҳлили омории унсурҳои матн, инчунин алгоритмҳо ва барномаҳои худкоркунии равандҳои татбиқи онҳо **асоснок карда шуданд** [7-М], [17-М], [24-М], [30-М].

7. Дар рисола асоси усули ҷамъоварӣ, таҳлил ва коркарди самарабахши иттилооти матнӣ бо забони тоҷикӣ **муқаррар карда шудааст**. Амсилаи бисёрсатҳаи равандҳои ба даст овардани симои рақамии матн **пешниҳод карда шудаас**т, ки дар заминаи он вежагиҳои асосӣ **муайян карда шуда**, таснифоти

унсурҳои матнии он **тартиб дода шуд** [14-М], [15-М], [18-М], [25-М], [33-М], [53-М], [58-М], [59-М].

8. Барои тарҳрезӣ, таҳия ва татбиқи вазифаи худкори санҷиши имлои матн бо забони тоҷикӣ механизмҳо, расмиёт ва алгоритмҳои коркарди маълумоти матнӣ **таҳия карда шудаанд**. Муҷтамеи низомҳои худкори компютерӣ татбиқ гардид, ки он аз луғатҳои электронӣ, тезауруси компютерӣ, табдил додани ҳуруфҳои ғайримеъёрӣ ба рамзгузории Unicode, модуль TajSpell бо имконияти ислоҳи имло, ҷойгиркунии гузаронидани вожаҳо аз сатр ба сатр, тезауруси забони тоҷикӣ дар муҷтамеи нармафзори MS Office иборат аст [12-М], [13-М], [19-М], [37-М], [54-М], [62-М]-[64-М], [66-М].

9. Барои ҳаллу фасли масъалаи тартиб додани тарҷумони худкори тоҷикӣ амсилаҳои математикии сохторҳои мантиқии артефактҳо, усулҳои тарҷумаи мошинӣ ва алгоритмҳои амалисозии онҳо **асоснок карда шудаанд**. Барои аз алифбои лотинӣ ва кириллӣ ба кириликии тоҷикӣ баргардонидани матнҳо низом **ташаккул дода шуд**. Барои таъмини иттилоотии низоми тарҷумаи мошинӣ корпусҳои мувозии тоҷикӣ-русӣ ва тоҷикӣ-англисӣ **ташкил карда шудаанд**. Дар заминаи технологияи Google муҷтамеи барномаҳои дуҷонибаи худкори тарҷумаи матн дар шакли Web-замима бо имконияти тарҷумаи бархати иттилооти матнӣ аз тоҷикӣ ба русӣ ва англисӣ **таҳия карда шуд** [4-М], [5-М], [11-М], [27-М], [32-М], [34-М], [36-М], [52-М], [55-М], [56-М].

10. Якумин бор низоми синтези худкори нутқ бо забони тоҷикӣ, ки дар заминаи усули пайвастани ҳиҷоҳо асос ёфтааст, **тарҳрезӣ шуд**. Амсилаҳои математикии сохтори ҳиҷоии вожаҳо дар забони тоҷикӣ **пешниҳод гардида**, дар асоси онҳо гуногунии ҳиҷоҳо ба даст оварда шуд, пойгоҳи ҳиҷо-овоз **ташаккул дода шуд**. Як қатор алгоритмҳои овоздиҳии матн бо забони тоҷикӣ бо назардошти ҳиҷоҳо, морфемаҳо, ададҳо, аломатҳои китобатӣ ва вожаҳои иқтибосии русӣ тартиб дода шудаанд. Натиҷаҳои бадастомада дар барномаҳои татбиқии овоҳдиҳии додани матн бо забони тоҷикӣ Tajik Text-to-Speech ва Computer Tajik Text Narrator **истифода шуданд** [9-М], [20-М], [23-М], [40-М], [57-М], [67-М].

11. Натиҷаҳои бадастомада дар конфронсҳои илмӣ-тадқиқотӣ дар сатҳи ҷумҳуриявӣ ва хориҷӣ **муаррифӣ шуда**, баҳои баланд гирифтанд. **Татбиқи** натиҷаҳои кор дар мақомоти давлатӣ ва муассисаҳои таҳсилоти олӣ имкон дод, ки масоили истифодаи самараноки забони тоҷикӣ дар ҷараёни коргузорӣ ҳаллу фасл гардад, инчунин метавонад ба рушди илми амсиласозии математикӣ, тарҳрезии низомҳои иттилоотӣ ва забоншиносии компютерӣ **мусоидат намояд** [41-М]-[50-М].

12. Натиҷаҳои кори диссертатсионӣ метавонад барои омӯзиши хусусиятҳои забони тоҷикӣ ҳам барои шаҳрвандони Ҷумҳурии Тоҷикистон ва ҳам барои ҳамаи онҳое, ки берун аз ҳудуди он қарор доранд, **заминаи устувору асоснок** шуда метавонад. Ҳамаи натиҷаҳои бадастомада ва лоиҳаҳои таҳияшуда дар Интернет дар www.tajlingvo.tj дастрас мебошанд [51-М].

ТАВСИЯҲО БАРОИ ИСТИФОДАИ АМАЛИИ НАТИҶАҲОИ ТАДҚИҚҲО

Натиҷаҳои дар диссертатсия бадастомада ҳалли масъалаҳои мубрам ва афзалиятноки омодакунии амсилаҳои математикӣ ва компютерӣ барои омӯзиши забон ва усулҳои худкори коркарди маълумоти матнӣ дар соҳаҳои санҷиши имло дар матн, тарҷумаи мошинии матн, синтез ва шинохти нутқ бо забони тоҷикӣ мебошанд. Муҷтамеи зикршудаи масъалаҳо барои баланд кардани сифати омӯзиши забони тоҷикӣ бо истифода аз имкониятҳои технологияҳои иттилоотӣ ва босуръатгардонии раванди коркарди ҳуҷҷатҳо дар Ҷумҳурии Тоҷикистон ва берун аз он аҳамияти калон дорад.

Инчунин натиҷаҳои таҳқиқоти диссертатсионӣ метавонанд дар ҷараёни таълим, дар пажӯҳишгоҳҳои илмӣ-тадқиқотӣ ва донишкадаҳои олии касбӣ ҳангоми таълими курсҳои махсуси фанҳои забоншиносии компютерӣ ва технологияҳои иттилоотӣ истифода шаванд. Илова бар ин, онҳо метавонанд ҳангоми таълифи корҳои курсӣ ва рисолаҳои хатми донишҷӯён, рисолаҳои илмии аспирантҳо ва ҷӯяндагони унвонҳои илмӣ дар соҳаи математика, технологияҳои иттилоотӣ ва забоншиносии компютерӣ васеъ истифода шаванд. Низомҳои худкори коркарди матн, ки бо забони тоҷикӣ таҳия шудаанд, барои истифода бо забони тоҷикӣ дар фаъолияти ҳуҷҷатгузорӣ дар ташкилоту корхонаҳои дохил ва хориҷи кишвар тавсия дода мешавад.

РӮЙХАТИ ИНТИШОРОТ АЗ РӮИ МАВЗӮИ ДИССЕРТАТСИЯ

Монографияҳо

[1-М] **Худойбердиев, Х.А.** Низомҳои худкори коркарди маълумот бо забони тоҷикӣ. [Матн] / З.Д. Усманов **Х.А. Худойбердиев** – Хуҷанд, ДДХБСТ, 2022. –186 с.

[2-М] **Худойбердиев, Х.А.** Комплекси барномаҳо барои талаффузи овози тоҷикӣ аз рӯйи матн. [Матн] / Усмонов З.Д., **Х.А. Худойбердиев** – Душанбе. Адиб, 2014. –158 с.

[3-М] **Худойбердиев, Х.А.** Опыт компьютерного синтеза таджикской речи по тексту. [Матн] / З.Д. Усманов, **Х.А. Худойбердиев** – Душанбе, Ирфон, 2010, –145 с.

Мақолаҳо дар маҷалла ва нашрияҳои илмӣ, ки аз тарафи КАО назди Президенти Ҷумҳурии Тоҷикистон ва КОА Федератсияи Россия тавсия шудааст

[4-М] **Худойбердиев, Х.А.** Оид ба низоми тарҷумони омории мошинӣ барои забони тоҷикӣ [Матн] / **Х.А.Худойбердиев** // Паёми донишгоҳи технологии Тоҷикистон. – 2023. № 3 (55). –С. 140-146.

[5-М] **Худойбердиев, Х.А.** Разработка и реализация системы машинного перевода на основе правил с русского на таджикский язык [Текст] / **Х.А.Худойбердиев** // Политехнический Вестник ТТУ имени академика М.С. Осими (Серия: Интеллект. Инновации. Инвестиции.). – 2023. – №2(62). – С. 33-36.

[6-М] **Худойбердиев, Х.А.** Моделирование системы автоматической обработки текста на таджикском языке [Текст] / **Х.А.Худойбердиев** // International Journal of Open Information Technologies. – 2023. – Т.11, № 3. – С. 27-33.

[7-М] **Худойбердиев, Х.А.** Цифровой портрет таджикского языка на основе статистических закономерностей кириллического алфавита [Текст] / **Х.А.Худойбердиев**, Ш.Н. Ашурова // Политехнический Вестник ТТУ имени академика М.С. Осими (Серия: Интеллект. Инновации. Инвестиции.). – 2022. – №4(60). – С. 29-32.

[8-М] **Худойбердиев, Х.А.** Вклад Усманова Зафара Джураевича в компьютерную лингвистику таджикского языка [Текст] / **Х.А.Худойбердиев** // Вестник Технологического университета Таджикистана. – 2022. № 4-1 (51). –С. 140-146.

[9-М] **Худойбердиев, Х.А.** Амсиласозии раванди шинохти нутқ дар заминаи нутқи забони тоҷикӣ [Матн] / Б.Х.Ашурзода, **Х.А.Худойбердиев** // Паёми Политехникӣ. ДДТ ба номи М.Осимӣ. (Бахши Интеллект, Инноватсия, Инвеститсия.) – 2022. – № 2 (58). – С. 39-42.

[10-М] **Худойбердиев, Х.А.** Масъалаҳои тарҳрезӣ ва коркарди луғатҳои электронӣ дар коркарди низомҳои худкори тарҷумон бо забони тоҷикӣ [Матн] / **Х.А.Худойбердиев** // Паёми Политехникӣ. ДДТ ба номи М.Осимӣ. (Бахши Интеллект, Инноватсия, Инвеститсия.) – 2022. – № 1 (57). – С. 41-47.

[11-М] **Худойбердиев, Х.А.** О проблемах художественного перевода и его взаимосвязь с машинным переводом на примере таджикского языка [Текст] / **Х.А.Худойбердиев** // Вестник технологического университета Таджикистана. – 2021. – № 4 (47). – С. 163-168.

[12-М] **Худойбердиев, Х.А.** Об алгоритме проверки орфографии на примере таджикского языка [Текст] / **Х.А.Худойбердиев** // Политехнический Вестник ТТУ имени академика М.С. Осими (Серия: Интеллект. Инновации. Инвестиции.). – 2021. – № 3 (31). – С. 48-53.

[13-М] **Худойбердиев, Х.А.** Система автоматической проверки орфографии таджикского языка – TajSpell [Текст] / О.М.Солиев, **Х.А.Худойбердиев**, Г.М.Довудов //. Вестник технологического университета Таджикистана. – 2021. – № 3 (46). – С. 188-193.

[14-М] **Худойбердиев, Х.А.** О распознавании автора текста на узбекском языке с помощью символьных униграмм [Текст] / **Х.А.Худойбердиев**, А.А.Косимов, П.Э.Зульфикарова // Проблемы вычислительной и прикладной математики. Научно-инновационный центр информационно-коммуникационных технологий Ташкентского университета информационных технологий имени М. аль-Хоразми. – 2020. – № 6 (30). – С. 49-55.

[15-М] **Худойбердиев, Х.А.** Оид ба монандкунии матн дар асоси басомади ҳиҷоҳо [Текст] / **Х.А.Худойбердиев**, А.А.Қосимов, Х.А.Тошхӯҷаев // Политехнический вестник. серия: интеллект. инновации. инвестиции. – 2020. – 2 (50). – С. 52-56.

[16-М] **Худойбердиев, Х.А.** О распознавании автора текста на основе частотности слогов [Текст] / **Х.А.Худойбердиев**, А.А.Косимов // Доклады Академии наук Республики Таджикистан. – 2019. – Т.62, № 11-12. – С. 641-645.

[17-М] **Худойбердиев, Х.А.** О статистических закономерностях слогового состава таджикского языка [Текст] / **Х.А. Худойбердиев //** Вестник Таджикского технического Университета, – 2015. – № 3 (31). – С. 48-53.

[18-М] **Худойбердиев, Х.А**. О соотношении словоформ и словоупотреблений в русском переводе произведения А.Фирдоуси «Шахнаме» [Текст] / **Х.А.Худойбердиев**, А.А.Косимов // Доклады Академии наук Республики Таджикистан. – 2015. – Т.58, № 9. – С. 786-792.

[19-М] **Худойбердиев, Х.А**. Об автоматическом конвертировании таджикского текста к стандартной графике [Текст] / **Х.А.Худойбердиев** // Доклады Академии наук Республики Таджикистан, – 2014. – Т.57, № 3. – С. 210-214.

[20-М] **Худойбердиев, Х.А.** О синтезе таджикской речи с русизмами [Текст] / З.Д.Усманов, **Х.А.Худойбердиев //** Доклады Академии наук Республики Таджикистан. – 2009. – Т.52, – № 5. – С. 358-361.

[21-М] **Худойбердиев, Х.А.** О синтезаторе таджикской речи по тексту [Текст] / З.Д.Усманов, **Х.А.Худойбердиев //** Доклады Академии наук Республики Таджикистан. – 2009. –Т.52, № 4. – С. 267-271.

[22-М] **Худойбердиев, Х.А.** Об автоматическом разложении слов на слоги. [Текст] / **Х.А.Худойбердиев //** Доклады Академии наук Республики Таджикистан. – 2007. – Т.50, № 5. – С. 417-419.

[23-М] **Худойбердиев, Х.А.** Алгоритм безударного озвучивания таджикского текста. [Текст] / З.Д.Усманов, **Х.А.Худойбердиев //** Доклады Академии наук Республики Таджикистан. – 2007. – Т.50, № 4. – С. 302-305.

[24-М] **Худойбердиев, Х.А.** О многообразии слогов таджикского языка. [Текст] / **Х.А.Худойбердиев //** Известия Академии наук Республики Таджикистан. – 2007. – №2 (127). – С. 31-34.

[25-М] **Худойбердиев, Х.А.** О слоговой структуре слов таджикского языка [Текст] / З.Д.Усманов, **Х.А.Худойбердиев //** Доклады Академии наук Республики Таджикистан. – 2006. – Т. 49, № 6. – С. 489-492.

Мақолаҳо дар дигар маҷаллаҳои илмӣ

[26-М] **Худойбердиев, Х.А.** Рушди илми лингвистикаи компютерӣ дар Ҷумҳурии Тоҷикистон [Матн] / О.М. Солиев, **Х.А. Худойбердиев**, Г.М. Довудов, Ш.Н. Ашӯрова // Паёми ДПДТТ ба номи академик М.С.Осимӣ. – 2022. – № 2 (23). – С. 17-24.

[27-М] **Худойбердиев, Х.А.** Проектирование и программная реализация автоматической транслитерации в цифровой библиотеке [Текст] / **Х.А. Худойбердиев**, М.П. Музаффаров, Ф.Э. Мирзозода // Вестник ПИТТУ имени академика М.С.Осими. – 2022. – № 1 (22). – С. 7-15.

[28-М] **Худойбердиев, Х.А.** Перспективы развития информационного пространства и цифровизации в Таджикистане: обзор основных тенденций [Текст] / Х.Т. Максудов, **Х.А. Худойбердиев**, Ш.Х. Максудов // Вестник ПИТТУ имени академика М.С. Осими. – 2021. – № 4 (21). – С. 7-18.

[29-М] **Khurshed A. Khudoyberdiev.** The Algorithms of Tajik Speech Synthesis by Syllable. Polytechnic institute of Tajik technical university named after academician M.S. Osimi, - Polytechnic institute of Tajik technical university named

after academician M.S. Osimi, Khujand. Tajikistan. International Forum “IT-Technologies for Engineering Education: New Trends and Implementing Experience” (ITEE-2019). Anthropological Dimension of Digital Technologies in Engineering Education ITM Web of Conferences 35, 07003 (2020).

[30-М] **Худойбердиев, Х.А.** Сравнительный анализ систем распознавания звука Sphinx и Mozilla Deepspeech [Текст] / **Х.А. Худойбердиев**, Р.М. Воситов // Вестник ПИТТУ имени академика М.С.Осими. – 2021. – № 1 (18). – С. 7-13.

[31-М] **Худойбердиев, Х.А.** Муаммоҳои тарҷумаи бадеӣ ва вобастагии он бо тарҷумаи мошинӣ дар Тоҷикистон [Матн] / З.А. Раҳмонов, **Х.А. Худойбердиев //** Паёми ДПДТТ ба номи академик М.С. Осимӣ. – 2020. – № 2 (7). – С. 7-11

[32-М] **Худойбердиев, Х.А.** Разработка параллельного корпуса таджикского и русского языков [Текст] / **Худойбердиев, Х.А.**, О.М. Солиев, П.А. Солиев // Новые информационные технологии в автоматизированных системах. – 2019. – № 22. – С. 179-181.

[33-М] **Худойбердиев, Х.А.** Информационная система и каталогизации кодексов республики Таджикистан [Текст] / **Х.А. Худойбердиев**, И.А. Джалолов // Вестник ПИТТУ имени академика М.С.Осими. – 2019. – № 3 (12). – С. 9-18.

[34-М] **Худойбердиев, Х.А**. Захираи мувозии забони точикӣ-русӣ: коркард ва тавсифи он [Матн] / **Х.А. Худойбердиев**, А.А. Назаров // Паёми ДПДТТ ба номи академик М.С.Осимӣ. – 2019. – № 1(10). – С. 7-12.

[35-М] **Худойбердиев, Х.А.** Сегментация речевого сигнала на базе слоговых структур таджикского языка [Текст] / **Х.А. Худойбердиев** // Новые информационные технологии в автоматизированных системах. – 2018. – № 21. – С. 181-182.

[36-М] **Худойбердиев, Х.А.** Сохтори мантиқӣ ва таҳлили артефактҳои тарҷумаи мошинӣ [Матн] / **Х.А. Худойбердиев**, З.А. Раҳмонов // Паёми ДПДТТ ба номи академик М.С. Осимӣ. – 2018. – № 2 (7). – С. 7-11.

[37-М] **Худойбердиев, Х.А.** Лингвистический тезаурус таджикского языка [Текст] / **Х.А. Худойбердиев**, О.М. Солиев // Новые информационные технологии в автоматизированных системах. – 2017. – № 20. – С. 103-105.

[38-М] **Худойбердиев, Х.А.** Модель анализа и сегментации речевого сигнала для послогового распознавания таджикской речи [Текст] / **Х.А. Худойбердиев** // Вестник Технологического университета Таджикистана. – 2017. – № 4 (31). – С. 85-87.

[39-М] **Худойбердиев, Х.А.** О множестве анаграмм в произведениях К.Худжанди [Текст] / **Х.А. Худойбердиев**, А.А. Косимов // Вестник ПИТТУ имени академика М.С. Осими. – 2017. – №2 (3). – С. 14-22.

[40-М] **Худойбердиев, Х.А.** О синтезаторе таджикской речи по тексту [Текст] / **Х.А. Худойбердиев** // Новые информационные технологии в автоматизированных системах. – 2013. – № 16 – С. 273-276.

Баромад ва тезисҳо дар конференсияҳо

[41-М] **Худойбердиев, Х.А.** О некоторых способах математического моделирования синтеза и распознавания речи [Текст] / **Х.А. Худойбердиев** // Материалы международной конференции «Современные проблемы

математики», посвящённой 50-летию Института математики им. А. Джураева Национальной академии наук Таджикистана. – Душанбе, Института математики им. А. Джураева НАНТ, 2023. – С. 253-255.

[42-М] **Худойбердиев, Х.А.** Формирование электронного словаря для системы автоматического перевода текста с таджикского языка на русский [Текст] / **Х.А. Худойбердиев**, А.А. Назаров, Ш.Н. Ашурова // Всероссийская научно-практическая конференция с международным участием «Информационный обмен в междисциплинарных исследованиях II». – Рязань, 2023. – С. 227-231.

[43-М] **Худойбердиев, Х.А.** Низомҳои худкор барои коркарди матн бо забони тоҷикӣ [Матн] / **Х.А. Худойбердиев //** Международная научно-практическая конференция «Новые достижения в области естественных наук и информационных технологий». – Душанбе, РТСУ, 2023. – С. 194-196.

[44-М] **Худойбердиев, Х.А.** Тарҳрезии низомҳои худкор барои коркарди матн бо забони тоҷикӣ [Матн] / **Х.А. Худойбердиев //** Конференсияи илмӣ-амалии ҷумҳуриявӣ бахшида ба рӯзи байналмилалии забони модарӣ таҳти унвони “Забони модарӣ – сарчашмаи худшиносӣ ва маънавиёти миллӣ”. – Душанбе, Кумитаи забон ва истилоҳоти назди Ҳукумати ҶТ, 2023.

[45-М] **Худойбердиев, Х.А.** Баланд бардоштани сифати корҳои хаттӣ бо истифодаи барномаи зидди асардуздӣ (Antiplagiat_TJ) [Матн] / **Х.А. Худойбердиев**, А.А. Косимов, М.Х. Файзуллозода, Х.М. Муродов, Ё.О. Зулфов // Конференсияи ҷумҳуриявӣ илмию амалӣ дар мавзӯйи «Тадбиқи технологияҳои иттилоотӣ ва коммуниткатсионӣ дар саноаткунонии кишвар», бахшида ба ҳадафи чоруми стратегии миллӣ. – Душанбе, Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ, 2022.

[46-М] **Худойбердиев, Х.А.** Современные тенденции в компьютерной лингвистике таджикского языка [Текст] / **Х.А. Худойбердиев** // Республиканская научно-практическая конференция «Актуальные проблемы лингвистики и лингводидактики в современных условиях». – Душанбе, Филиал Московского государственного университета имени М.В. Ломоносова в городе Душанбе, 2022. – С. 279-284.

[47-М] **Худойбердиев, Х.А.** О проблеме автоматической транслитерации текста на таджикском языке [Текст] / **Х.А. Худойбердиев** // IV Международная научно-практическая конференция «Наука и технологии». – Алматы, Казахстан, 2022. – С. 101-106.

[48-М] **Худойбердиев, Х.А.** Таҳлили масъалаҳои асосии пешбарии тарҷумаи мошинӣ дар мисоли забони тоҷикӣ [Матн]. / **Х.А. Худойбердиев** // Конференсия ҷумҳуриявии илмӣ-амалӣ Масъалаҳои мубрами тарҷума ва забоншиносӣ дар замони муосир”. – Душанбе, Донишкадаи давлатии забонҳои тоҷикистон ба номи Сотим Улуғзода, 2019.

[49-М] **Худойбердиев, Х.А.** Методҳо ва алгоритмҳо барои шинохти овоз [Матн] / Н.С. Маҳмудов, **Х.А. Худойбердиев**, Ғ.Ҳ. Сафаров // Конференсияи илмӣ-амалии омӯзгорон, муҳаққиқони ҷавон бахшида ба 30-солагии Истиқлолияти давлатии Ҷумҳурии Тоҷикистон. – Хуҷанд, ДПДТТХ ба номи академик М.С.Осими, 2019.

[50-М] **Худойбердиев, Х.А.** Алгоритмы послогового распознавания

таджикской речи в амплитудно-временном пространстве [Текст] / **Х.А. Худойбердиев** // Научно-практическая конференция «Применение информационно-коммуникационных технологий для инновационного развития Республики Таджикистан». – Душанбе, ТУТ, 2017.

Шаҳодатномаҳои муаллифӣ ва бақайдгирии давлатии захираҳои иттилоотӣ

[51-М] **Худойбердиев, Х.А.** Web-приложение “Автоматические системы обработки информации на таджикском языке – www.tajlingvo.tj" [SOFT] / Х.А. Худойбердиев **//** – 28.04.2022. – № 4202200496.

[52-М] **Худойбердиев, Х.А.** Web-приложение таджикский переводчик (tarjumon.tj) [SOFT] / **Х.АХудойбердиев**, О.М.Солиев, П.А.Солиев, Г.М.Довудов, А.А.Назаров // – 03.12.2021/ –№ 4202100482.

[53-М] **Худойбердиев, Х.А.** Web-сайт “Электронный каталог кодексов Республики Таджикистан” [SOFT] / **Х.А. Худойбердиев**, И.А. Джалолов // – 25.02.2021. –№ 4202100470.

[54-М] **Худойбердиев, Х.А.** Автоматическая система TajSpell-2.0. для проверки орфографии таджикского языка в офисном пакете приложений MS Office 2010-2019 [SOFT] / З.Д. Усманов, О.М. Солиев, **Х.А. Худойбердиев**, Г.М. Довудов // – 30.07.2020. – № 4202000456.

[55-М] **Худойбердиев, Х.А.** Web-приложение Tajik-Russian-Parallel Corpus [SOFT] / **Х.А. Худойбердиев**, О.М. Солиев, Г.М. Довудов, А.А. Косимов // – 30.04.2019. – № 4201900402.

[56-М] **Худойбердиев, Х.А.** Web-приложение Tajik-English-Parallel Corpus [SOFT] / **Х.А. Худойбердиев**, О.М. Солиев, А.А. Назаров, П.А. Солиев // – 30.04.2019. – № 4201900401.

[57-М] **Худойбердиев, Х.А.** Компьютерный Диктор таджикского текста Computer Tajik Text Narrator [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, А.А. Худойбердиев // –10.06.2018. – № 4201800386.

[58-М] **Худойбердиев, Х.А.** База данных «Единица измерений текстов произведений таджикских современных поэтов и писателей» [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, А.А. Косимов // – 16.05.2018. –№ 4201800381.

[59-М] **Худойбердиев, Х.А.** База данных «Единица измерений текстов произведений таджикских классических поэтов и писателей» [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, А.А. Косимов // – 16.05.2018. – № 4201800380.

[60-М] **Худойбердиев, Х.А.** Web-приложение проверки уникальности текста на таджикском языке Taj_Text_Plagiat [SOFT] / З.Д. Усманов, О.М. Солиев, **Х.А. Худойбердиев**, П.А. Солиев // – 16.05.2018. – № 4201800378.

[61-М] **Худойбердиев, Х.А.** База данных αβ-кодирования для распознавания анаграмм [SOFT] / З.Д. Усманов, О.М. Солиев, **Х.А. Худойбердиев**, Г.М. Довудов, А.А. Косимов // – 16.05.2018. – № 4201800377.

[62-М] **Худойбердиев, Х.А.** Таджикский языковой пакет для тезауруса в Microsoft Office [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, О.М. Солиев, Г.М. Довудов // – 04.10.2012. – № 4201200237.

[63-М] **Худойбердиев, Х.А.** Таджикский языковой пакет для расстановки

переносов в Microsoft Office [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, О.М. Солиев, Г.М. Довудов // – 04.10.2012. – № 4201200236.

[64-М] **Худойбердиев, Х.А.** Таджикский языковой пакет для проверки орфографии в Microsoft Office [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, О.М. Солиев, Г.М. Довудов // – 04.10.2012. – № 4201200235.

[65-М] **Худойбердиев, Х.А.** Компьютерный мультиязыковый словарь MultiGanj. [SOFT] / З.Д. Усманов, С. Холматова, **Х.А. Худойбердиев**, О.М. Солиев // – 12.11.2008. – № 077TJ.

[66-М] **Худойбердиев, Х.А.** Компьютерный русско-таджикский словарь [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, О.М. Солиев // – 29.01.2008. – № 054TJ.

[67-М] **Худойбердиев, Х.А.** Компьютерное озвучивание таджикского текста Tajik Text-to-Speech [SOFT] / **Х.А. Худойбердиев** // – 04.09.2007. – № 041TJ.

[68-М] **Худойбердиев, Х.А.** Таджикский текстовый редактор Tajik Word (TW) [SOFT] / З.Д. Усманов, **Х.А. Худойбердиев**, О.М. Солиев // – 05.07.2007. – № 030TJ.

АННОТАЦИЯ

на автореферат диссертации **Худойбердиева Хуршеда Атохоновича** на тему «ПРОЕКТИРОВАНИЕ И РЕАЛИЗАЦИЯ АВТОМАТИЧЕСКИХ СИСТЕМ ОБРАБОТКИ ИНФОРМАЦИИ НА ТАДЖИКСКОМ ЯЗЫКЕ» на соискание ученой степени доктора технических наук по специальности 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

Ключевые слова: таджикский язык, математическое моделирование, математическая статистика, численные методы, компьютерная лингвистика, компьютерное моделирование, автоматическая проверка орфографии, машинный перевод, синтез речи, информационные системы, база данных, технология программирования.

Цель диссертационной работы состоит в разработке моделей, методов и алгоритмов, позволяющих создавать информационные системы автоматической обработки информации на таджикском языке для их дальнейшего использования в человеко-машинных системах управления в естественно-языковом диалоге.

Методы исследования: для решения задач, стоящих перед исследованием, были использованы методы систематического анализа, математической статистики, основы представления и обработки наборов данных, а также теория алгоритмов, математическая и компьютерная лингвистика, синтез данных, компьютерное моделирование автоматических информационных систем, технологии программирования и обработки данных.

Научная новизна: в результате научно-исследовательской работы и разработки автоматических систем предложен ряд методических подходов к исследованию, анализу и автоматической обработке текстовой информации на таджикском языке.

Положения, выносимые на защиту:

1. Представлена концепция автоматической обработки текстовой информации на таджикском языке как объект научного исследования и программные средства для систематического анализа, на основе которых определены понятия и теоретические термины.
2. Представлен и экспериментально проверен научно-практический подход к разработке электронных словарей и компьютерных тезаурусов, в рамках которого сформированы примеры решения поисковых задач и методы применения этих примеров в процессе реализации компьютерных словарей.
3. Впервые предложен подход автоматического синтеза речи на таджикском языке, основанный на использовании метода конкатенации слогов. При этом получены всевозможные структуры слогов и слоговых структур слов таджикского языка. Разработано программное обеспечение автоматического синтеза речи на основе собственных алгоритмов и базы данных «слог-звук». В результате синтеза речи формируется целая звуковая дорожка на основе предложенной текстовой информации в формате цифрового звукового файла.
4. Разработаны новые методы извлечения, представления и обработки данных, составляющих отдельные элементы текста, а также предложен новый способ решения проблемы автоматического правописания текста на таджикском языке.
5. Впервые изучен вопрос автоматического перевода текста с таджикского языка на русский, разработаны модели, методы и алгоритмы, которые дают возможность эффективно решать практические задачи.
6. Исследовано совместное использование системного анализа, структурированного подхода к обработке данных, объектно-ориентированного программирования и компьютерной лингвистики для разработки систем автоматической обработки текстовой информации на таджикском языке.
7. Для реализации всех представленных автоматических систем разработан комплекс компьютерных программ TajLINGVO, проведена его экспериментальная апробация на территории Республики Таджикистан.

Программный комплекс TajLINGVO зарегистрирован в Национальном патентном центре Министерства экономического развития и торговли Республики Таджикистан, а разработанные проекты доступны на сайте www.tajlingvo.tj.

АННОТАТСИЯ

ба автореферати рисолаи илмии **Худойбердиев Хуршед Атохонович** дар мавзӯи «БАЛОИҲАГИРӢ ВА АМАЛИГАРДОНИИ НИЗОМҲОИ ХУДКОРИ КОРКАРДИ МАЪЛУМОТ БО ЗАБОНИ ТОҶИКӢ» барои дарёфти дараҷаи илмии доктори илмҳои техникӣ аз рӯйи ихтисоси 05.13.11 - Таъминоти математикӣ ва барномавии мошинҳои ҳисоббарор, муҷтамаъҳо ва шабакаҳои компютерӣ

Калимаҳои калидӣ: забони тоҷикӣ, амсиласозии математикӣ, омори математикӣ, усулҳои ададӣ, забоншиносии компютерӣ, тарҳрезии компютерӣ, тафтиши худкори имло, тарҷумони мошинӣ, синтез нутқ, низомҳои иттилоотӣ, манбаи додаҳо, технологияи барномарезӣ.

Ҳадафи тадқиқот - таҳияи амсилаҳо, усулҳо ва алгоритмҳое мебошад, ки барои эҷоду таҳияи низомҳои иттилоотии коркарди худкори иттилоот бо забони тоҷикӣ барои истифодаи минбаъдаи онҳо дар низомҳои идоракунии инсону мошин дар муколамаи табиии забонӣ имкон медиҳанд.

Усулҳои тадқиқот. Барои ҳалли вазифаҳои дар назди таҳқиқот гузошташуда усулҳои таҳлили низомманд, омори математикӣ, асосҳои пешниҳод ва коркарди муҷтамеи додаҳо, инчунин назарияи алгоритмҳо, лингвистикаи математикӣ ва компютерӣ, синтези маълумот, амсиласозии компютерии низомҳои худкори иттилоотӣ, технологияҳои барномасозӣ ва коркарди маълумот истифода шуданд.

Навгонии илмии тадқиқот. Дар натиҷаи кори илмӣ-тадқиқотӣ ва таҳияи низомҳои худкор як қатор равишҳои методии таҳқиқ, таҳлил ва коркарди худкори иттилооти матнӣ бо забони тоҷикӣ пешниҳод шудааст.

Муқаррароте, ки барои дифоъ пешниҳод карда мешавад:

1. Мафҳуми коркарди худкори иттилооти матнӣ бо забони тоҷикӣ ҳамчун объекти тадқиқоти илмӣ ва воситаҳои нармафзор барои таҳлили низомманд пешниҳод гардида, дар асоси онҳо мафҳумҳо ва истилоҳоти назариявӣ муайян карда шуданд.

2. Равиши илмӣ-амалии таҳияи луғатҳои электронӣ ва тезаурусҳои компютерӣ пешниҳод шудаанд, ки дар доираи он намунаҳои ҳаллу фасли масъалаҳои ҷустуҷӯ ва усулҳои истифодаи намунаҳои зикршуда дар раванди татбиқи луғатҳои компютерӣ ташаккул дода шуданд.

3. Бори аввал равиши синтези худкори нутқ бо забони тоҷикӣ пешниҳод шудааст, ки дар заминаи усули пайвандкунии ҳиҷоҳо асос ёфтааст. Ҳамзамон дар забони тоҷикӣ тамоми сохторҳои имконпазири ҳиҷо ва таркиби ҳиҷоии вожаҳо ба даст оварда шуданд. Нармафзори синтези худкори нутқ дар асоси алгоритмҳои худӣ ва пойгоҳи додаҳои “ҳиҷо-садо” таҳия шудааст. Дар натиҷаи синтези нутқ дар асоси иттилооти матнии пешниҳодшуда дар шакли файли овозии рақамӣ, мавҷи садоии пурраи овозӣ ташкил карда мешавад.

4. Усулҳои нави истихроҷ, пешниҳод ва коркарди додаҳое, ки унсурҳои алоҳидаи матнро ташкил медиҳанд, таҳия шуда, тарзи нави ҳалли масъалаи имлои худкори матн дар забони тоҷикӣ пешниҳод шудааст.

5. Бори аввал масъалаи аз забони тоҷикӣ ба забони русӣ ба таври худкор тарҷума кардани матн мавриди таҳқиқ қарор гашта, амсилаҳо, усулҳо ва алгоритмҳое таҳия карда шуданд, ки барои ҳалли самараноки масъалаҳои амалӣ имкон медиҳанд.

6. Истифодаи муштараки таҳлили низомманд, равиши сохторӣ нисбати коркарди додаҳо, барномасозӣ ба объект нигаронидашуда ва забоншиносии компютерӣ барои таҳияи низомҳои коркарди худкори иттилооти матнӣ ба забони тоҷикӣ таҳқиқ карда шуд.

7. Барои татбиқи ҳамаи низомҳои худкори пешниҳодшуда маҷмӯи барномаҳои компютерии TajLINGVO тартиб дода шуда, дар ҳудуди Ҷумҳурии Тоҷикистон озмоиши таҷрибавии он гузаронида шуд.

Муҷтамеи нармафзори TajLINGVO дар Маркази миллии патентии Вазорати рушди иқтисод ва савдои Ҷумҳурии Тоҷикистон ба қайд гирифта шудааст ва дар сомонаи www.tajlingvo.tj дастрас аст.

ANNOTATION

on the abstract of the dissertation thesis **Khudoyberdiev Khurshed Atokhonovich** on the topic "DESIGN AND IMPLEMENTATION OF AUTOMATIC SYSTEMS OF INFORMATION PROCESSING IN TAJIK LANGUAGE" for fulfillment of the requirements for the degree of Doctor of Technical Sciences on specialty 05.13.11 - Mathematical and software of computing machines, complexes and computer networks.

Keywords: Tajik language, mathematical modelling, mathematical statistics, numerical methods, computer linguistics, computer modelling, automatic spelling check, machine translation, speech synthesis, information systems, database, programming technology.

The objective of the dissertation work is to develop models, methods and algorithms that allow to create information systems of automatic information processing in the Tajik language for their further utilization in human-machine control systems in natural-language dialogue.

Methods of research: for solving the problems facing research, it was used methods of systematic analysis, mathematical statistics, fundamentals of representation and processing of data sets, as well as the theory of algorithms, mathematical and computer linguistics, data synthesis, computer modelling of automatic information systems, programming and data processing technologies.

Scientific novelty: as a result of research work and development of automatic systems, it has been proposed a number of methodological approaches to the research, analysis and automatic processing of text information in the Tajik language.

Statements put forward for defenses:

1. The concept of automatic processing of text information in Tajik language as an object of scientific research and software tools for systematic analysis is presented, on the basis of which the concepts and theoretical terms are represented.

2. The scientific and practical approach to the development of electronic dictionaries and computer thesauruses is presented and experimentally tested, within the framework of which examples of solving search problems and methods of applying these examples in the process of implementing computer dictionaries are formed.

3. For the first time the approach of automatic speech synthesis in Tajik language based on the use of syllable concatenation method has been proposed. At the same time all possible syllable structures and syllable structures of words of Tajik language are obtained. Software for automatic speech synthesis based on own algorithms and "syllable-sound" database has been developed. As a result of speech synthesis, a whole sound track is formed on the basis of the proposed textual information in the format of a digital sound file.

4. New methods of extraction, representation and data processing constituting separate text elements have been developed, as well as a new way of problem solving of automatic text spelling in Tajik language has been proposed.

5. For the first time the question of automatic text translation from Tajik into Russian has been studied; models, methods and algorithms, which make it possible to solve effectively practical problems have been developed.

6. The joint use of system analysis, structured approach to data processing, object-oriented programming and computer linguistics for the development of systems of automatic processing of text information in Tajik language has been investigated.

7. For realization of all presented automatic systems the complex of computer programs TajLINGVO is developed, its experimental approbation on the territory of the Republic of Tajikistan is carried out.

TajLINGVO software complex is registered in the National Patent Centre of the Ministry of Economic Development and Trade of the Republic of Tajikistan, and the developed projects are available at www.tajlingvo.tj.