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

ДИССЕРТАЦИЯ

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

Научный консультант: **доктор физико-математических наук, Академик НАНТ, профессор**

Усманов Зафар Джураевич

ДУШАНБЕ - 2024

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Актуальность темы исследования. Системы автоматической обработки текстовой информации на естественном языке функционируют посредством набора программных пакетов и компьютерных приложений, в основе которых лежит математическая модель. 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.

Правильное и эффективное использование информационных технологий требует от пользователей выполнения ряда обязательств с учетом защиты национального языка [3] и соблюдения Закона «О государственном языке Республики Таджикистан» [5]. Кроме того, Закон Республики Таджикистан «О распространении информации» [7] и Закон Республики Таджикистан «Об информации» [6] являются одним из основных механизмов внедрения информационных технологий во все сферы жизни общества. В связи с этим, планирование и построение ряда автоматизированных информационных систем обработки данных на таджикском языке представляется одной из наиболее актуальных задач современности.

В рамках научных исследований в области компьютерной лингвистики и выполнения Государственной стратегии «Информационно-коммуникационные технологии для развития Республики Таджикистан» [8], Стратегии инновационного развития Республики Таджикистан на период до 2020 года [10], Национальной стратегии развития Республики Таджикистан на период до 2030 года [9] достигнуты значительные результаты, а именно для формирования и развития электронных словарей, синтеза речи, автоматической проверки орфографии и машинного перевода текста были спроектированы и внедрены системы автоматической обработки информации на таджикском языке.

Результаты исследования направлены на реализацию Постановления Правительства Республики Таджикистан о «Государственной программе внедрения информационно-коммуникационных технологий в общеобразовательных учреждениях Республики Таджикистан» [1], Постановления Правительства Республики Таджикистан о «Концепции формирования электронного правительства в Республике Таджикистан» [2] и Постановления Правительства Республики Таджикистан «Государственный стандарт размещения таджикского алфавита на компьютерной клавиатуре» [4], что в свою очередь подчеркивает особую актуальность выбранной темы исследования.

Степень научной разработанности изучаемой проблемы. 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 необходимыми 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. и Damerau F.J. [46] Grishman R. [43], Hutchins W.J. [45], Hausser R.R. [44], Cohen M. и Massaro D. [42], Liberman A.M. [49], Black A.W. и Taylor P.A. [41], Johnson M. [47], Nirenburg S., Somers H.L., Wilks Y. [51], Koehn P. [48], Mercer R.L. [50], Schroeder M. [52], Zen H. [53] 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.

Разработка методологии, методов и базовых моделей разработки систем автоматической обработки текстовой информации имеет давнюю историю. В трудах таких ученых, как Д.Ш.Сулейманов [36], В.А.Фомичев [38], А.В.Анисимов [17], Т.В.Батура ва Ф.А.Мурзин [18], О.Ф.Кривнова [23], С.В.Лесников [26], А.А.Марченко [27], Р.К.Потапова [29], С.Б.Потемкин ва Г.Е.Кедрова [30], Н.Н.Сажок [33], А.И.Солонина [34], В.Н.Сорокин [35], Л.А.Чистович [39] 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.

В исследованиях Д.Ш. Сулейманова [159] и В.А. Фомичева [267] представлены методы разработки основных принципов, структуры и технологии, направленные на создание практических лингвистических моделей, которые затем применяются в информационных системах обработки текстов на естественных языках, синтаксического и семантического анализа.

Автоматические системы обнаружения и исправления орфографических ошибок в русских текстах, а также автоматического индексирования русских текстов по ключевым словам были предложены группой учёных под руководством И.В. Ковалёва [160] для исследования обширного текстового материала, алгоритмов и программ на основе морфологического, синтаксического и семантического анализа. Эти методы в компьютерной лингвистике создали совершенно новые возможности для создания систем с перспективой автоматической обработки текстовой информации.

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.

Вместе с тем обработка текстовых материалов на таджикском языке в вышеперечисленных системах не отвечают требованиям орфографии таджикского языка. В связи с этим, следует признать, что практические и информационные возможности упомянутых систем обеспечивают пользователю действенный процесс исследования основ естественного языка. Поэтому возникает вопрос моделирования и разработки систем автоматической обработки текстовых данных на таджикском языке.

Широкое использование информационно-коммуникативных технологий в Таджикистане привлекло внимание исследователей в области информатики 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, 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.

Системное решение одной из важнейших задач государственной программы – создание автоматической системы обработки информации на таджикском языке, с которой тесно связаны как независимость, так и информационная безопасность Таджикистана, вполне возможно, т.к. страна обладает достаточным количеством признанных специалистов во многих областях науки и приемлемой научно-технической базой.

Различные области компьютерной лингвистики, считающейся важной стороной связи информации с общественной жизнью в нашей стране, уже исследованы как отдельными специалистами, так и небольшими коллективами.

Проводимое исследование в диссертационной работе, наряду с достижениями таджикских ученых, обеспечивает высокий уровень научности рассматриваемых проблем. В частности, таджикские исследователи в сфере своей научной деятельности предложили большое количество систем автоматической обработки элементов текстовой информации на таджикском языке, в том числе: компьютерную раскладку таджикского алфавита; N-граммы букв; слоговую структуру; слоговую структуру слов; формы слов и их произношение; анаграммы; N-граммы слов; морфы; предлоги и суффиксы; корни слов; фразы и типы предложений. Все научные достижения в виде интеллектуального продукта или информационных ресурсов получили свидетельство о государственной регистрации [317-322]. Для реализации необходимых задач были созданы пакеты компьютерных программ и веб-приложения: система автоматической проверки орфографии, пакет синтеза и распознавания речи, система голосового управления конечным устройством, а также система автоматического компьютерного перевода.

Следует отметить, что предлагаемая диссертационная работа выполнена на основе трудов З.Д. Усманова. В ней использованы основные понятия моделирования систем автоматической обработки информации на таджикском языке. Компьютерные модели и алгоритмы, представленные в диссертации,

позволяют использовать информационные технологии для обработки информации на таджикском языке и помогают изучать таджикский язык современными методами.

Связь исследования с программами (проектами), научной тематикой. 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.

ОБЩЕЕ ОПИСАНИЕ ИССЛЕДОВАНИЯ

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.

Методы исследования. Для решения задач, стоящих перед исследованием, были использованы методы систематического анализа, математической статистики, основы представления и обработки наборов данных. В работе также применены 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 поход 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.

Уровень достоверности результатов. Надежность спроектированных автоматических систем и программного комплекса подтверждена

соответствующими актами о практическом внедрении, 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. и 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.

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 методы, модели и алгоритмы обработки информации, описанные на таджикском языке и представленные на

защиту, подготовлены им при участии и руководстве его научного консультанта.

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 2023 г.), г. Душанбе; всероссийская научно-практическая конференция с участием международных представителей по теме «Обмен информацией в междисциплинарных исследованиях II», (14 апреля 2023 г.), 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 2023 г.), 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.), 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; третья республиканская конференция «Практические информационные системы: проблемы моделирования, внедрения в развивающихся странах» – АИС-2022, ХПИТТУ имени академика М.С. Осими, (26 октября 2022 г.), 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 Улугзода, (2019 г.), г. Душанбе;

материалы 22-го научно-практического семинара «Новые информационные технологии в автоматических системах», Институт прикладной математики им. М.В. Келдыша РАН, (19 апреля 2019 г.), г. Москва, Российская Федерация; научно-практическая конференция преподавателей, молодых исследователей, посвященная 30-летию Государственной Независимости Республики Таджикистан, ХПИТТУ имени академика М.С. Осими, (2019 г.), г. Худжанд; региональная научно-практическая конференция, посвященная 90-летию Темурхана Максудова, Филиал Технологического университета Таджикистана в городе Исфаре, (2018 г.), г. Исфара; материалы 21-го научно-практического семинара «Новые информационные технологии в автоматических системах», Институт прикладной математики им. М.В. Келдыша РАН, (20 апреля 2018 г.), г. Москва, Российская Федерация; научно-практическая конференция «Применение информационно-коммуникационных технологий для инновационного развития Республики Таджикистан», Технологический университет Таджикистана, (2017 г.), г. Душанбе; материалы 20-го научно-практического семинара «Новые информационные технологии в автоматических системах», Институт прикладной математики им. М.В. Келдыша РАН, (21 апреля 2017 г.), г. Москва, Российская Федерация; вторая международная конференция «Практические информационные системы: проблемы моделирования, внедрения в развивающихся странах» – АИС-2017, ХПИТТУ имени академика М.С. Осими, (14-15 апреля 2017 г.), г. Худжанд; 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. Худжанд; материалы 17-го научно-практического семинара «Новые информационные технологии в автоматических системах», Институт прикладной математики им. М.В. Келдыша РАН, (18 апреля 2014 г.), г. Москва, Российская Федерация; 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; материалы 16-го научно-практического семинара «Новые информационные технологии в автоматических системах», Институт прикладной математики им. М.В. Келдыша РАН, (19 апреля 2013 г.), г. Москва, Российская Федерация; первая международная конференция «Практические информационные системы: проблемы моделирования, использование в развивающихся странах» – АИС-2012, ХПИТТУ им. академика М.С. Осими, (30 апреля 2012 г.), г. Худжанд.

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 интеллектуального продукта.

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.

ГЛАВА 1. ПЕРСПЕКТИВНЫЕ ЗАДАЧИ РЕАЛИЗАЦИИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ В ТАДЖИКСКОМ ЯЗЫКЕ

§ 1.1. Общая информация

Широкое распространение компьютерной техники в Таджикистане позволило организациям и учреждениям страны осуществить переход на совершенно новую технологию подготовки печатных текстов на таджикском языке. В связи с этим возникли важные задачи, связанные с созданием таджикских графических драйверов и решением вопроса стандартизации печатной продукции.

Другой проблемой являлось формирование оригинального таджикского контента в Интернете. Современные информационные технологии обработки информации прочно вошли в практику большинства стран мира и стали неотъемлемой частью современной цивилизации. Поэтому создание информационных систем обработки информации на таджикском языке для использования широкого круга таджикско-адаптированной компьютерной графики доказало свою очевидную необходимость.

Модели и методы создания электронных словарей для формирования компьютерного тезауруса, представленные Devadason F.J. [214], Collins M. [204], Daille B. [210], De Luca E.W. [212], Hausser R.R. [221], Kutuzov A. [233], Church K. [197] могут быть использованы как в качестве теоретических основ, так и для решения задач морфологического анализа в прикладной лингвистике.

Вопросы системы автоматической проверки правописания в текстовых данных рассмотрены в исследованиях учёных дальнего зарубежья, таких как Damerau F.J. [211], Abney S. [192], Angell R.C. [195], Levenshtein V.I. [235], Hausser R. [220], Church K.W. [198]. Следует отметить, что предложенные ими методы нашли свое практическое применение в процессе разработки алгоритмов и автоматических средств проверки орфографии в текстах на иностранных языках.

Проблемы автоматического перевода на основе статических моделей были в центре внимания зарубежных учёных. В частности, данные вопросы нашли свое

решение в работах Masterman M. [242-243], Simard M. [247], Ahmed F. [193], Lopez A. [238], Pujianto E. [246].

Модели, предложенные Cooper F.S. [206-208], George E.B. [216], Hunt, A.J. [223], Stolcke A. [250], Klatt D.H. [227-230], Van Santen J.P.H. [255], Massaro D.W. [240-241], Olive J. [245], Vitale T. [256], составляют основу синтеза речи.

В Российской Федерации отмечаются значительные достижения в области компьютерной лингвистики с использованием математического и компьютерного моделирования. Большой вклад в развитие данного направления внесли такие ученые, как А.Г.Сбоев [138-139], Д.Ш.Сулейманов [145-147], Ю.Г.Зеленков [93], Е.В. Котельников [111-112], Д.В.Михайлов [125-126], Ф.А.Мурзин [127], И.И.Быстров [64], А.В.Заболеева-Зотова [81], И.А.Минаков [123-124], С.В.Смирнов [140], А.В. Пруцков [134-135], В.Я.Чучупал [185-186], В.Н.Сорокин [143-144], Н.Г.Загоруйко [182-183], О.Ф.Кривнова [113-114], А.А.Зализняк [84], А.И.Евсеева [78-79], Р.К.Потапова [133], К.А.Дроздова [76], В.И.Галунов [65].

В странах СНГ в отдельных областях компьютерной лингвистики известны труды белорусских ученых Б.М. Лобанова [116-118], Л.И. Сирульника [182-183], Е.Б. Карневской [101], украинских исследователей Т.В. Людовика [119-120], Н.Н. Сажка [137], А.В. Анисимова [55]. По вопросам компьютерных словарей, тезаурусов и машинного фонда русского языка заметными исследованиями являются работы Ю.Н. Караулова [100], А.С. Нариньяни [129], В.Ш. Рубашкина [136], С.В. Лесникова [115].

Опираясь на возможности современных информационных технологий и средств крупномасштабных вычислений во втором тысячелетии, ученые Е.Б.Козеренко [102-103], Г.Г.Белоногов [63], С.О.Шереметьева [187], Е.Г.Жиляков [80], В.Н.Захаров [92], И.В. Ковалева [104], Ф.А. Мурзин [127], Н.Н. Черник [184] представили математические модели и компьютерные методы обработки текстовых данных. Результаты исследования показали, что вопрос распространения компьютерной лингвистики до сих пор не решен до конца.

Для автоматической обработки информации на естественном языке, относящейся к категории «*Natural Language Processing, NLP*», разработаны

различные методы, математические и компьютерные модели, комплекс алгоритмов. Такие программы по желанию пользователя могут обработать и текстовую, и голосовую информацию, сохраненную в электронной памяти. Ниже приведен список наиболее популярных компьютерных синтезаторов речи: word2vec [285]; Лемматизатор правописания русского языка до исправлений (А.Е. Поляков) [307]; Apache OpenNLP (The Apache Software Foundation, Incubator)) [308]; Mystem (Илья Сегалович, Виталий Титов, компания Яндекс) [276]; Ngram Statistics Package - NSP (Ted Pedersen) [278], Langsoft [272]; Программный продукт компании LingSoft (Финляндия) [274]; Система StarLing (С.А. Старостин)) [283]; MonoConc/ParaConc (Michael Barlow Dept of Linguistics, Rice University, Texas, USA) [275]; WordSmith Tools (Mike Scott, School of English, University of Liverpool) [286]; Galaktika-ZOOM (корпорация Галактика, Москва) [270], netXtract (Relevant Software Inc.) [277]; Paai's text utilities (Dr. J.J. Paijmans, Нидерланд) [293].

Передовые системы обработки информации на естественном языке и программные продукты, такие как Portal Language bab.la (Андреас Шретер, Патрик Укер) [294], онлайн-словари издательства «ЭТС» (ETS Publishing House) [301], словари Ожегова и Зализняка (С.А. Старостин) [299], Lexical FreeNet (Datamuse Corporation) [300], WordNet (Cognitive Science Laboratory, Princeton University) [295], Babylon.com (Babilon.com Ltd.) [290], ABBYY Lingvo-11 (ABBYY Software House) [289], электронный словарь PROMT (ОАО «ПРОект МТ») [297], Яндекс. Словари (Яндекс) [298], Lexical FreeNet (Datamuse Corporation) [273] обеспечивает как офлайн, так и онлайн многоязычный словарь. Из этого следует, что нам необходимо решить задачу проектирования и создания систем электронных словарей и тезаурусов таджикского языка, многоязычный словарь, отражающий все возможные толкования слов данного языка и определяющий связи между ними в виде компьютерного тезауруса. Эффективному процессу изучения основ естественного языка пользователю способствуют практические возможности и информационная поддержка этих систем.

Список наиболее известных систем, относящихся к области синтеза речи: Sakrament Text-to-Speech Engine (кампания “Сакрамент”) [280], CSLU Toolkit

(Center for Spoken Language Understanding) [292], CMU Artificial Intelligence Repository (Carnegie Mellon University, School of Computer Science) [291], Речевые программы в Websound.ru (Александр Радзишевский) [288], Speech technology (Центр речевых технолоий, С-Петербург) [282], Fonix Speech (Fonix Co) [269], Text-To-Speech Converter for MS Word (Exiton) [284], Govorilka (Anton Ryazanov) [271], BookMania (Sergey Shishmintzev, Киев) [268], Speech Synthesis and Recognition Laboratory (Минск) [281]. Отмечается, что некоторые системы адаптированы для чтения текстов вслух на любом языке. Но в процессе непосредственной работы с ними выяснилось, что приписываемые им возможности совершенно не подтверждаются, поскольку высокое качество синтезированной речи напрямую связано со спецификой звучности языка. Отсюда следует, что программная система, разработанная для конкретного языка, не может одинаково успешно выполнять свои задачи на любом другом языке. Такая ситуация ставит вопрос о проектировании и организации систем компьютерного синтеза для таджикской речи.

Также важно отметить, что первые системы проверки орфографии стали доступны на персональных компьютерах в конце 70-х годов прошлого века. Группа из шести лингвистов из Джорджтаунского университета разработала первую упомянутую систему для IBM. Эта система была установлена на персональные компьютеры CP/M и TRS-80 в 80-е годы, а затем в 1981 году появились первые пакеты для IBM PC. Эти системы проверки представляли собой автономные программы, многие из которых можно было запускать из пакетов текстовых процессоров. В настоящее время все пакеты, работающие с текстовой информацией, имеют встроенные или устанавливающие системы автоматической проверки орфографии, поддерживающие более ста языков. В России широко используется популярный пакет «ОРФО» [287], проверяющий орфографию на шести языках: русском, английском, немецком, французском, испанском и украинском. Для всех языков есть удобный инструмент для добавления новых слов со всеми формами этих слов. Пакет «ОРФО» поддерживает все продукты Microsoft, а также подключается к компьютерным продуктам PageMaker [312], WordPerfect

[314], WordPro [315] и Quark XPress [313]. Во всех случаях используется один и тот же пользовательский словарь. Например, MS Office [310] и OpenOfficeOrg [279] по умолчанию поддерживают более 100 языков. Каждая система проверки опирается на базу данных словарей, словоформ и элементов слов, таких как префиксы, корни и суффиксы. К сожалению, в этот пакет не входит проверка орфографии на таджикском языке. На этом основании вопрос проектирования и организации систем автоматической проверки орфографии таджикского языка становится актуальной проблемой для научных кругов.

Системы автоматического перевода Natural Language Projects at ISI (Univ. of Southern California/Information Science Inst.) [311], «Автоматический словарь Мулитран» (pom@aha.ru) [306], Translate.Ru (ООО ПРОМТ) [304], LEO (Department of Informatics, Technische Universitat, Munchen) [309], Perevodov.net (Ectaco) [303], Проекты PIT ZS (Исследовательский институт искусственного интеллекта) [296], Computer Aided Translation (Google) [302], Яндекс Переводчик (translate.yandex.com) [305] в основном обеспечивают онлайн-перевод текста или документа. Эти системы поддерживают парные переводы с разных языков. Основным средством создания машинных переводов на основе статистических методов является проектирование и разработка параллельных ресурсов для двух языков. На основе указанных случаев необходимо решить вопросы машинного перевода текста с таджикского языка. В первую очередь необходимо разработать параллельные таджикско-российские ресурсы и внедрить в систему перевода.

§ 1.2. Обзор результатов исследований компьютерной лингвистики таджикского языка

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 [148-181] обратились к совершенно новой отрасли

информационно-коммуникативных технологий – компьютерной лингвистике таджикского языка.

Научные задачи, связанные с компьютерной лингвистикой таджикского языка, нашли свое решение в исследованиях таких таджикистанских ученых, как М.А. Исмоилов [94-99], С.А. Зарипов [85-91], О.М. Солиев [142], Г.М. Довудов [74-75], Л.А. Гращенко [68-72], А.А. Косимов [105-110], А.Г. Гуломсафдаров [73; 96; 98], К. С. Бахтеев [60-62; 106; 110], К. А. Евазов [189-191], Ш. Н. Ашурова [56-59], А. А. Назаров [128].

Академик З.Д. Усманов, возглавивший группу исследователей, определил перспективные цели и задачи изысканий в области компьютерной лингвистики таджикского языка. К ним он отнес моделирование простого двусоставного предложения, создание таджикских графических драйверов, решение задачи стандартизации печатной продукции, автоматический синтез таджикского текста, обмен системами графических линий, анализ и автоматическое морфологическое распознавание, распознавание авторов таджикских текстов.

В разработке указанного научного направления принял участие коллектив молодых исследователей:

- в работе кандидатской диссертации С.А. Зарипова предложены методы моделирования простого двусоставного предложения на английском языке для создания системы автоматического перевода [263];

- в кандидатской диссертации О.М. Солиева представлены математические модели пропорционального размещения символов на клавиатуре свободной конфигурации, новые варианты эргономического представления таджикского шрифта на клавиатуре компьютера для создания драйвера TajGraph [265];

- в кандидатской диссертации Л.А. Гращенко разработан эффективный расчетный алгоритм преобразования линейных графических систем в таджикско-персидский язык для его применения в виде проблемно-ориентированного программного комплекса [259];

- в кандидатской диссертации Г.М. Довудова разработаны и реализованы в виде программного комплекса модели и алгоритмы автоматизации процесса морфологического анализа таджикских словоформ [262];

- в кандидатской диссертации А.А. Косимова разработаны модели, численные методы и алгоритмы процесса автоматического распознавания авторства таджикских текстов, которые реализованы в виде комплексного компьютерного программного обеспечения [264].

Для системы автоматической обработки элементов текстовой информации на таджикском языке были разработаны компьютерный алфавит; буква N-грамма; слоговая структура; слоговая структура слов; формы слов и их употребление; анаграммы; N-грамм слов; морфы; префиксы и суффиксы; корни слов; фразы и типы предложений. Для реализации актуальных задач был создан комплекс компьютерных программ и веб-программ: пакеты автоматической проверки орфографии, синтеза и распознавания речи, системы голосового управления конечными устройствами, а также системы автоматического машинного перевода.

В процессе разработки и внедрения автоматизированных информационных систем обработки текстовой информации на таджикском языке был проанализирован комплекс таджикских правил правописания [11-14]. Вопрос совместимости текстовых элементов был учтен в источнике данных по интерпретационной культуре таджикского языка [15-16]. В процессе обработки компьютерных моделей в качестве основных требований использовались также научные результаты В.С. Расторгуевой [20], Д.М. Искандаровой [21], Т.С. Шокирова [54, 188], А.А. Одинаева [130-131], А.Ю. Фомина [266], Д.Д. Собирова [141], А. Мамадназарова [121-122].

Все вышеперечисленные вопросы являются фундаментальной основой развития научной области компьютерной лингвистики таджикского языка и подтверждают важность темы диссертации.

Значительные результаты были достигнуты в процессе научных исследований с 2005 года в научной школе компьютерной лингвистики при

Институте математики А. Джураева НАНТ и ХПИТТУ имени академика М.С. Осими.

Основные результаты указанной научной школы можно разделить на три основные группы.

Первую группу составляют работы, предназначенные для непосредственного практического использования:

- драйвер TajGraph 1.0 для размещения таджикского шрифта на клавиатуре компьютера;

- русско-таджикско-русские компьютерные словари;

- компьютерное звучание таджикского текста;

- автоматическое преобразование таджикского текста в стандартную графику;

- таджикский языковой пакет для системы Open Office Org;

- таджикский словарь для проверки орфографии в Microsoft Office.

На практическом уровне все упомянутые исследования доказали свою верность и надежность

Вторую группу составляют исследования, нацеленные на практическое применение в ближайшем будущем. Их внедрение, к сожалению, не было осуществлено должным образом, хотя они были реализованы на современном уровне. Это произошло из-за недостаточной рекламы и продвижения соответствующих программных продуктов, а также недопонимания ответственных сотрудников, не заинтересованых в использовании новых информационных технологий. Эти исследования включают в себя:

- эргономичное размещение английских, русских и таджикских букв, а также алфавита эсперанто на клавиатуре компьютера;

- автоматическое преобразование таджикского текста в стандартную графику;

- таджикско-персидский конвертер линейных графических систем;

- эргономичное размещение таджикского алфавита на клавиатуре мобильного телефона.

Третья группа исследований состоит в основном из теоретических исследований, которая решает такие вопросы, как:

- частота буквенных униграмм, диаграмм и триграмм в таджикском языке;
- слоговая основа и частотность таджикских словоформ, слоговая структура;
- автоматическое распознавание глаголов в предложениях на таджикском языке;
- αβ-кодирование слов и предложений на естественных языках;
- «стоп-слова» из таджикского словаря;
- кодирование слов естественного языка в алфавитном порядке;
- морфемная и частотно-морфемная словарная база таджикского литературного языка;
- автоматическая система морфологического анализа словоформ таджикского языка;
- связи форм слов и их употребление.

Важнейшим результатом всех теоретических исследований третьей группы может стать создание компьютерной системы морфологического анализа таджикских словоформ. Он основан на разработанных авторами позиционной кодировке грамматик словоформ таджикского языка. Это открывает чрезвычайно разнообразные возможности для теоретических и практических решений в различных областях компьютерной лингвистики.

К важным проектам, в которых система компьютерного морфологического анализа выполняет большое количество подготовительных технических процедур, относятся:

- разработка словаря корней и основ таджикского языка;
- разработка грамматического словаря таджикского языка;
- разработка инверсионного словаря (обратного алфавита);
- создание компьютерного тезауруса таджикского языка;
- создание автоматической системы выделения ключевых слов;
- изучение грамматических конструкций таджикских предложений;

- разработка гипотезы автоматического компьютерного перевода с таджикского языка на таджикский язык;

- разработка автоматической системы оценки сложности текста;

- создание базы данных таджикских фраз.

Другой важный вопрос, который следует решить исходя из вышесказанного – регулярное сотрудничество с Microsoft для облегчения доступа к нашим программным продуктам для их работы в новых версиях операционных систем Windows. Этому мы научились из неудачного опыта разработки и распространения в отрасли пакета проверки правописания на таджикском языке. Наш программный комплекс, работавший на Windows XP/7/Vista и получивший множество положительных отзывов, оказался непригоден для работы с более новыми версиями Windows.

За последние годы достигнут значительный прогресс в области компьютерной лингвистики и продвижения технологий разработки Государственной стратегии в Республике Таджикистан. Разработаны и внедрены системы автоматической обработки информации на таджикском языке для создания электронных словарей, синтеза речи, автоматической проверки орфографии и компьютерного перевода текстов.

§ 1.3. Описание результатов исследований по информационным системам 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:

$$TajLINGVO = \{T, P, A, TE, I, R\} \quad (1.1)$$

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 (рис. 1.1), 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 информационных ресурсов*» 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.

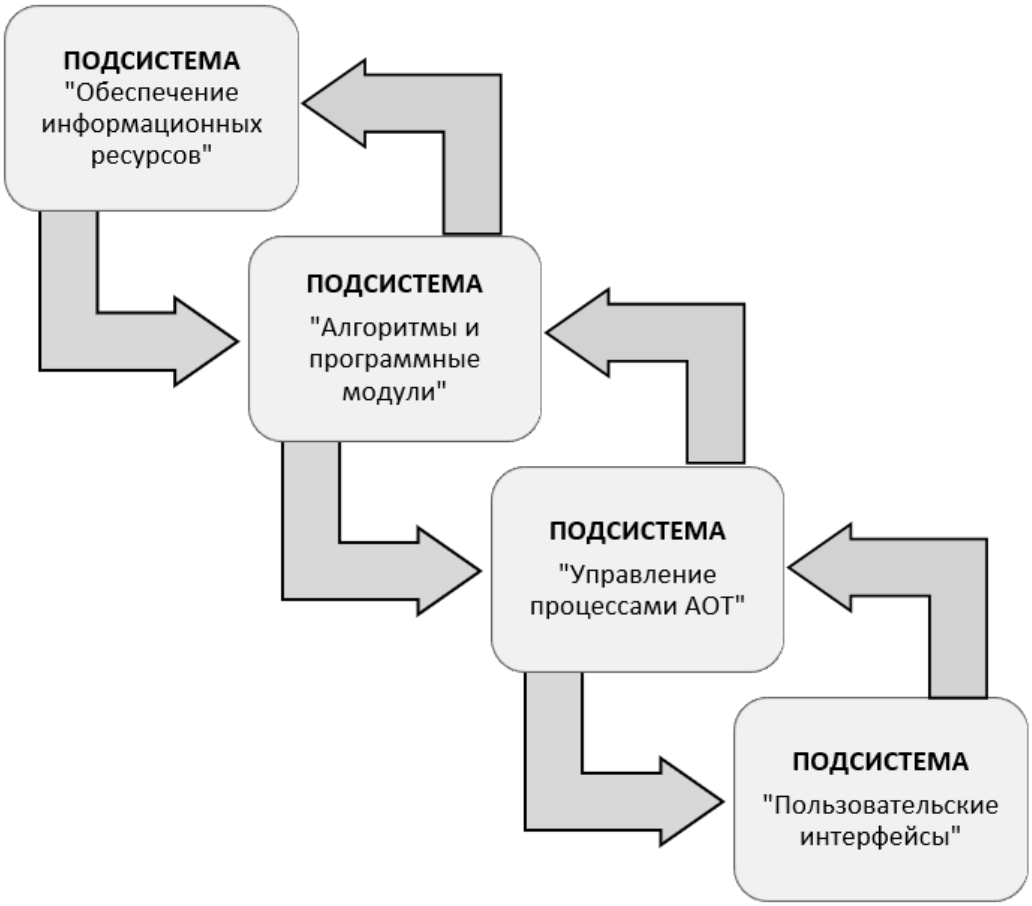

Рисунок 1.1. - Схема логической структуры TajLINGVO

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

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:

$$FM = \{LC, SW, SS, DS, GS, CS\} \quad (1.2)$$

где,

LC (лингвистическая структура) – источник текстовой информации для формирования языковых ресурсов;

SW (word structure) – набор составленных слов, образованных от LT;

SS (semantic structure) – совокупность семантических структур, описывающих SW;

DS (data structure) – набор лингвистических структур, SS образовались в SW;

GS (grammar structure) – комплекс грамматических явлений, основанных на грамматических правилах естественного языка;

CS (code structure) – 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 (рис. 1.2).

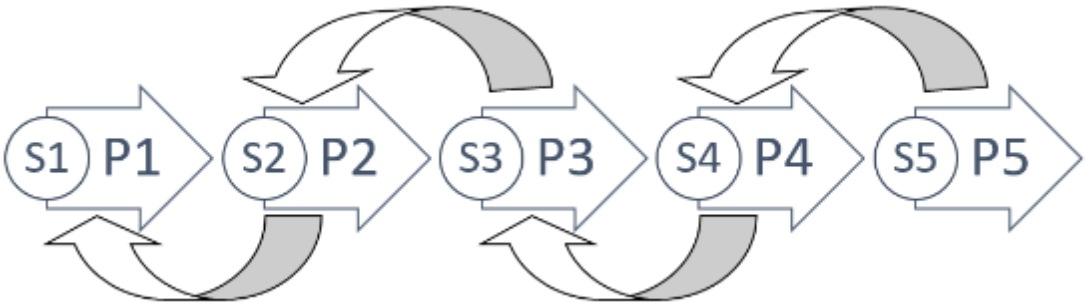

Рисунок 1.2. - План информационной модели TajLINGVO

Проанализируем другие функции, используемые в информационной модели системы TajLINGVO:

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

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.

Функциональная модель системы. В настоящее время для создания функциональных программных моделей используются стандартные методологии и языки моделирования, такие как IDEF, DFD, UML. Универсальный язык моделирования используется для графического описания и объектного моделирования информационных систем. В рамках визуального моделирования язык UML широко использует объектно-ориентированные и предметно-ориентированные методы. В UML определены четыре основных типа моделей:

- статическая модель(неподвижная) (static model);
- динамическая модель (движущаяся) (dynamic model);
- модель взаимодействия объектов (interaction model);
- физическая модель (physical model).

Практическая модель информационной системы представляется в виде комбинации трех типов диаграмм: вариантов использования, деятельности и классов.

Описание вариантов использования определяет общие границы и содержательный объем предметной области, подлежащей моделированию на начальных этапах проектирования системы. Описание планирует общие требования к практическому поведению системы и раскрывает ее в виде логических и реальных объектов.

С помощью диаграммы деятельности анализируется работа системы в контексте потоков данных и процессов управления. Изображение отражает абстрактный алгоритм этапа жизни системного объекта, однако в другой схеме, чем в которой описаны основные шаги алгоритма.

Для описания структуры объектов системы, взаимосвязи объектов, символов, действий и процедур используется изображение классов.

Надежная работа компьютерной системы морфологического анализа позволяет обеспечить быстрое освоение корпуса национального таджикского языка, разработку второй и относительно полной гипотезы двунаправленного машинного перевода между русским и таджикским языками, а также внедрение и развитие информационные технологии обработки семантики.

Наличие морфологического анализатора не мешает прогрессу, а наоборот, помогает развитию других направлений компьютерной лингвистики, например, созданию системы автоматического распознавания автора текста и различных высокочастотных словарей (корнеплоды) и корни слов, формы слов и их грамматика, части речи и грамматические категории и т.д.

Структурная модель системы состоит из классов и их взаимодействия:

- сущностный класс Text_Element реализует технологию обработки текстовой информации составной структуры;

- класс объектов Text_Element реализует технологию обработки и управления элементами текстовых данных типа: буквы, биграммы и триграммы, слоги, слова, словосочетания, предложения;

- класс объектов Lingustic_Corpus осуществляет технологию выражения структуры текстовых элементов в параллельном предложении;

- класс таджикского интерфейса преобразования текста в речь предоставляет доступ к внешней библиотеке TTextSpeechLib, выполняющей синтез речи на основе слоговых звуков в списке классов типа DB_Slog_Zvuk;

- класс интерфейса MultyGANJ обеспечивает доступ к внешней библиотеке TThesaurusLib, которая выполняет действия по управлению лингвистическим тезаурусом на основе информации в списке класса Taj_Thesaurus;

- класс интерфейса TajSPELL обеспечивает доступ к внешней библиотеке TspellCheckLib, которая выполняет операции автоматической проверки правописания на основе заданных данных в виде списка классов Spell_Dictionary;

- класс интерфейса www.tarjumon.tj обеспечивает доступ к внешней библиотеке TTranslateLib, которая выполняет автоматический машинный перевод

текстов на основе параллельного хранения текстов на таджикском, русском и английском языках в списке таких классов, как tj_ru_en_Parallel_Corpus.

Программные модули управления процессами синтеза речи, компьютерный тезаурус, автоматическая проверка орфографии и машинный перевод входят в состав графических интерфейсов аппаратно-программного обеспечения системы TajLINGVO.

Информационное обеспечение системы TajLINGVO соответственно реализуется на основе данных «слог-голос», «таджикского компьютерного тезауруса», XML-файлов структур слов, данных «корпуса параллельных текстов». Управление данными осуществляется посредством языка запросов SQL в системе управления базами данных MySQL, процессы управления реализуются с помощью средств программирования MS Visual Studio.Net и веб-программирования PHP, MySQL.

В процессе исследования рассмотрены методы проектирования систем автоматической обработки текстовой информации на примере таджикского языка – TajLINGVO. На основе методов SADT предложена логическая структура системы. С целью отражения возможностей информационного обеспечения системы и взаимодействия с процессами обработки информации предложен план работы информационной модели системы. На основе возможностей языка UML были созданы схемы вариантов использования, функционала и классов, в которых представлено описание практического примера системы TajLINGVO.

Результаты исследования послужили основой для проведения других научных исследований в области компьютерной лингвистики, включая анализ основных процессов автоматической обработки текстовой информации. Выявлена эффективность использования модели при решении конкретных практических задач, включая компьютерную обработку тезаурусов, автоматическую проверку орфографии, синтез речи и машинный перевод. Результаты исследований и разработанные информационные системы доступны в Интернете по адресу www.tajlingvo.tj

В контексте процесса развития компьютерной лингвистики можно сделать вывод, что выбранная тема исследования важна, теоретические и практические аспекты достоверны, инновации реальны и необходимы, объем и степень их применения всеобъемлющи и убедительны, работа направлена на развитие таджикского языка с использованием возможностей информационных технологий. Разработка теоретических основ и полученные практические результаты в исследовании достигнуты благодаря широкому использованию математических моделей на высоком уровне программирования. В частности:

1. «*Электронные словари*», предлагаемые составителями, полны, просты и очень удобны для использования в Интернете. Безусловно, они могут внести реальный вклад в повышение доступности электронных энциклопедий культуры, остро нуждающиеся в настоящее время в обогащении.

2. Проект «*Тезаурус таджикского языка*» с точки зрения охвата словарного состава таджикского языка по сравнению со «Словарем таджикского языка» и «Толковым словарем таджикского языка» в части электронного использования имеет множество технических преимуществ (отображение и чтение разнообразных слов, ссылки, многозначные слова, повторы, виды переносов и др.).

3. Известно, что с появлением компьютерной печати проблема орфографии таджикского языка и проблемы ее исправления, к сожалению, не решены до сих пор. С этой точки зрения обсуждаемая инициатива исследователей и программистов достойна поддержки, ведь предлагаемый ими проект «*Программный комплекс TajSpell – проверка правописания таджикского языка*» может устранить часть этих проблем.

4. Вопрос произношения является одной из наиболее актуальных проблем таджикской орфоэпии (фонетики), которая до сих пор рассматривается традиционно-историческим методом. Внедрение проекта «*Компьютерный синтез речи таджикского текста*» может оказать существенную помощь в компьютерном произношении звуков и слов таджикского языка, в его совместимости с орфографией таджикского языка.

5. Проект «*Web--приложение параллельных корпусов таджикского и английского языков*» является относительно новым явлением в традиционном таджикском языкознании. Поскольку в советское время он считался одним из второстепенных вопросов национального языкознания. Благодаря независимости Республики Таджикистан данный вопрос стал приоритетным, в связи с чем Web-приложение задуманных параллельных структур таджикского и английского языков может послужить практическим руководством.

6. Приложение параллельных корпусов таджикского и русского языков – это проект, который давно должен был быть реализован. Во многих постсоветских государствах эта проблема была решена очень быстро. Однако в Таджикистане, несмотря на его богатую лексикографическую историю (издание большого количества таджикско-русских и русско-таджикских словарей в период советской эпохи и период независимости) данный вопрос все еще находится на стадии рассмотрения.

7. «*Автоматический перевод текста на таджикский язык*» можно считать одним из необходимых проектов, поскольку он может заполнить многие пробелы в существующей программе.

Выводы по первой главе

В жизни современного общества автоматизированные информационные технологии играют важную роль.

Широкое использование информационно-коммуникативных технологий в Таджикистане вызвал большой интерес у исследователей в области математики, информационных технологий и лингвистики.

Огромная заслуга в исследовании компьютерной лингвистики в Таджикистане принадлежит академику АНТ, профессору, д.ф.-м.н. З.Д. Усманову: он собрал вокруг себя талантливых молодых ученых и создал целую научную школу, главным направлением которой была компьютерная (математическая) лингвистика.

Основная деятельность указанной научной школы была направлена на решение трех основных задач:

1) создание трудов, предназначенных для непосредственного практического использования;

2) проведение исследований, нацеленных на практическое применение в ближайшем будущем;

3) ведение теоретических исследований.

Научные задачи, связанные с компьютерной лингвистикой таджикского языка, нашли свое решение в исследованиях таких таджикских ученых, как З.Д. Усманов, М.А. Исмоилов, С.А. Зарипов, О.М. Солиев, Х.А. Худойбердиев, Л.А. Гращенко, Г.М. Довудов, А.А. Косимов, А.Г. Гуломсафдаров, К.С. Бахтеев, К.А. Евазов, Ш.Н. Ашурова, А.А. Назаров.

В результате активной деятельности научного коллектива, был решен ряд задач в области компьютерной лингвистики: моделирование простого двусоставного предложения, создание таджикских графических драйверов, решение задачи стандартизации печатной продукции, автоматический синтез таджикского текста, обмен системами графических линий, анализ и автоматическое морфологическое распознавание, распознавание авторов таджикских текстов и т.п.

ГЛАВА 2. МЕТОДОЛОГИЯ КОМПЬЮТЕРНОГО АНАЛИЗА И СИНТЕЗА ЕСТЕСТВЕННОГО ЯЗЫКА

§2.1. Методы и функции анализа текста

Обработка текстовых данных является объектом многих специальностей, связанных с информационными технологиями. Как в процессе сбора, обработки и передачи текстовых данных, так и в процессе анализа и синтеза данных, знание методов лингвистического анализа, прежде всего, необходимо для успешной практики.

Понимание процесса формирования текста расширяет возможности формирования проблемных и решаемых задач относительно структурных элементов текста. Основы методов анализа текстовой информации могут быть использованы для автоматической определения смысла информации и в процессе принятия решений в профессиональной деятельности.

Текст (от лат. textus – ткань, плетенный, непрерывный) – это такая последовательность символических единиц, объединенных смысловой связью, основными характеристиками которых являются последовательность и целостность; человеческий речевой продукт, имеющий полный и подлинный смысл в виде документа, подвергнутого литературной обработке по виду этого документа; произведение, состоящее из названия и ряда специальных единиц, объединенных различными видами лексических, грамматических, логических, стилистических связей, имеет определенное значение.

Лингвистический анализ текста – это:

- анализ по определению системы языковых средств;
- анализ с целью определения особенностей стилистических элементов текста;
- анализ, изучающий структуру рабочих стилей и их речевую систему;
- анализ языковых средств с точки зрения теории информации;

- анализ комплекса текста языковых средств, которые объединены вокруг аспектов речевого общения;

- анализ языковых средств и их взаимосвязи в структурной части текста;

- комплексный анализ языковых средств в зависимости от языка, жанра и спектра текстовой информации.

В контексте проведенного исследования это понятие можно представить следующим образом: «лингвистический анализ текста – это метод исследования текста, который можно рассматривать как структурно-семантический анализ, направленный на тему содержания текста и цифровую информацию, содержащуюся в текстовой информации».

Компьютерная лингвистика – научное направление в области математического и компьютерного моделирования логических процессов человека для описания естественных языков.

Область деятельности компьютерных лингвистов – разработка алгоритмов, процедур и прикладных программ обработки лингвистической информации представлены в следующих работах [201; 209; 213; 215; 218; 224-225; 234; 244; 249; 251-252; 254].

Направлениями и задачами компьютерной лингвистики, нуждающиеся в научном исследовании, являются:

1. Обработка естественного языка (Natural language processing).

2. Проведение этапов обработки и анализа текста: синтаксического, морфологического, семантического.

3. Формирование языковых ресурсов (корпусов), создание и использование электронных текстовых ресурсов (корпусов).

4. Организация электронных словарей, тезаурусов, описаний (онтологий).

5. Автоматическая проверка орфографии.

6. Автоматический перевод текстов.

7. Автоматическое извлечение фактов из текста.

8. Автоматический вывод и маркировка текста.

9. Организация систем управления знаниями.

10. Организация вопросно-ответных систем.

11. Оптическое распознавание знаков.

12. Автоматическое распознавание речи.

13. Автоматический синтез речи.

14. Автоматический и машинный перевод.

15. Создание словаря с помощью компьютера и электронные словари.

В современном мире на базе точных наук возникает необходимость создания компьютерных программ для компьютерной или автоматической обработки текстов на естественном языке. Это связано с активным ростом текстовой информации в мире. Примеры такой обработки включают поиск, сортировку, фильтрацию и сбор информации из различных доступных источников, получение знаний и принятие решений. При разработке программ при интеграции и интерпретации программных обеспечений используются алгоритмы обработки текста на естественном языке. Они имеют дело с различными элементами текста, например, с обработкой предложений, слов, слогов, букв и т.д., которые представляют трудности.

Обработка естественного языка интересовала учёных на протяжении многих десятилетий. С начала 60-х годов прошлого века многие исследователи внесли значительный вклад в область анализа и осмысления понятий естественных языков. В частности, ученые В.Н. Сорокин, О.Ф. Кривнова, Г.Г. Белоногов, А.В. Анисимов, Д.Ш. Сулейманов представили методы и модели обработки текстовой информации на русском языке. Однако лучшие способы решения проблем, связанных с этой отраслью науки, показали учёные Р.Л. Мерсер, Р. Гришман, М. Джонсон, А.М. Либерман, В.Дж Хатчинс. Они внесли значительный вклад в создание методов и моделей в обработки текстовой информации на английском языке. Следовательно автоматическая обработка естественных языков, например, таджикского языка, в области компьютерной лингвистики до сих пор остается трудной, требующего решения задачей. В исследованиях по обработке таджикских слов под руководством З.Д. Усманова решены многие вопросы в этой области, но остается еще много научных вопросов, находящихся за пределами изучения. Причины

отсутствия исследований вопроса автоматической обработки на таджикском языке, следующие:

1. Компьютерная модель таджикского языка, как и других языков мира, например, английского, русского, арабского, не определена.

2. Фонетические, орфографические и грамматические правила таджикского языка не структурированы полностью с точки зрения математических и компьютерных средств.

3. Не сформирован единый структурный резерв (корпус) таджикского языка с возможностью обеспечения как текстовой, так и звуковой информацией.

Обработка естественного языка включает в себя ряд операций, в том числе анализ текста – графемный, лексический, морфологический, синтаксический и семантический анализ. Анализ смысла текста не учитывался при выполнении данного исследования. По этой причине семантический анализ включен в планы будущих научных работ. В данном научном исследовании больше внимания уделяется нескольким методам анализа, таким как графемный, морфологический и синтаксический.

В приведенной ниже диаграмме представлено описание подразделов этапа анализа, составляющих основу исследования (рис. 2.1.).

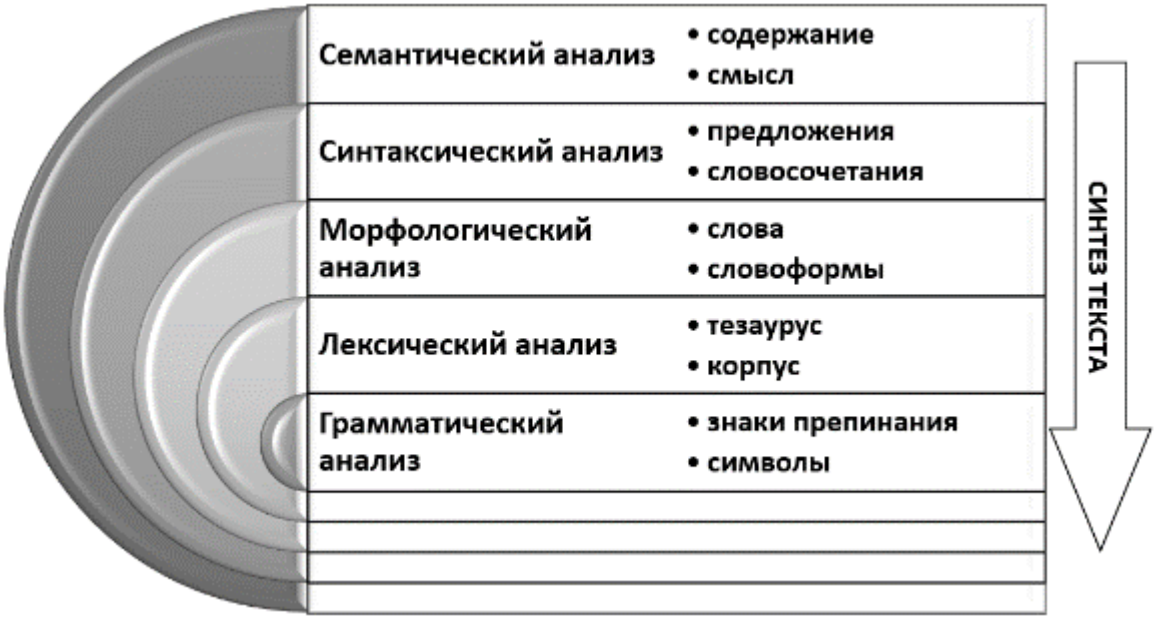

Рисунок 2.1. - Схема этапов обработки текстовых данных

*1. Грамматический анализ текста***,** в ходе которого элементы текстовой информации выделяются из совокупности информации в виде отдельных элементов: предложений, слов, слогов, букв, знаков препинания.

Сегментация текста – это основной этап обработки естественного языка, который представляет собой лингвистическую концепцию, используемую в информатике для разделения потока текста на важные элементы. В информатике процесс сегментации текста является частью лексического анализа. При сегментации текста фрагмент текста сначала делится на предложения – этот процесс называется сегментацией предложения. После этого каждое предложение делится на слова, и так деление продолжается до появления символов.

Задача графемного анализа текста – создать основу для дальнейшего морфологического и синтаксического анализа на основе выбора слов, словоформ и предложений. Кроме того, метод графемного анализа позволяет уделить внимание выделению и организации ненормативных элементов и обработать эти элементы, например: элементы формирования текста – толщина шрифта, форма наклона письма, подчеркивание; структурные элементы текста – заголовки, абзацы, сноски; различные текстовые элементы, не являющиеся словами – цифры, даты в числовой форме, буквенные и числовые комбинации и т.п.

В данной научной работе методы лексического анализа были использованы в процессе формирования синтеза речи в таджикском языке. Значимые результаты были достигнуты в процессе разработки алгоритмов деления слова на слоги и последующего соединения слогов в программном обеспечении для озвучивания текста через компьютер на таджикском языке.

2. Лексический анализ текста – анализ слова во всех видах системных отношений, существующих в языке. Анализируются синонимы, антонимы, полисемичные слова, омонимы, анаграммы, которые служат наиболее точными средствами передачи информации.

В процессе словарного анализа выделяются все лексические особенности языка, наблюдаемые в тексте: словоформы, словоупотребление и слова.

В данной диссертационной работе методы лексического анализа использовались в процессе разработки электронных словарей, таджикского электронного тезауруса, а также при разработке алгоритмов и программных модулей проверки правописания текста.

3. Морфологический анализ, в ходе которого выделяется грамматическая основа, определяется часть речи, слова представляются в виде словарной статьи.

После сегментации текста следующим этапом является морфологический анализ текста, который проводится с целью изучения образования слов из сравнительно небольших смысловых единиц, называемых морфемами. Морфемы имеют два основных типа: основу и аффикс. Основа является главной морфемой слова и выражает основное значение. Аффиксы вносят дополнительные значения разного типа и делятся на префиксы, суффиксы и постфиксы. Приставки стоят перед корнем, суффиксы – после корня, а инфиксы – внутри основы. Процесс «лемматизации» – ключевой этап морфологического анализа, на котором определяется лемма или основа каждой лексемы.

В научном исследовании наиболее хорошие результаты морфологического анализа были получены и применены при автоматической проверке правописания и машинном переводе текста на таджикский язык. Лемматизация как сравнение нелемматизированного текста с представлением класса UML может вызвать несогласованность и повлиять на точность перевода. Кстати, в процессе лемматизации аффиксы, вырезаемые из леммы, используются для грамматических целей, и по сравнению с представлением классов UML никакой грамматической информации не требуется.

В данном исследовании разработаны методы, компьютерные модели и программные модули для лемматизации на основе правил таджикского языка.

4. Синтаксический анализ, в процессе которого определяются синтаксические отношения между словами в предложении, в разных частях предложения создается синтаксическая и грамматическая структура.

Дерево синтаксического анализа формируется как особый способ графического изображения грамматических связей в предложении. Традиционно

существует два типа разделения; «сверху вниз» (целевой) и «снизу-вверх» (ориентированный на данные). Анализатор ищет такое дерево, которое можно построить от корневого узла до листьев. С другой стороны, восходящий анализ начинает с листьев (ввода слова) и пытается построить дерево, используя набор грамматических правил.

Синтаксический анализ текста используется в научной работе по нескольким причинам: анализ может предоставить исследователю дерево анализа и набор зависимостей. Такие зависимости на самом деле представляют собой отношения в различных синтаксических структурах ограничений естественного языка. В исследованиях такие логические связи сравниваются с эквивалентными отношениями в словарях и устоявшимися правилами таджикского языка, отбором предложений в процессе машинного перевода.

5. Семантический анализ, в ходе которого устанавливаются смысловые связи между синтаксическими группами и словами, определяются семантические отношения.

Семантический анализ используется для определения различных компонентов предложения и анализа входного текста для извлечения его известных значений, то есть прямого или ясного значения предложения. В ходе семантического анализа значения предложений выражаются через логическую форму. Существуют разные способы анализа семантики текста на естественном языке, и обычно он выполняется в два этапа: уровневый семантический анализ и глубокий семантический анализ.

В научной работе не использовались методы семантического анализа текста, поскольку определение смысла и логического содержания текстовой информации не являлось объектом исследования.

Каждый из вышеописанных методов анализа представляет собой набор независимых задач. Каждое задание в большинстве случаев не имеет практического применения изолированно, а используется как составная часть комплексных задач, таких как:

1. Разделение текста.

2. Определение формы слова и словоупотребления.

3. Морфологический анализ и подбор морфем.

4. Исправление ошибок, проверка орфографии.

5. Лемматизация.

6. Анализ синтаксиса предложения.

7. Автоматический подбор терминов.

8. Распознавание субъектов, имеющих название.

9. Определение фактов «объект-субъект-действие».

10. Выбор значений текстовых элементов.

11. Оценка смыслового сходства.

12. Выявление связей между текстовыми объектами.

Исследования упомянутых выше методов и задач не охватывают полностью все области анализа текста. Расширение круга методов анализа текста, а также поиск зависимостей и связей единиц анализа текста с комплексными процессами образовали различные исследовательские подходы и методы. Были исследованы различия между лингвистическими единицами анализа текста и систематическим описанием разработанных методов анализа текста и задач, решаемых при реализации информационных систем.

Некоторые современные подходы и методы обработки текстов на естественном языке: машинное обучение, искусственный интеллект, преобразование последовательностей, статистическая обработка данных могут быть использованы для решения задач компьютерной лингвистики. Некоторые из перечисленных подходов были использованы в данном научном исследовании, а именно: преобразование последовательностей, обработка статистических данных.

В работе нашли свое решение некоторые задачи компьютерной лингвистики таджикского языка, например, автоматическая проверка орфографии, машинный перевод и автоматический синтез речи на основе лингвистических правил и ресурсов, в том числе: несортированные текстовые ресурсы (корпус); тезаурусы, онтологии и электронные словари.

§2.2. Математические модели обработки текста на естественном языке

При разработке набора терминов и понятий для описания структуры естественных языков можно положиться на математическую лингвистику. В связи с возникшей необходимостью исследования в области математической лингвистики получили широкое распространение в Республике Таджикистан начиная с 90-х годов прошлого века.

В Республике Таджикистан основоположником научных исследований в области математической лингвистики таджикского языка считается академик НАНТ, доктор физико-математических наук, профессор З.Д.Усманов. Предложенные им математические методы и модели стали главным достижением таджикской научной школы в области компьютерной лингвистики.

В более чем ста научных статьях и монографиях З.Д. Усманова исследованы математические модели и методы обработки естественного языка.

К наиболее важным исследованиям З.Д. Усманова можно отнести: применение γ-классификатора для распознавания однородных объектов; автоматический поиск и агрегирование статистических закономерностей анаграмм; об анаграммах словоформы n-грамм; использование γ-классификатора для разметки печатного текста; о цифровом изображении текста и его применении; моделирование мозга анаграмматически искаженного текста; об эвристическом кластеризаторе, о проблеме автоматического распознавания значения согласных; применение n-граммы в распознавании схожих текстов и кодировании предложений; моделирование нахождения словоформ и использования слов для решения задач морфологического анализа; задача алфавитного кодирования слов естественных языков; моделирование слоговой структуры для реализации синтеза речи; моделирование размещения алфавита таджикского языка на клавиатуре и мобильных устройствах.

В последующих главах и отдельных разделах подробно описаны математические модели и методы обработки текстовой информации на таджикском языке.

О слоговом строе слов таджикского языка. Слог – это единица речи, состоящая из одного или нескольких звуков, образующих тесное фонетическое единство. Согласно другой лингвистической концепции, слог – это звук или совокупность звуков в слове, произносимый одним выдыхательным толчком.

С учетом задач исследования, представляется важным изучение закономерностей таджикского языка, связанных с понятием *слога*, *слоговой структуры слова*.

Каждое слово представляет собой заданную последовательность букв. Заменяя гласные на единицы, а согласные на нули (букву «й» мы считаем согласной), мы превращаем слово в упорядоченный набор нулей и единиц. Такое преобразование называется кодированием слова, а полученный результат, то есть запись, слоговой структурой слова.

Размер структуры $W_{0,1}^{*}$ — это количество букв, составляющих слово W, или количество символов (двоичных символов), используемых при написании $W_{0,1}^{*}$. Два слова имеют одинаковую структуру, если их представление в двоичной записи одинаково, в противном случае говорят, что они имеют разную структуру. Понятно, что структуры могут быть идентичными, если они имеют одинаковый размер. Известно, что каждому слову W соответствует один и только один образ $W_{0,1}^{*}$. Фактически каждому естественному языку соответствует несколько слов $W_{0,1}^{*}$ одновременно. Это означает, что разные слова с одинаковым количеством букв могут иметь одинаковую слоговую структуру. Например, такие слова, как «дилшод», «кардам» соответствуют структуре типа «010010».

Результаты, которые будут представлены далее в виде показателей, основаны на статистической обработке выборки из произведений, объём которых составляет 458628 слов. Образы этих слов, т. е. соответствующие слоговые структуры, представленные набором $\{W_{0,1}^{*}\}$, были затем предметом статистического анализа.

Всего в комплексе обнаружено 2978 таджикских слов с разной слоговой структурой, причем 1 и 14 представляют собой размеры самой маленькой и самой большой структур соответственно.

Определено статистическое распределение структур, т.е. установлено соответствие между слоговой структурой слов и частотой их встречаемости в текстах таджикского языка.

Установлено, что 17 таджикских текстов покрывают 50%, а 34 - 75%. Обнаружено, что 89 структур охватывают 90% таджикских текстов. При этом 170 структур встретились в тексте 429 843 раза, что составляет 95% охвата текста.

Каждое из 170 сложных слов разбивается на слоги «вручную» (путем деления на слоги слов, соответствующих конкретному сложному слову). Всего определены шесть различных структур слогов — фиксированы 1, 10, 01, 010, 100 и 0100. Частота встречаемости сочетаний, представленных в тексте в 985 768 слогов, возникающих в результате разделения 429 843 слов на 170 различных слогов таджикского языка.

Из полученных результатов следует, что значительно распространены двухбуквенные слоги, как «да», «ба», «ро», «на», «ни», «та», «ме», «ва», «ки» (в символической надписи – 01) и т. д., относительно редки слоги, состоящие из трех букв: «абр», «илм», «ашк», «ишқ», «умр», «орд» (в символическом написании - 100). Следует заметить, что два двухбуквенных слога 10 и 01 вместе с трехбуквенным слогом 010 составляют большинство слогов таджикского языка и что средний размер слогов в таджикском слове составляет 2, 3.

Об алфавитном кодировании слов естественных языков. В ходе исследовательской работы был решен ряд задач, связанных с вопросом алфавитного кодирования слов естественных языков. В частности, было дано определение алфавитного кодирования, согласно которому «слово представляет собой ряд букв, расположенных в алфавитном порядке». Отмечаются особенности прямого и обратного отражения набора слов по отношению к набору образных символов.

Задачи статистического изучения предлагаемому отражению устанавливаются на примере естественных языков.

1. Пусть L – любой естественный язык с алфавитом A ва $W = "\alpha_1\alpha_2 \ldots\alpha_n"$ – любое слово длины n, состоящее из букв $\alpha_k \in A$, $k = 1, 2,\ldots,$ будет n.

Рассмотрим последовательность $CW = "\alpha_{s1}\alpha_{s2} \ldots\alpha_{sn}"$, составленную из тех же букв W, но расположенных в алфавитном порядке.

Определение. Выражение $F: W \rightarrow CW$ называется упорядоченной алфавитной кодировкой (αβ - кодировкой) слова W, а соединение букв CW – αβ-кодом.

Для пояснения определения отметим, что αβ- кодировка, например, слово W = «*муҳаррик*» ведет к соединению CW = «*аикмрруҳ*», слова W = «аббос» - на то же соединение CW = «аббос», потому что в этом слоге буквы уже расположены в алфавитном порядке.

Слово W и его образ CW можно представить в виде двух фиксированных алгебраических сдвигов из множества n! Объясните возможные перестановки n букв, составляющих слово W. Следует отметить, что такая интерпретация понятна, если все буквы разные. Если же буква встречается в слове более одного раза, повторение букв следует различать по порядку букв в слове.

2. Видно, что для каждого слова W выражение F соответствует уникальному изображению CW, но обратное выражение (декодирование), как правило, не уникально. Фактически устранение неоднозначности вызвано анаграммами, которые есть в большинстве языков. Анаграмма – это по крайней мере одна пара слов естественного языка, состоящая из одного и того же набора букв: (например, для таджикского языка: «фарзи» – «зарфи» – «зариф» – «фариз», «рустами» – «мастури», «замин» – «назми», «хумор» – «хурмо» – «рухом» и др.). Каждой анаграмме соответствует один образ, αβ-кодировка.

Вышеизложенное ставит перед исследованием ряд задач.

Задача 1. Необходимо рассчитать относительную частоту встречаемости анаграмм в естественных языках, опираясь на статистические данные.

3. Полученные результаты дают представление об общем потенциале анаграмм и о том, насколько их присутствие нарушает взаимность предлагаемого выражения в естественных языках.

4. Помимо рассмотренного выражения F, интерес представляют четыре его типа, обозначаемые F(*), F(t), F(f) и F(f,l). Как и F, они определяются во множестве {W} естественных слов L.

Придадим выражению F* следующие характеристики:

- F* определяется во множестве {W }, $W \in L$;

- F* : W $\rightarrow$ CW, т.е. отображает слова в их αβ - символах;

- обратное выражение F^{*-1} соответствует набору одиночных символов, кодируемых F^{-1}, в комплексе анаграмм каждому изображению CW соответствует одно слово W*, которое имеет наибольшую частотность в текстах по сравнению с рассматриваемым набором слов-анаграмм.

В случае $F(t)$ паре (CW, NW) соответствует слово W, где CW, как указано в пункте 1, αβ–кода слова W, а NW – количество транспозиций, посредством которых осуществляется переход с канала W на канал CW (или наоборот с канала CW на W).

Здесь предполагается, что n! всех возможных перестановок букв слова W расположены так, что каждая последующая перестановка является производной от предыдущей.

Понятно, что выражение $Ft: W \rightarrow (CW, N)$, осуществляет относительно «успешное» кодирование, чем $CW \rightarrow W$, хотя оно и не является взаимоисключающим.

В случае $F^{(f)}$– выражение слова W связано с цепочкой α $C(W/\alpha)$, где α_1 – первая буква в слове W и $C(W/\alpha_1)$ αβ – символ ссылки W/α_1, т.е. слово W без первой буквы.

Как и в выражении предыдущего пункта, кодирование отмены $\alpha_1 C(W/\alpha_1) \rightarrow W$ имеет в определенном смысле лучшие свойства, чем $CW \rightarrow W$.

Другой способ кодирования выражается как $F(f, l)$ $\alpha_1 C(W/\{\alpha_1,\alpha_n\})\alpha_n$. В нем первые буквы α_1 и последние α_n слова W остаются неизменными, а связь букв между ними, т.е. $W/\{\alpha_1,\alpha_n\}$ подвергается αβ-кодированию. Несомненно,

упомянутое предложение немного сложнее предыдущего, но в плане расшифровки оно более удачно.

5. В заключение следует отметить, что все рассмотренные выше словосочетания произвольно придают слову W единственно подходящий образ. С другой стороны, противоположные им выражения (отмена декодирования) обычно не имеют того же значения. Как упоминалось в пункте 1, F-обнаружение обратного выражения происходит благодаря анаграммам: каждое изображение анаграммы соответствует как минимум двум фоновым изображениям в наборе {W}.

Использование обработанных выражений $F^{(*)}$, $F^{(t)}$, $F^{(f)}$ и $F^{(f,l)}$ - это попытка различить анаграммы и распознать их фоновые изображения с помощью дополнительных символов, закодированных αβ-кодировкой.

Эффективность входных выражений для данного естественного языка можно оценить только экспериментально посредством статистической обработки выборочных данных. Поэтому оно заслуживает внимания.

Задача 2. Исследование статистических свойств выражений F(*), F(t), F(f) и F(f,l), в том числе оценка эффективности декодирования изображений-анаграмм.

Сосредоточение внимания на F-выражениях, таких как F(*), F(t), F(f) и F(f, l), основано на множественных взаимосвязанных представлениях исходного набора, объясненных словами {W}, что значительно упрощает решение задачи ряд проблем с обработкой текстовых данных.

Например, рассмотрите возможность использования оператора F для автоматического исправления ошибки замены букв при вводе слова W. Упомянутый процесс выглядит следующим образом. Сначала слову W присваивается символ CW. Затем фоновое изображение ищется в базе данных «W ↔CW» по коду CW. Если W не является элементом анаграммы, то W является базовым изображением CW.

Поэтому ошибка, связанная со перемещением шрифта (и не обязательно соседнего шрифта), исправляется. Если CW αβ-код анаграммы (например, CW = «*зиорҳ*» – код автора анаграммы - «*ҳозир*», «*зоҳир*», «*изҳор*» выражение), то выбор

основного изображения, например, по самой его высокой частотности (как при выражении F (*), возможен. Но в этом случае может случиться ошибка в принятии решения.

Кодирование предложений. Для упорядочения предложения внутри текста, а также определения анаграмм в составе текста предлагается специальный метод алфавитного кодирования буквенных связей.

Для естественного языка L с буквенным алфавитом A обозначим через $W = "\alpha_1\alpha_2 \ldots\alpha_n"$ любое звено букв длиной n ($\alpha_k \in A,\ k = 1, n$). Рассмотрим связь $CW = "\alpha_{s1}\alpha_{s2} \ldots\alpha_{sn}"$, состоящую из буквы, присутствующей в той же W, но составленный по алфавиту A (например: если W = “малоик”, то CW = “аиклмо”).

Определение 1. Выражение $F: W \rightarrow CW$ называется упорядоченной буквенной кодировкой ($\alpha\beta$ -) соединения W, в свою очередь CW называется $\alpha\beta$-кодом W соединения.

Выражение F и ряд «смежных» выражений предложены для автоматизации процесса кодирования словоформ и поиска анаграмм в текстовых предложениях. Статистический анализ позволил выявить эффективность αβ-кодирования (в смысле возможности совпадения значений между словоформами и αβ-кодами) для английского, литовского, русского и таджикского языков, а также язык – эсперанто. Также собрана статистика по количеству различных анаграмм определенного размера (количества элементов) для текстовых ресурсов (корпуса) таджикского, английского и русского языков. Были найдены и представлены многочисленные анаграммы с большим количеством элементов.

Результаты были получены в рамках обработки данных двумя методами, а именно:

- созданием списка различных словоформ с указанием частоты их появления в базе данных;

- кодированием производных словоформы и созданием списка различных символов с указанием частоты их появления.

Словоформа является частным случаем понятия буквосочетания, но это не мешает описанной процедуре автоматически распространяться на множество {W}, элементами которого являются W-соединения. Результатом первой процедуры является список соединений с их частотами в множестве {W}, а результатом второй процедуры – список $\alpha\beta$-кодов множества соединений.

Эта процедура, формальная по отношению к абстрактным привязкам, рассматривается в следующем параграфе.

Кодирование предложений. Пусть S – любое предложение языка L, состоящее из p слов. Удаляя из S все знаки препинания и пробелы между словоупотреблениями, получаем связь W(S) из буквы предложения S.

Определение 2. Связь CW(S), образованная выражением F: W(S) → CW(S), называется αβ -символом предложения S.

Положения. Выражение F и ряд соседних выражений предложены для автоматизации процесса кодирования словоформ и поиска анаграмм в текстовых предложениях. Статистический анализ позволил выявить эффективность $\alpha\beta$ -кодирования (в смысле возможности совпадения значений между словоформами и αβ-кодами) для английского, литовского, русского и таджикского языков, а также язык – эсперанто. Кроме того, собрана статистика по количеству различных анаграмм определенного размера (количества элементов) для текстовых ресурсов (корпуса) таджикского, английского и русского языков. Были найдены и представлены многочисленные анаграммы с большим количеством элементов.

Эта процедура, формальная по отношению к абстрактным привязкам, рассматривается в следующем разделе.

Пусть S_0 и S — два разных предложения с $CW(S_0) = CW(S_1)$. Тогда S_0 и S_1 элементы одной и той же анаграммы, то есть одно получается из другого перемещением соответствующей буквы.

Пример 1. Пусть S_0 – «абри монда», а S_1 – «бари доман». Поскольку оба предложения имеют одинаковый αβ-код, т.е. $CW(S_0) = CW(S_1)$ = «аабдимнор», то по указанным выше правилам рассматриваемые предложения являются

элементами одной и той же анаграммы, так как состоят из одного и того же набора букв.

Эти примеры показывают, что когда S_0 и S_1 берутся из набора данных, сравнение их αβ-кодов позволяет нам ответить на вопрос, являются ли они анаграммами друг друга.

Задача А. По заданному S_0 определить, содержит ли данное текстовое предложение такой S1 или нет, или $CW(S_0) = CW(S_1)$.

Очевидно, что диапазоны S_0 и S_1 рассматриваются как множества, элементами которых являются предложения и их части. Если задача имеет решение ровно для S_0, то естественно, что S_1 будет получена точно. Сложность решения этой задачи, вероятно, будет заключаться в разработке метода систематического анализа заданного набора текстов всех кандидатов на роль S_1.

Индексирование элементов. αβ-кодирование можно использовать (без конкретных целей) для формального упорядочивания словоформ внутри предложений, а также предложений внутри текста. Фактически из двух элементов, будь то слово или предложение, мы считаем элемент с наименьшим количеством букв, а если они одинаковы, считаем элемент в начале алфавита в L. Единицы перевода (формы слов и предложений) и определить соответствующий порядок элементов.

Исследована проблема моделирования лингвистических задач на основе математического аппарата на примере английского и русского языков.

Метод математического моделирования был использован для исследования вопросов обработки текстовой информации на естественном языке.

Были созданы общие способы структурирования и кодирования текстовых элементов, такие как слоговая структура слова, кодирование слов и предложений.

Математические методы З.Д. Усманова использованы для исследования проблем обработки текстовой информации на таджикском языке.

Также с помощью математических методов обработки информации были исследованы статистические закономерности некоторых элементов текста: слогов, слов, анаграмм, предложений.

Слоговые структуры слов таджикского языка были созданы с целью синтеза речи. Определены математические методы кодирования текстовых элементов для решения задач автоматической проверки правописания и машинного перевода текста на таджикский язык.

§2.3. Методы обработки системы проверки правописания текстовых данных

Автоматическая проверка орфографии является одним из основных требований шифрования естественного языка. Система проверки орфографии анализирует весь объем введенных слов и при наличии орфографических ошибок предлагает конкретный список наиболее подходящих вариантов. Возможные ошибки могут возникнуть при использовании клавиатуры или при несоблюдении правил написания текста на естественном языке.

Проблема автоматического определения правописания слов является важной исследовательской задачей в области компьютерной лингвистики. Этому способствует развитие информационных технологий и теории алгоритмов, а также возможность оцифровки текстовой информации. Имеется ряд причин для продолжения научно обоснованных усилий, направленных на развитие этой области – разработка автоматических информационных систем, внедрение программ проверки правописания текста [194; 196; 205; 217; 222; 236; 257].

Исследования в области проверки орфографии начались еще в 60-х годах прошлого века. Особенно ярко оно нашло свое выражение в работе Дамерау [132], где были предложены методы предсказания слов с ошибками в документах. Проверка орфографии очень важна для ряда компьютерных программ, таких как текстовые процессоры, веб-браузеры, поисковые системы и т.д. В процессе исследования установлено два метода проверки орфографии: обнаружение ошибок и исправление ошибок.

В начальном периоде изучались математические модели, алгоритмы и методы разработки систем проверки орфографии и автоматического исправления ошибок текста на естественном языке.

Разработка и реализация функций компьютера обеспечивают редактирование текста, доступ к поисковым системам с возможностью исправления орфографии. В большинстве случаев орфографические ошибки не важны для пользователей, но для автоматизированных систем орфографические ошибки могут стать проблемой, ограничивающей их эффективность.

Наиболее распространенный способ поиска ошибок в тексте – поиск каждого слова в словаре: отсутствие слов в словаре считается ошибкой. Но прежде, чем перейти к описанию этого вида метода, рассмотрим несколько методов без использования словаря.

Первый метод, предложенный в 1974 году Райсманом [62] и Хенсоном [63], использует словари второстепенным способом. Этот метод начинается с проверки словаря и сортировки всех трехбуквенных последовательностей – триграмм, встречающихся при поиске. Например, в таджикском языке, триграмма «*анд*» встречается очень часто, а «*рпт*» вообще не встречается. Используя список триграмм с наибольшей частотой встречаемости, система проверки правописания делит текст на триграммы и ищет их в списке. Если встречается триграмма, которой никогда не было в словаре, то слово, содержащее такую триграмму, фактически должно быть написано с ошибкой. Например, если система обнаружит «*зиёрпт*», которое может быть ошибочно написано вместо «*зиёрат*». Для системы обнаружения орфографических ошибок этот метод имеет ограниченную ценность, поскольку большинство ошибок не содержат невозможных триграмм. Этот метод может быть эффективно использован для задач обнаружения ошибок в процессе автоматического чтения текста с распечатанных страниц сканером, где каждая буква «читается» отдельно согласно методу.

Второй метод, предложенный Моррисом и Черри в 1975 году, вообще не использует словарь. Как и предыдущий метод, он делит текст на триграммы, но составляет их список и отмечает, как часто каждый из них встречается в одном и

том же фрагменте текста. Затем он снова считывает текст с помощью сканера и вычисляет уникальный индекс каждого слова на основе содержащихся в нем триграмм. На примере слова «*зиёрат*» можно выявить, что это единственное слово в тексте, имеющее «*зиё*», «*иёр*», «*ёра*», «*рат*». В этом случае, если в списке появится триграмма «рпт», система может ошибочно оценить слово по слову «*зиёрпт*». А вот в случае введенного слова «*корманд*», состоящего из триграмма «кор», «орм», «рма», «ман» и «анд», наоборот, оно получает низкую оценку, поскольку в результате полученные триграммы, вероятно, в анализируемом фрагменте текста встречаются где-то еще. После завершения анализа метод обращает внимание пользователя на слова с высоким показателем.

Второй метод, как и первый, нельзя использовать для поиска наиболее частых орфографических ошибок. Второй метод также эффективен для обнаружения опечаток, поскольку метод был разработан именно для этой цели. Его преимуществом перед всеми методами, основанными на словарях, является поиск орфографических ошибок в многоязычных текстах, например, в русских или английских словах, имеющихся в таджикском языке.

Для проверки орфографии, то есть поиска и исправления ошибок, широко и эффективно используется заранее определенный словарь. Существует два распространенных подхода к обнаружению ошибок: поиск по словарю и анализ *n-граммы*. Большинство методов проверки правописания используют словари в качестве списков правильного написания, помогая алгоритмам находить правильно написанные слова. Так, один из методов, предложенных Геханом Далкиличем и Ялчином Чеби, основан на анализе *n-грамм*. Этот метод ищет слова с ошибками по всему тексту. На первом этапе необходимо создать список *n-грамм* на основе ресурса (корпуса). В этом случае ресурс должен быть достаточно большим, чтобы найти все возможные варианты *n-граммы*. На втором этапе методом долгосрочной стабилизации рассчитывается частота появления неизвестной *n-граммы* в ресурсе (корпусе). Если обнаруживается недостающая *n-грамма*, слово определяется как написанное с ошибкой. Вышеупомянутые методы легли в основу большинства алгоритмических методов поиска и исправления орфографических ошибок в

тексте. Подходы к исправлению ошибок с использованием *n-грамм* как элемента обработки текста были впервые предложены исследователем Шенноном [77] в 1994 году. Идея использования *n-грамм* была применена ко многим задачам, таким как распознавание речи, переведенное слово, коррекция слов, предсказание и исправление правописания. Этот чисто статистический метод не требует знания языка документа. Еще одним преимуществом *n-грамм* является автоматическое покрытие часто встречающихся корней. N-граммы можно использовать двумя способами: без словаря или со словарем.

Использование *n-граммов* без словаря позволяет узнать, в какой позиции неправильного слова произошла ошибка. Этот метод позволяет заменить неправильное слово без использования словаря так, чтобы оно содержало только правильные n-граммы, вместе с тем этот метод отличается низкой производительностью.

С другой стороны, n-грамма используется вместе со словарем для определения расстояния между словами, но слова всегда сравниваются со словарем. Есть много способов сделать это, например, проанализировать, сколько n-грамм общего имеет слово с ошибкой и словарное слово по длине.

Сравнительно распространенный и эффективный способ найти и исправить слова с ошибками – использовать расстояние Левенштейна [132] для количественной оценки сходства с признанным стандартным словарем. Хотя существует множество связанных мер расстояния и их применений, расстояние Левенштейна по-прежнему остается относительно надежным способом использования операции редактирования для сравнения двух строк. Но это односторонне, поскольку берется во внимание только расчет расстояния редактирования между словами с ошибками и правильно написанными словами.

Блок-схема алгоритма проверки орфографии на основе n-грамм показана на рисунке 2.2.

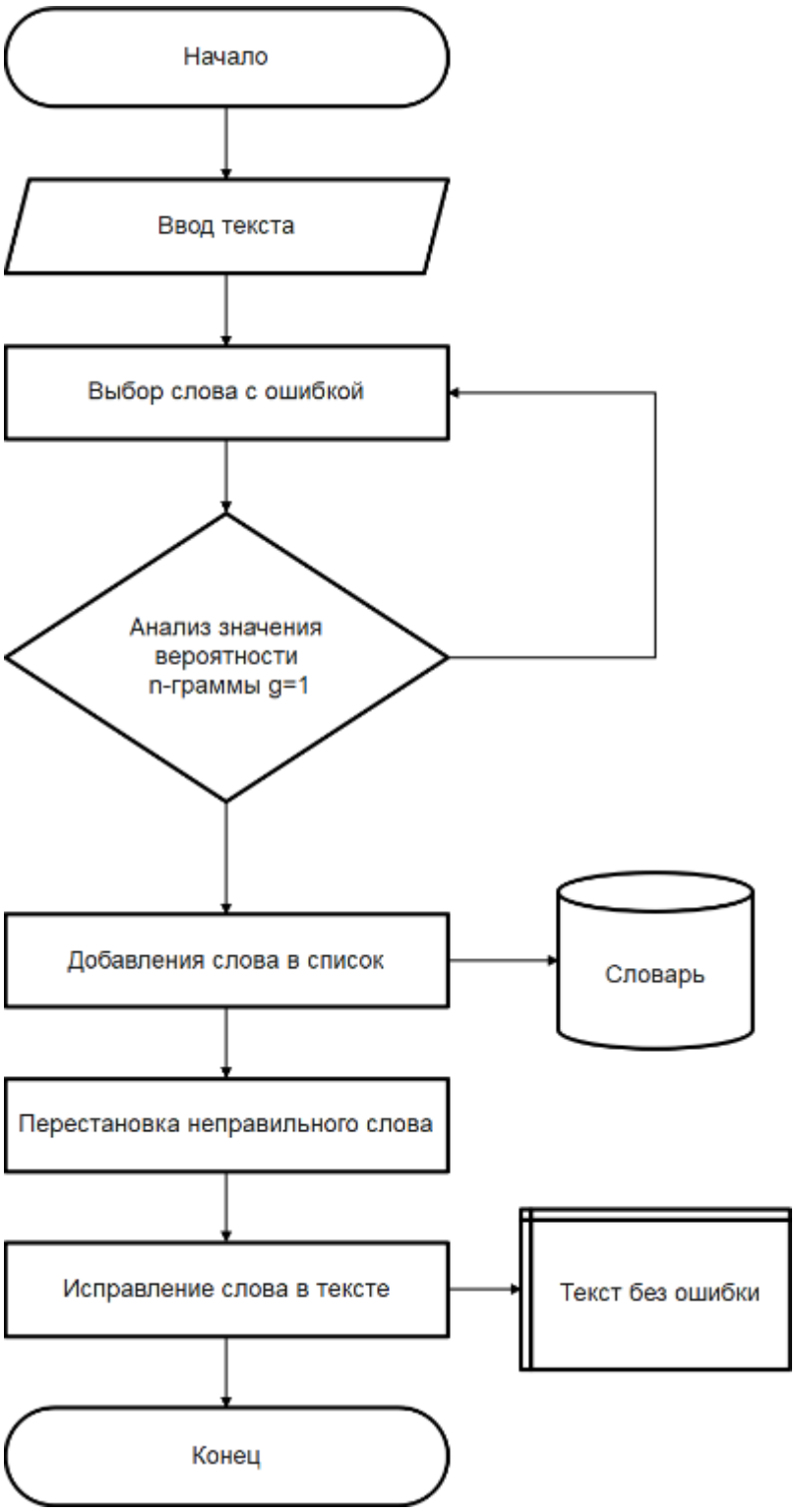

Рисунок 2.2. - Схема алгоритма проверки орфографии на основе n-грамм

Лучший метод обнаружения ошибок на основе словаря был предложен в 2009 году ученым Де Аморимом [66]. Этот метод требует формирования словаря из лексических источников, содержащего список правильных слов данного языка. Однако и этот метод имеет ограниченную продуктивность из-за внутренней

архитектуры ресурсов и использования словарей, организованных как автоматические машины окончательной обработки.

Согласно проведенным исследованиям, существует два типа орфографических ошибок: когнитивные и печатные, т.е. типографические ошибки.

Когнитивные ошибки – это ошибки, которые возникают, когда неизвестно правильное написание слова. В этом виде произношение слова с ошибкой такое же или похоже на произношение предположительно правильного слова, например «*кумак*» вместо «*кӯмак*», где опечатки составляют около 80% ошибок. При анализе опечаток можно выделить четыре относительно распространенные группы. Например, для слова «*истиқлол*» возможен один из следующих вариантов:

1. Ввод одной буквы: «*исстиқлол*» (ошибка x1).
2. Написание одной буквы: ««*итиқлол*»» (ошибка x2).
3. Изменение одной буквы: «*истиклол*» (ошибка x3).
4. Смещение двух соседних букв: «*итсиқлол*» (ошибка x4).

Алгоритм Левенштейна. Для выполнения задачи по исправлению первых трех типов ошибок в слове при вводе текста широко используется расстояние Левенштейна [30]. 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:

$$D(i,j) = f(x) = \begin{cases} 0, \ i = 0, j = 0 \\ i, \ j = 0, i > 0 \\ j, \ i = 0, j > 0 \\ min\begin{Bmatrix} D(i,j-1)+1, \\ D(i-1,j)+1, \\ D(i-1,j-1)+m(w1[i],w2[j]) \end{Bmatrix}, \ J > 0, i > 0 \end{cases} \quad (2.3)$$

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).

Для определения ошибки четвертого типа, т.е. смещения двух соседних букв в одном слове, предлагается использовать метод расстояния Дамерау-Левенштейна, который является формой расширения алгоритма расстояния Левенштейна.

$$D(i,j) = \begin{cases} \max(i,j), \ \min(i,j) = 0 \\ \min \begin{cases} D(i-1,j)+1 \\ D(i,j-1)+1 \\ D(i-1,j-1)+1_x \\ D(i-2,j-2)+1 \end{cases}, \ i,j > 1, w1[i] = w2[j-1], w1[i-1] = w2[j] \\ \min \begin{cases} D(i-1,j)+1 \\ D(i,j-1)+1 \\ D(i-1,j-1)+1_x \end{cases} \end{cases} \quad (2.4)$$

здесь x= $w1[i] \neq w2[j]$ принимается во внимание.

Расмотренный выше метод предполагает наличие базы данных с миллионами слов. Необходимо найти правильный вариант неправильного слова, которое предлагается ввести. Поиск совпадения путем сравнения искомого слова с каждым словом в базе данных для определения сходства требует длительной операции.

Вместо того, чтобы повторять совпадение каждого слова в базе данных, можно заранее подготовить набор возможных вариантов с помощью алгоритма хеширования и выполнить поиск совпадения в хешированной базе данных.

Однако исследования показали, что большинство инструментов проверки правописания так или иначе используют тот или иной способ охвата максимального словарного запаса конкретного языка. Такие словари составлены для некоторых языков. Одним из недостатков использования словаря является

необходимость наличия большого объема постоянной памяти компьютера. Но есть способ экономии места, согласно которому сохраняется только база слов. В случае с таджикским языком данная задача также признана подходящей. Например, вместо того, чтобы писать в словаре слова «*корманд*», «*коргоҳ*», «*корӣ*», «*корро*» и «*корҳо*», мы можем хранить только слово «*кор*», учитывая набор правил удаления аффиксов перед поиском слов. Удаление аффиксов продолжается до тех пор, пока не будет получена основа, а она обязательно появится. Этот процесс известен как удаление аффиксов.

Приведенные исследования легли в основу разработки методов, математических моделей и их применения в виде алгоритмов решения задач проверки правописания текстовой информации на многих языках, таких как: английский, русский, казахский, узбекский. Они позволили разработать и внедрить систему автоматической проверки правописания на таджикском языке, которая подробно рассмотрена в четвертой главе диссертационной работы.

§2.4. Алгоритмы и методы применения машинного перевода

Понятие машинный перевод можно отнести к информационным системам, которые напрямую связаны с компьютерными технологиями, обеспечивающими перевод текста с участием человека или без него. Автоматический перевод, являясь относительно новым средством компьютерного перевода, поддерживает переводчиков, предоставляя доступ к справочникам, различного вида словарям, удаленным базам терминологических ресурсов, способствует передаче и приему текста, являясь хранилищам ранее переведенных текстов в виде «памяти переводов» [203; 231-232; 239; 248; 356].

В ходе исследования общей структуры основных методов машинного перевода было установлено следующее:

Первый метод обычно называют прямым переводом или «вторым (бинарным) переводом», который разработан со всеми возможностями и специально для конкретной пары языков. Перевод выполняется непосредственно

из начального текста в целевой текст с минимальным синтаксическим или семантическим анализом.

Основные недостатки таких систем заключаются в следующем:

- синтаксический анализ лексики и текстов основного языка для определения порядка слов в переводимом языке не реализуется;

- для правильного распознавания фраз на языке перевода не определяются все возможные варианты многозначности текстов на языке исходного текста;

- система нацелена только на создание представления текста с исходного языка, подходящего для целевого языка.

Обычно такие системы состоят из большого двуязычного словаря, программных модулей анализа и генерации текстов. Такие системы иначе можно назвать системами «прямого перевода» только для двух языков в одном направлении.

Вторая методология опирается на стратегию проектирования *интерлингвистического подхода*, который ориентирован на внутреннюю структуру основного языка и предлагает возможность преобразования текстов в форме семантико-синтаксического представления. В рамках конверсионного подхода рассматривается перевод более чем на один язык. На основе такого межъязыкового представления создаются тексты для других языков.

В этих системах перевод осуществляется в два этапа. На первом этапе перевод с исходного языка осуществляется на основе межъязыкового подхода. На втором этапе – с интерлингвистического подхода к языку перевода. В то же время межъязыковой подход требует анализа текста на исходном языке только для конкретного языка, подлежащего переводу.

Распространенной причиной поддержки межъязыкового подхода является экономия усилий в многоязычной среде, т.е. программа анализа одного исходного языка может использоваться для более чем одного языка при переводе, а программа генерации может многократно использоваться для конкретного языка. Для реализации интерлингвистического подхода можно использовать язык эсперанто,

который искусственно создан на основе набора общих для всех языков лексики и правил, что делает его относительно подходящим языком.

Третий основной метод – это трансферный подход. Вместо двух этапов межъязыкового представления есть три этапа, которые направлены как на тексты исходного языка, так и на тексты языка перевода. На первом этапе тексты с языка оригинала преобразуются в абстрактные представления, ориентированные на свои правила. Второй этап преобразует полученное представление в эквивалентные представления, ориентированные на язык перевода. На третьем этапе создаются окончательные тексты с исходного языка на переводимый язык.

Если межъязыковой подход требует полного разрешения всех неясностей в тексте исходного языка, чтобы был возможен перевод на любой другой желаемый язык, то трансферный подход разрешает только те неясности и двусмысленности, присущие самому языку: омонимы и многозначные сложные синтаксические структуры. На втором этапе решаются проблемы лексических различий между языками.

Для реализации таких систем перевода необходима разработка следующих типов словарей:

- словари на языке оригинала, содержащие подробные грамматические и семантические правила;

- аналогичные словари на переводимом языке;

- двуязычный «переводной» словарь, в котором даются основные формы как языка оригинала, так и языка перевода.

Методом с хорошей перспективой в решении проблемы машинного перевода является так называемый нейронный машинный перевод. Этот подход потенциально может преодолеть большинство недостатков предыдущих систем машинного перевода: метода «согласно правилу» и автоматического статистического перевода. Этот подход, основанный на статистике, моделирует работу нейронных сетей в человеческом мозгу и определяет новые способы компьютерной обработки естественного языка посредством нейронных сетей.

Модели нейронного машинного перевода в основном основаны на словарях с заранее определенными и записанными словами, хотя перевод в основном опирается на открытый словарь.

Каждый из рассмотренных методов способствовал созданию ряда моделей, методов и математических алгоритмов машинного перевода. Наиболее популярными на сегодняшний день являются следующие методы:

1. Метод машинного перевода, основанный на правилах. Основная идея данного метода выражается в повторном использовании ранее существовавших образцов перевода в качестве основы нового перевода. Описанный метод сравнивает входные данные с базой данных реальных образцов и определяет наиболее близкое совпадение. Процесс автоматического создания совпадений перевода происходит путем сопоставления соответствующих фрагментов перевода и последующего их итеративного объединения для создания текста перевода.

Алгоритм машинного перевода на основе правил. Первые подходы к машинному переводу основывались на лингвистических правилах, которые использовались для анализа исходного предложения и создания промежуточного представления целевого языка. Такие подходы полезны для перевода между языками из близкородственных языковых семей. Методы машинного перевода, основанные на правилах, включают машинный перевод на основе словаря, машинный перевод на основе преобразователей и межъязыковый перевод (рис. 2.3).

Инструмент на основе словаря использует записи языкового словаря для поиска эквивалентных слов на целевом языке. Использование словаря в качестве единственного источника информации для перевода означает, что слова переводятся точно так, как они указаны в словаре. Поскольку во многих случаях это не так, применяются грамматические правила.

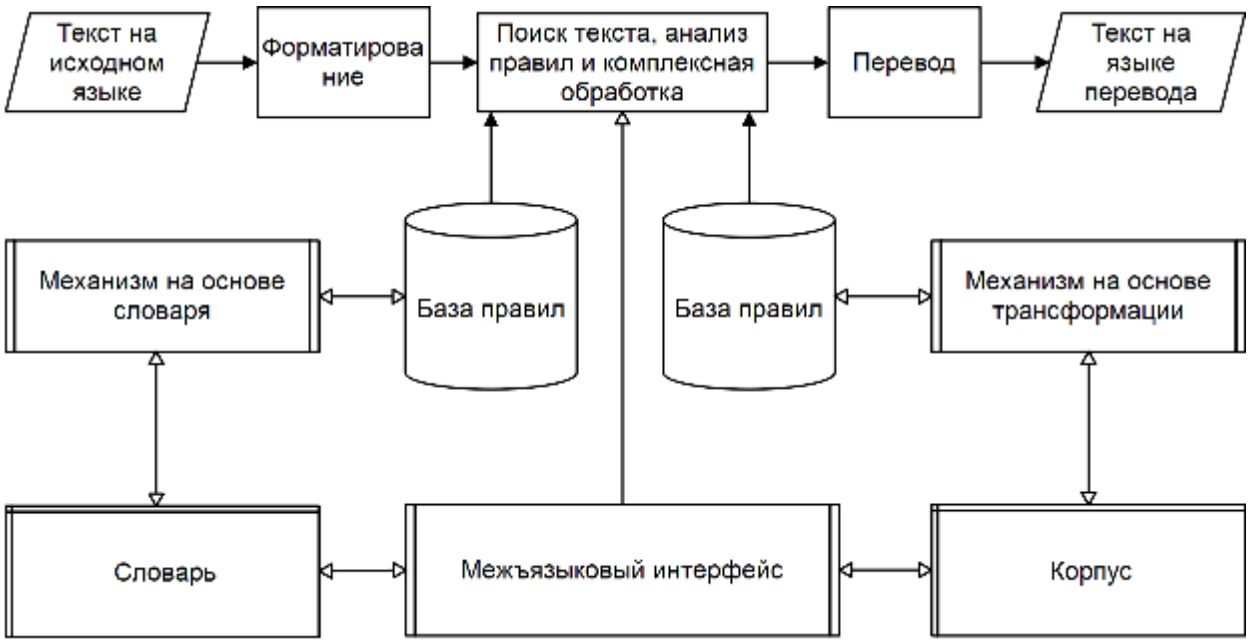

Рисунок 2.3. - Алгоритм машинного перевода на основе правил

Механизм, основанный на методе трансфера, преобразует исходный текст в промежуточный текст с помощью специальной структуры языковых правил. Данная структура затем переносится в аналогичную структуру на целевом языке и создается на переведенном языке. Механизм использует заранее определенный и полный корпус, основанный на гравиметрических, морфологических, синтаксических и семантических правилах как исходного, так и переводящего языков.

В межъязыковом интерфейсе текст преобразуется в промежуточный язык, так называемый искусственный язык. Это нейтральное представление независимо от любого языка. Упомянутый механизм подходит при задействовании более двух языков, поскольку не требует, чтобы каждый из языков был связан с набором правил перевода друг с другом в обоих направлениях с возможностью реконструкции предложений и словосочетаний. Чтобы добавить новый язык, необходимо указать анализатор текста и генератор вариантов для преобразования в промежуточный язык. При этом полученные варианты проверяются лингвистами, которые оценивают сбалансированность машинного перевода на основе правил.

2. Статистический метод машинного перевода, основанный строго на методе статистического перевода, не использует традиционные лингвистические

правила. Суть этого метода заключается в том, что сначала сопоставляются фразы, группы слов и отдельные слова параллельных текстов, затем рассчитывается вероятность совпадения слова с каждым словом одного языка или словами переведенного предложения, которым они соответствуют на другом языке. Особенностью статистического метода является формирование большого двуязычного корпуса переводов.

Алгоритм статистического машинного перевода в основном использует две возможные модели: модель перевода и языковую модель. В первой модели изучаются двуязычные корпуса и определяется оценка вероятности правильного перевода текста с языка оригинала на целевой язык. Языковая модель изучается на одноязычном корпусе и используется для улучшения результатов перевода (рис. 2.4.).

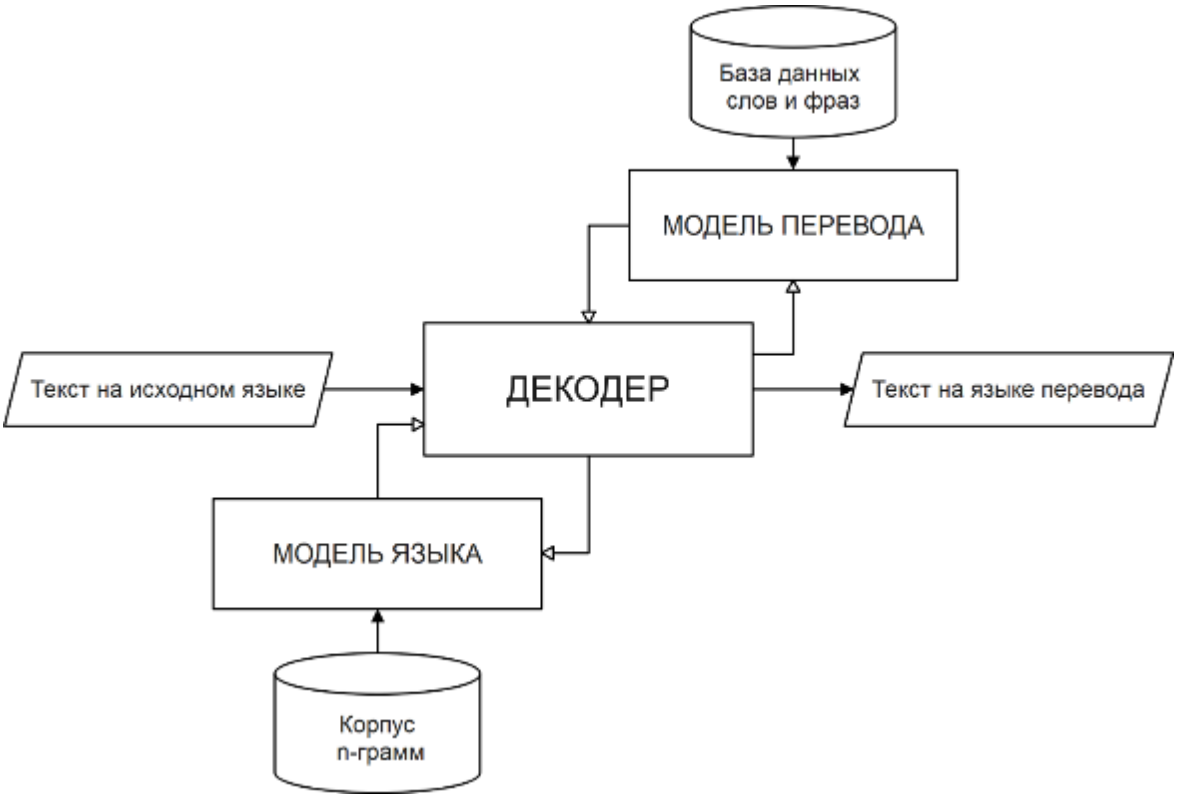

Рисунок 2.4. - Алгоритм машинного статистического перевода

Основная проблема реализации алгоритма статистического перевода – машинное обучение для изучения шаблонов перевода на основе образцов

человеческого перевода и обучающих корпусов. В этом случае также необходимо определить наибольшую вероятность использования соответствующего текста как пропорционального результата перевода с языка оригинала. 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:

$$t_{\text{беҳт}} = argmax_t\, P(t|s) = argmax_t\, P(s|t) \times P(t) \qquad (2.5)$$

где,

P(s|t) – модель перевода,

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.

Модели перевода в основе слов. 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:

$$P(t,a|s) = \prod_{i=1}^{l_t} P(t_i|s_{ai})P(a_i|a_{i-1},i,l_t,l_s) \quad (2.6)$$

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

Модель перевода на основе словосочетаний. 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:

$$P(s|t) = \prod_{i=1}^{I} \emptyset(s_i|t_i)d(a_i - b_{i-1} - 1) \quad (2.7)$$

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* отметим 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.

$$P(w_1, \dots, w_m) = \prod_{i=1}^{m} P(w_i | w_1, \dots, w_{i-1}) \quad (2.8)$$

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

$$P(w_1, \dots, w_m) \approx \prod_{i=1}^{m} P\left(w_i \middle| w_{i-(n-1)}, \dots, w_{i-1}\right) \quad (2.9)$$

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:

$$P\left(w_i \middle| w_{i-(n-1)}, \dots, w_{i-1}\right) = \frac{count(w_{i-(n-1)}, \dots, w_{i-1}, w_i))}{count\left(w_{i-(n-1)}, \dots, w_{i-1}\right)} \quad (2.10)$$

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

Механизм декодирования. Основная цель декодера – определение лучшего варианта текста на языке перевода, который максимизирует вероятность перевода P(t|s). Эвристические методы можно признать наиболее эффективным способом декодирования. При применении статистического алгоритма машинного перевода процесс пословного декодирования сравнительно сложнее, так как существует

возможность перестановки отдельных слов в исходном языковом тексте. В то же время процесс декодирования может быть реализован с использованием алгоритмов пропорционального поиска, программирования целых чисел или алгоритмов поиска с конечным результатом.

В зависимости от сложности использования перевода каждого слова процесс декодирования с использованием модели фразового перевода является относительно успешным, поскольку в нем используются относительно более крупные единицы перевода: словосочетания, структуры, короткие предложения. Наиболее часто используемыми алгоритмами декодирования являются декодирование стека поиска вектора, поиск вектора на основе интегрируемых стеков, декодирование с ограниченным размахом подъема и декодирование с преобразованием конечного состояния. При декодировании с поиском вектора декодер начинает с поиска всех возможных переводов в таблице фраз со всеми вероятными типами фраз в тексте на языке оригинала. Декодирование исходного текста начинается с пустого исходного предположения, затем слева направо создаются предполагаемые варианты перевода.

В результате все предположения расширяются за счет выбора доступных вариантов перевода.

3. Бинарные методы машинного перевода. В определенных случаях каждый из двух предложенных методов показал определенную эффективность. В результате возник ряд бинарных и гибридных систем. Метод, основанный на шаблонах, подходит для решения конкретных задач, связанных с определением всех возможных правил языка, подлежащего переводу. Элементы аналитического раздела могут больше опирается на статистический анализ, тогда как передача и генерация больше подходят для подхода, основанного на правилах.

Иными словами, бинарный метод считается «*многомоторной*» системой. В этом случае исходный текст проходит через ряд различных систем машинного перевода, каждая из которых использует разные методы. Один может быть основан по сути на словарном запасе, другой – на анализе и образовании правил, третий – на основе образцов или чисто статистических данных.

4. Метод машинного перевода, основанный на машинном обучении, в основном включает в себя модуль предварительной обработки, модуль кодирования, декодирования и модуль внимания. Последовательность каждого из упомянутых модулей согласно системе включает в себя модуль предварительной обработки, модуль кодирования, модуль внимания и модуль декодирования. Для реализации процесса интеграции нейрологического машинного перевода все четыре упомянутых модуля работают вместе в соответствии с алгоритмом, показанным на рис.2.5.

Модуль предварительной обработки основан на методе извлечения текста на естественном языке, синтаксическом анализаторе структуры фраз, извлечении логической структуры и, наконец, генерации структуры текста на языке перевода. Метод многопоточного кодирования преобразует текст исходного языка в набор фиксированных элементов, которые анализируются непосредственно в соответствии с правилами синтаксиса языка из заданного корпуса.

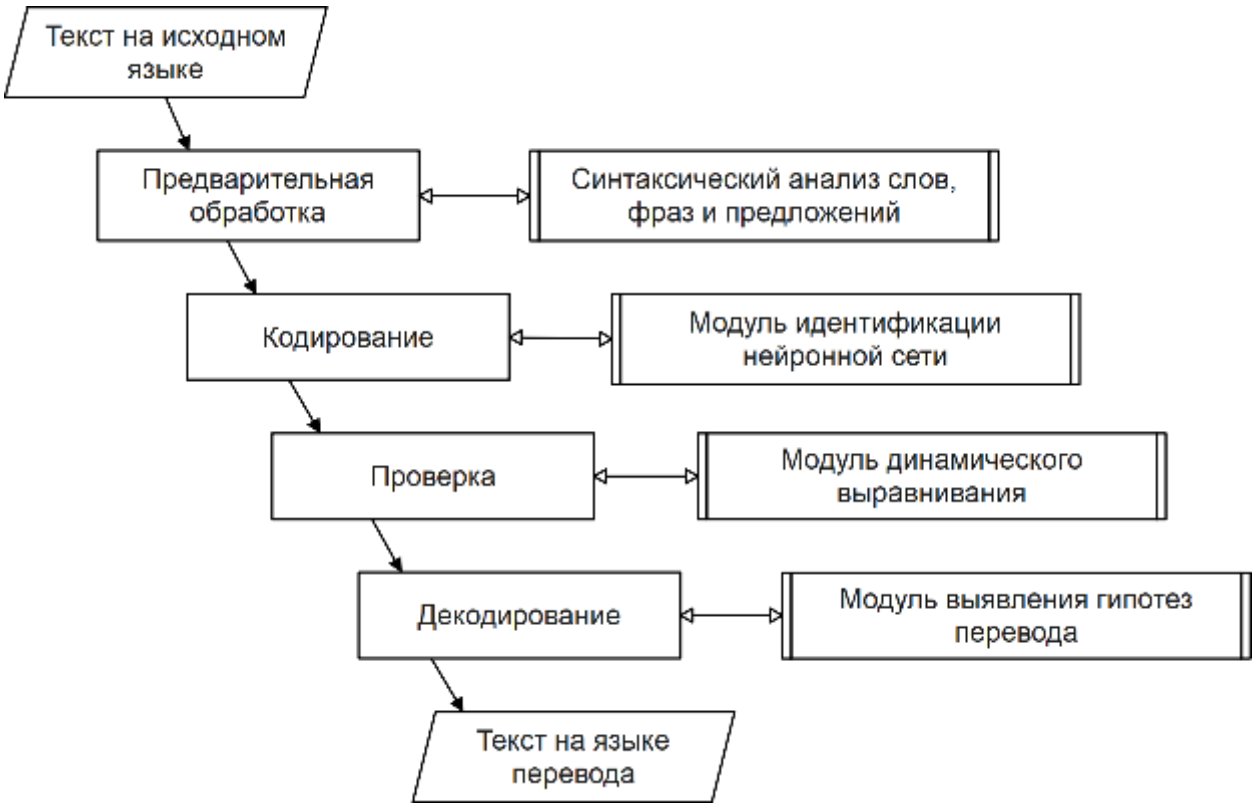

Рисунок 2.5. - Алгоритм машинного перевода на основе машинного обучения

Модуль шифрования является частью структуры шифрования-дешифрования интегрированного машинного перевода. Двунаправленная нейронная сеть используется для преобразования последовательности идентификаторов в текст исходного языка в соответствии с непрерывным векторным представлением больших размерностей.

Тестовый модуль в основном реализует динамическое и выборочное внимание к каждой части текста на языке оригинала при генерации слов языка перевода для получения точного содержательного вектора в качестве основы для перевода.

Модуль декодирования соответствует модулю кодирования в обратном порядке. Целью указанного модуля является отражение семантического представления векторной формы в тексте на естественном языке. Для решения задачи перевода больших объемов текстовой информации целесообразнее использовать линейные неравенства, например статистический машинный перевод.

В результате вместо последовательной сложной обработки скрытых структур можно получить единую сложную нейронную сеть. Этот метод значительно упрощает процесс перевода и решает проблему распространения ошибок в процессе совместного действия многих модулей статистического машинного перевода. Используя метод перестановки рекурсивной нейронной сети, можно генерировать бесконечную длину из содержимого окна определенного размера.

Результаты исследования основаны на разработке и внедрении системы автоматического перевода текстов на таджикский язык на базе бинарной модели машинного перевода, которые более подробно описаны в пятой главе диссертации.

Достижению этой цели способствуют следующие основные задачи:

- разработать модели, методы и математические алгоритмы на основе методологии бинарного машинного перевода таджикского языка;

- создать логичную и реальную параллельную структуру ресурсов, в первую очередь «русско-таджикскую» и «англо-таджикскую» для информационного обеспечения системы машинного перевода;

- определить эффективные алгоритмы поиска, выделения и сортировки текстовых элементов на таджикском языке и пути их реализация в программных модулях для параллельной обработки ресурсов.

§2.5. Методы и алгоритмы текстового синтеза речи

Разработка системы текстового синтеза речи на основе текста предполагает вычисление речевого сигнала по заданному тексту. Система состоит из большого набора компонентов, таких как абстрактно-лингвистический анализ структур текстовых элементов и набора методов кодирования речи. Разработка дизайна и реализация таких систем является одной из наиболее актуальных задач компьютерной лингвистики.

Следует отметить, что синтез речи представляет собой сочетание нескольких отраслей, в том числе инженерии, лингвистики, информатики, математики и управления звуком. Современный научный прогресс в указанных областях используется для решения проблемы синтеза речи на неравномерном уровне. Проблемы, с которыми сталкивается каждая из этих упомянутых областей, принципиально различны по своей сложности. Создать цифровой звук из больших объемов текста чрезвычайно сложно. Однако задача разделения отрезка звука на кратчайшие по длительности компоненты, соответствующего определенным значениям текстовых элементов, не представляет большой сложности [199-200; 202; 219; 226; 237; 253; 258].

Одной из основных задач диссертационного исследования является изучение возможностей указанных направлений для теоретической постановки проблемы и ее решения, а также их практическое применение для создания системы автоматического синтеза речи на таджикском языке.

Для достижения указанной цели необходимо изучить природу человеческого голоса и механизм его формирования, описать цифровой образ звука на основе математических моделей, а также оценить существующие методы и алгоритмы синтеза речи заданного текста. Для создания концептуальной модели синтеза речи

необходимо исследовать практические возможности и недостатки современных синтезаторов.

Речь – это уникальный звук, производимый голосовым аппаратом человека. Воздух выталкивается из легких через голосовой тракт, создавая акустические волны, которые отлетают с губ как поля давления. Особенности этого процесса хорошо изучены, что дает нам важные сведения о речевом общении. Поперечное сечение речевого тракта человека состоит из площади поперечного сечения ротовой полости от отверстия глотки до губ и определяется следующими частями: высотой тела языка; переднее/заднее положение языка; высота кончика языка; ротовая полость; отверстие гортани и акустические импеданцы на губах. Модель речевого аппарата состоит из трех компонентов: полости рта, полости глотки и акустического сопротивления губ.

Одним из важнейших открытий в области компьютерной лингвистики является то, что человеческую речь можно моделировать на основе импульсной функции источника. Легкие направляют поток воздуха через голосовые связки, а передаточная функция улавливает воздействие шума речевого тракта, преобразуя его в речевые спектрограммы. Соответственно, цель синтеза речи и цифровой передачи речи состоит в том, чтобы представить источник голоса как одиночный импульс за определенный период времени.

В случае цифрового кодирования речи она распадается, а затем повторно синтезируется по отдельным фрагментам. На последнем этапе происходит фильтрация, в результате которой большая часть голосовых единиц скрывается. Но в то же время спектральная модель представляет собой лишь реальную анализируемую речь, не соответствующую просодическим измерениям речи. Определение цифрового образа человеческой речи основано на методах, разделяющих запись голоса на фонемы. Затем анализируются амплитудные и частотные характеристики каждой фонемы с целью поиска фонем отдельных букв на основе их классификации по определенному набору частотных характеристик. Такие методы рассматривают каждую фонему как единую неразложимую единицу речевого сигнала с частотными характеристиками. Подходы к анализу

аналитического описания фонем и построению математической модели реально могут быть использованы как для задач распознавания, так и для задач синтеза звуков речи.

Цифровое изображение речи. Известно, что слово состоит из одного или нескольких слогов, которые, в свою очередь, состоят из одной или нескольких фонем. Фонема – мельчайшая единица языка, не имеющая ни лексического, ни грамматического значения. Чтобы понять первичные единицы текста, например слово в речи, необходимо проанализировать расположенные последовательно фонемы.

Описание амплитуды и времени произношения фонемы буквы «О» в трех разных временных интервалах: начало речи; описание фонемы; окончание речи (рис. 2.6).

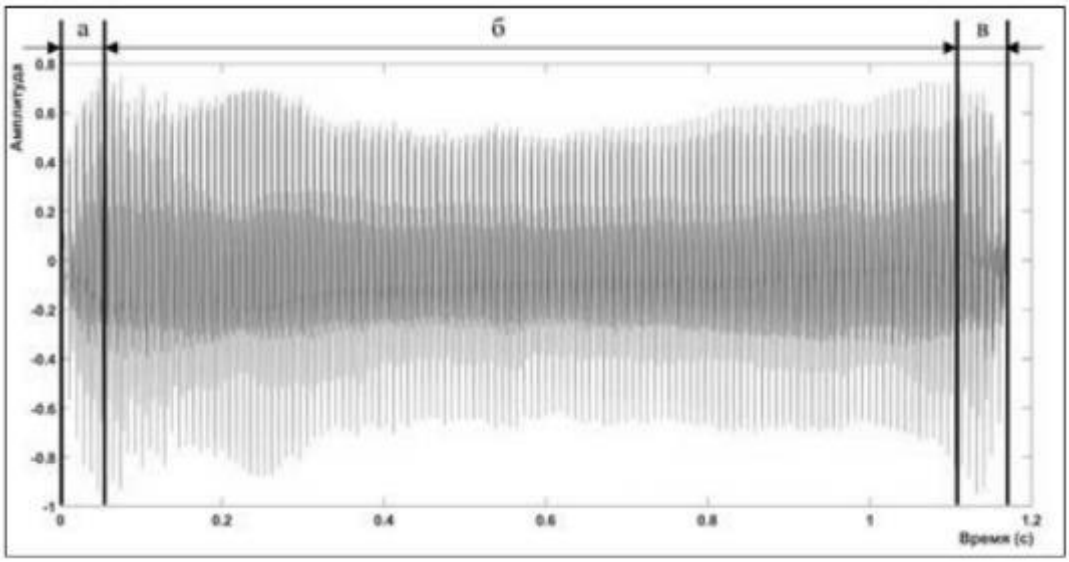

Рисунок 2.6. - Амплитудно-временное описание фонемы буквы «О»

При исследовании особенностей амплитуды и времени произношения фонем в определенный период можно предположить, что состояние фонем как компонентов голосового спектра остается неизменным. Как показано на рисунке 2.7, для анализа и описания звука необходимо разделить его на спектральные части.

Каждая верхняя точка на вышеприведенной схеме соответствует фонеме – форманте. Поэтому каждую фонему можно описать, исходя из простейших

структурных мер: частоты и относительной амплитуды. Математически эти два измерения образуют вектор.

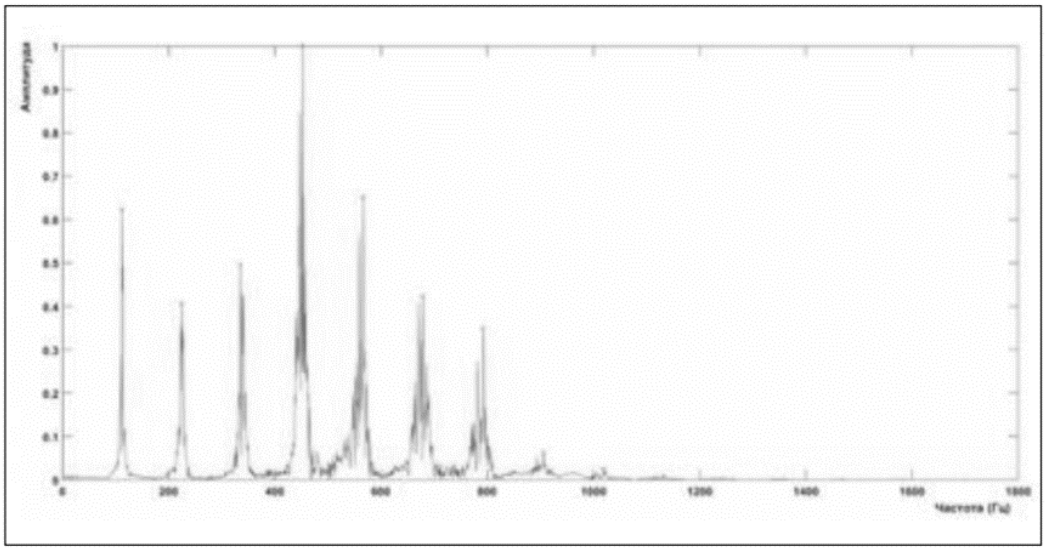

Рисунок 2.7. - Спектральные компоненты фонемы буквы «О»

Набор таких векторов, соответствующих существующим важным формантам, соответствует матрице параметров. В качестве метода оценки описания цифрового звукового образа (фонем) предложен пример реконструкции фонем человеческой речи для решения задачи автоматического синтеза речи.

Тогда фонему можно описать набором следующих параметров:

$$f(t) = \sum_{i=1}^{N} A_i \sin(2\pi v_i t) \qquad (2.11)$$

где, официальная запись форманты указана под знаком плюс.

Соответственно, используя значения амплитуд и частоты выбора модели звука, например, создать букву «у» и синтезировать его.

Набор мер значений фонемы зависит от ее характеристик. Например, для правильного синтеза голосовой записи гласной буквы «у» использовалась матрица, состоящая из восемнадцати числовых тактов и описывающая девять важных формант. Для создания более точной модели следует учитывать все важные

форманты фонем. Точность сравнения исходного сигнала (рис. 2.8.а) и синтезированного (рис. 2.8.б) выражается в одинаковой длительности звуков.

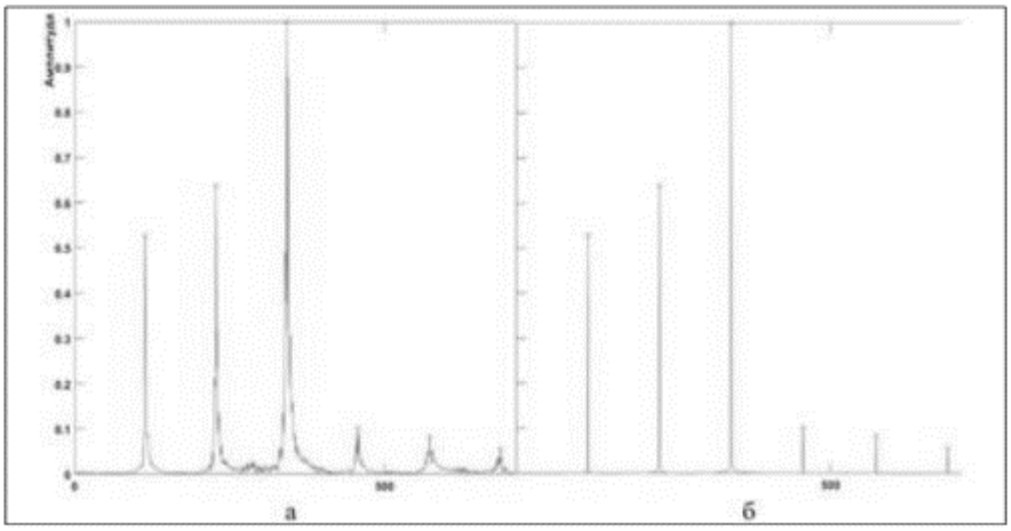

Рисунок 2.8. - Модель звука «у» а) оригинальный; б) синтетический (искусственный)

Методы синтеза речи. В настоящее время синтезированную речь производят различными методами, каждый из которых имеет свои преимущества и недостатки.

Синтезатор речи характеризуется двумя основными особенностями: естественностью голоса и следовательно, сложностью языка. Именно эти две характеристики учитываются при разработке проекта синтезатора. Некоторые синтезаторы речи характеризуются естественной передачей звука, другие отличаются высотой тона. Каждый метод синтеза речевого сигнала отличается сложностью алгоритма и основными принципами синтеза, используемыми в каждом отдельном приложении. В зависимости от поставленных целей для их построения используются различные методы синтеза речи, которые по основным принципам можно разделить на три группы:

- параметрический синтез;
- конкатенационный синтез;
- полный синтез по правилам.

Параметрический синтез речи – это конечная работа систем кодирования речи, в которой речевой сигнал представляется набором плавно изменяющихся

параметров. Этот метод рекомендуется использовать, когда набор сообщений ограничен и не меняется очень быстро.

Преимущества этого метода заключаются в следующем:

- возможность записи выступления на нужном языке и с нужным ведущим;

- параметрическое изображение имеет высокое качество в зависимости от степени сжатия информации;

- удобство для всех типов пользователей.

Основным недостатком этого метода является то, что его нельзя использовать для входящих сообщений, которые не определены заранее.

Конкатенативный синтез речи состоит в построении сообщения из заранее записанного словаря первичных элементов синтеза. Размер элементов синтеза состоит как минимум из одного слова. Содержание синтезируемых сообщений определяется размером словаря. Традиционно количество словарных единиц не превышает нескольких сотен слов. Основная проблема компиляционного синтеза - объем памяти для хранения словаря. В зависимости от этого используются разные методы сжатия/кодирования речевого сигнала.

Преимущества этого метода заключаются в следующем:

- удобство для пользователя;

- умение использовать произвольные выражения для синтеза.

Недостатками метода являются:

- необходимость большого объема памяти для хранения словаря;

- необходимость использования сжатия речевого сигнала.

Полный синтез речи по правилам. Синтез речи на основе правил основан на запрограммированном знании акустических и лингвистических ограничений и не использует напрямую элементы человеческого языка. В памяти сохраняются измеренные величины, полученные в результате анализа речевого сигнала, например правила соединения звуков со словами и словосочетаниями, правила формулирования.

В системах, основанных на методе синтеза, существует два подхода. Первый подход направлен на создание модели системы формирования языка человека,

называемой «синтез произношения». Второй метод – формантный синтез по правилам. Четкость и естественность таких синтезаторов можно повысить до значений, соответствующих свойствам естественного языка.

Метод конкатенативного синтеза является оптимальным для решения задачи проектирования и развития синтеза речи в таджикском языке. Учитывая особенности фонетических правил таджикского языка, для алгоритма синтеза речи используется наименьший слог, который образует последовательность слогов с приемлемой точностью. Комбинируя отдельные слоги, можно создать последовательность звуков с фиксированными размерами и паузами.

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. 2.9).

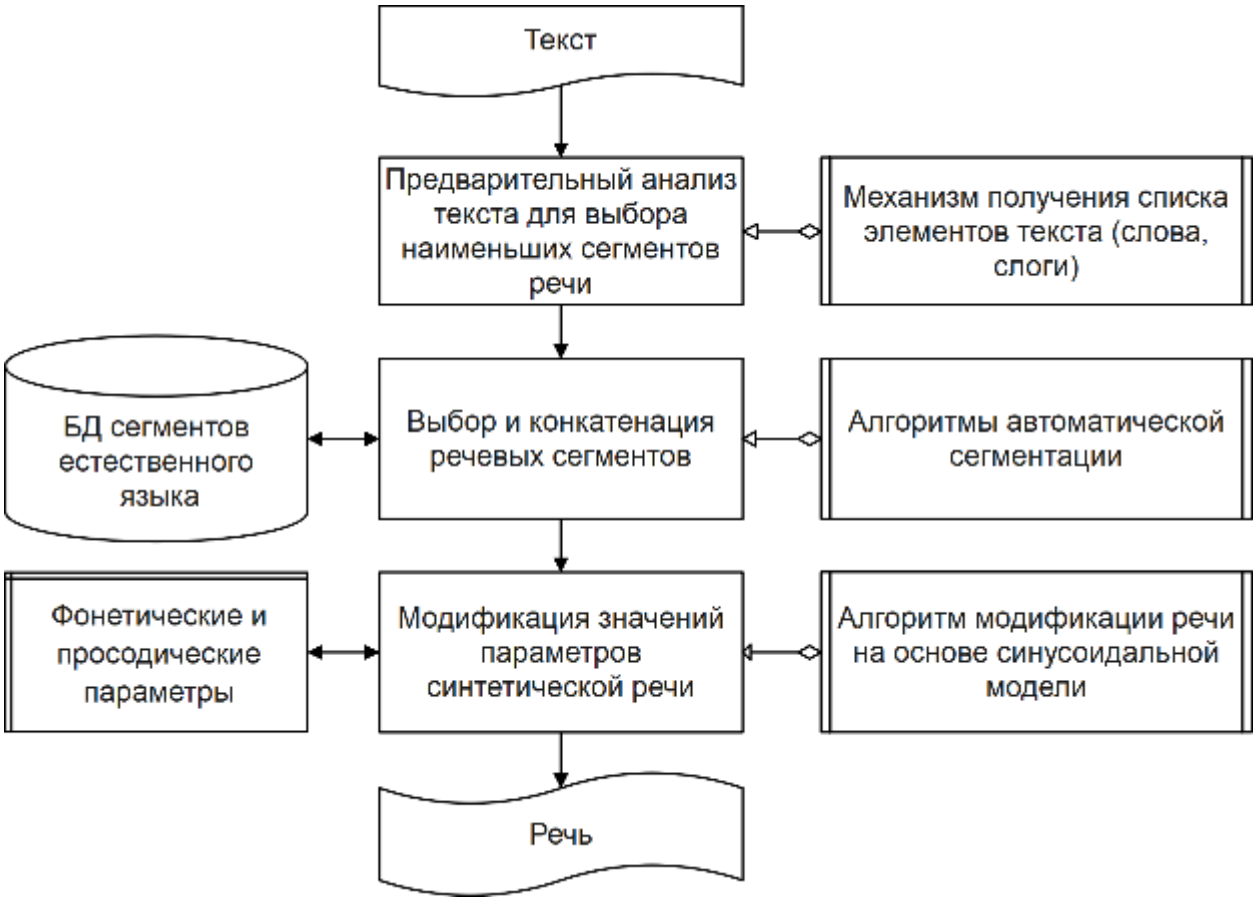

Рисунок 2.9. - Алгоритм синтеза речи на основе конкатенации элементов

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.

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:

$$\{w^*, t^*\} = argmax_{wt=T} g(w) \qquad (2.12)$$

где, ставится условие {w, g(w)} Є 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.

$$S^* = argmin_{w_1 \dots w_i \dots w_n = T} (n) \qquad (2.13)$$

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:

$$P_{j_{p+1}}(m,\tau) = \sum_{l \in L_m} P_{j_p}(l,\tau) b_{j_l}\left(O_{p+1}\right) \quad t-\tau+1 \le p < t \qquad (2.14)$$

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.

При определении нормальной рекурсии второго порядка для формулы мы предоставляем прямую вероятность наилучшего пути оценки от исходного списка до момента времени t, заканчивающегося на элементе j, предполагая, что правильным выбором является элемент k, начинающийся в момент времени t + 1. В этом случае вероятность наблюдения компонента j за период времени от t − τ + 1 до t равна Pjt (M, τ).

$$\alpha_t(k) = max_\tau \alpha_{t-\tau}(j) a_{jk} d_j(\tau) P_{j_t}(M,\tau) \qquad 1 \le t \le T \qquad (2.15)$$

В конце каждой рекурсии можно восстановить более вероятную последовательность списков, где j и τ большую часть времени обеспечивают вывод искомой части k.

Изменения синтетических речевых показателей. Одним из наиболее успешных и широко используемых методов создания синтетической речи в системах синтеза речи является объединение речевых компонентов. При таком подходе необходимо изменять просодические параметры (длительность, основная частота, амплитуда) речевых компонентов.

Алгоритм *модуляции речи* на основе синусоидальной модели используется в процессе кодирования, хранения, синтеза компонентов и защиты речевого сигнала от разрывов в виде пауз на границах компонентов.

Синусоидальная модель рассматривает речевой сигнал как результат прохождения функции возбуждения голосовой связки e(n) через изменяющуюся во времени линейную систему h(n), которая представляет особенности пути речи.

Модель учитывает, что упомянутая система письма включает в себя впечатления от речевых импульсов и реакцию на импульсы речевого тракта:

$$\sum_{k=1}^{L} a_k(n) \cos[(n - n_0)w_k] \qquad (2.16)$$

где,

ωk – частота каждой синусоиды;

ak(n) – соответствующая ей амплитуда;

L – количество синусоидов в речевом диапазоне частот;

n0 – время начала импульса основного тона.

Импульс основного тона возникает при суммировании всех синусоидальных волн в момент времени n = 0, который является центром кадра анализа.

С помощью преобразования Фурье [71] h(n) можно изменить время прохождения звука, которое выражается следующим образом:

$$H(w,n) = M(w,n) \cdot \exp\left(j \cdot \Psi(w,n)\right) \quad (2.17)$$

где, M(ω, n) и Ψ(ω, n) – амплитуда и фаза передаточной функции системы.

В результате пропускания функции возбуждения e(n) через нестационарную систему h(n) речевой сигнал представляется как сумма других синусоид:

$$s(n) = \sum_{k=1}^{L} A_k(n) \cdot \cos\left(\Theta_k(n)\right) \quad (2.18)$$

где, Ak(n) = ak(n) • Mk(n) и 2k(n) = (n − n0) • ωk + Θk(n).

Значения $M_k(n)$ и $\Theta_k(n)$ представляют собой амплитуду и фазу функции системы вблизи частоты, определяемой ωk.

Такое разложение компонентов смешанной системы на амплитуды и фазы сигналов позволяет рассматривать их независимо друг от друга при просодических изменениях (см. рис. 2.10).

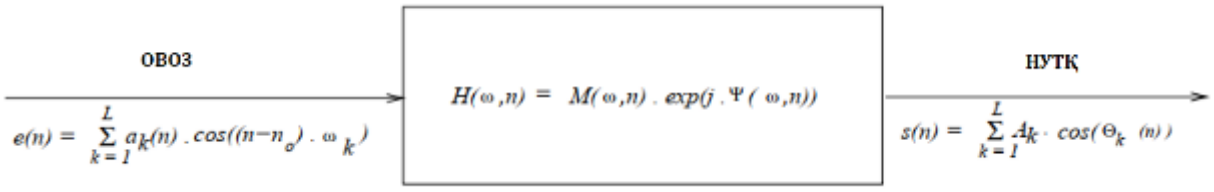

Рисунок 2.10. - Синусоидальная модель речеобразования

Синусоидальная модель может выполнять масштабные и качественные просодические изменения фонетических размеров. Длительность можно изменить, не затрагивая длительность основных периодов тона исходных компонентов и без необходимости цитирования или настройки. Частично невокализованные компоненты учитываются при смешанных стимулах, что позволяет избежать типичных ошибок, возникающих при бинарном разрешении звонких или невокализованных компонентов. При этом используется фиксированная частота

компонентов, поэтому нет необходимости определять их местонахождение перед анализом.

По результатам анализа характеристик человеческого голоса и цифрового образа речи определены новые формулы функции механизма синтеза речи на основе метода конкатенации речевых компонентов естественного языка. Установлено, что фрагмент звука можно представить в цифровом виде определенными размерностями звука, такими как амплитуда и частота дискретизации. На основе системы дифференциальных уравнений создана математическая модель цифрового образа речи.

Проведен сравнительный анализ возможностей и недостатков существующих методов синтеза речи. На основании полученных результатов был выбран метод объединения речевых элементов для разработки системы автоматического синтеза речи с естественным языком. 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 для проектирования и реализации механизма синтеза речи в таджикском языке, который подробно описан в шестой главе диссертационной работы.

Выводы по второй главе

Изучены проблемы компьютерной лингвистики таджикского языка, таких как автоматическая проверка орфографии, машинный перевод и автоматический синтез речи на основе математических моделей и методов их обработки. Нашли свое решение вопросы процесса решения проблем, основанный на правилах и

ресурсах естественного языка, то есть неклассифицированных корпусах текста, тезаурусе, онтологиях и электронных словарях.

Проблема моделирования лингвистических задач на основе математического аппарата исследована на примере английского и русского языков. Методом математического моделирования исследованы вопросы обработки текстовой информации на таджикском языке.

Для исследования проблем обработки текстовой информации на таджикском языке были использованы математические методы З.Д.Усманова, а также созданы общие способы построения и кодирования текстовых элементов, такие как слоговая структура слов, кодирование слов и предложений. Также с помощью математических методов обработки информации исследовались статистические закономерности некоторых элементов текста: слогов, слов, анаграмм, предложений.

Слоговые структуры слов таджикского языка были созданы с целью синтеза речи. Определены математические методы кодирования текстовых элементов для решения задач автоматической проверки правописания, машинного перевода текста и голосового синтеза на таджикском языке.

Проведен сравнительный анализ практических возможностей и недостатков существующих методов и на основе полученных результатов разработан относительно сбалансированный механизм системы автоматической обработки данных на таджикском языке, в частности:

1) *для автоматической проверки орфографии таджикского языка*: обработка электронных словарей; компьютерная обработка тезауруса; обработка математических моделей, специальных методов и алгоритмов обеспечения словарного и тезаурусного управления; инструмент для замены символами на основе стандартного таджикского алфавита; разработка алгоритма проверки орфографии; разработка комплекса компьютерных программ, обеспечивающих автоматическую проверку орфографии в тексте на таджикском языке;

2) *для машинного переводчика:* параллельная обработка корпуса для предоставления информации; разработка математических методов,

обеспечивающих поиск, сортировку и выделение текстовых элементов; разработка специальных алгоритмов на основе статистического метода машинного перевода; разработка алгоритмов замены букв в процессе машинного перевода; разработка комплекса компьютерных программ, обеспечивающих машинный перевод текста на таджикский язык;

3) *для синтеза речи:* предварительный анализ текста; выбор и соединение компонентов речи естественного языка из базы данных на основе автоматического алгоритма декомпозиции; изменение значений фонетического и просодического размеров синтетической речи с использованием синусоидальной модели синтеза речи; разработка комплекса компьютерных программ, обеспечивающих синтез речи на таджикском языке.

ГЛАВА 3. ОБЪЕКТНО-ОРИЕНТИРОВАННОЕ МОДЕЛИРОВАНИЕ СИСТЕМ ОБРАБОТКИ ТЕКСТА ЕСТЕСТВЕННОГО ЯЗЫКА

§3.1. Моделирование процессов

Методологической основой моделирования информационных систем является определение и анализ общей взаимосвязи взаимосвязанных объектов, а также достижение общих целей всеми рабочими группами. Информационная система характеризуется изменениями состояния объектов, происходящими в результате их совместного действия в различных процессах и во внешней среде с течением времени.

С точки зрения ученых и специалистов [19; 24; 25; 28; 31; 37] в области информационных технологий процесс управления информационной системой, направленный на ее объекты для достижения цели, можно показать в виде информационного процесса. В этом случае объекты связывают внешнюю среду и информационную систему воедино. Внешняя среда и объект управления предоставляют в информационную систему данные о своем состоянии. Информационная система анализирует данные, реагирует на процессы внешней среды определенным ответом, меняет свою цель и структуру при необходимости.

Информационная система – совокупность организационных, технических, программных и информационных средств, созданных вместе с целью сбора, хранения, обработки и представления информации, необходимой для выполнения управленческих операций. Информационная система связывает совокупность объектов и внешнюю среду посредством информационного процесса:

- информационный процесс сообщает о внешней ситуации в информационную систему, созданной на основе критериев особенностей закономерностей;

- информационный процесс передает из информационной системы во внешнюю среду отчетные данные с целью принятия решений пользователями;

- информационный процесс передает от информационной системы к объекту управления, который представляет собой совокупность плановой, нормативной и распределительной информации для реализации процесса управления.

Методология проектирования и интегрированная архитектура программ. Как определено в формуле 1.1., модель проекта информационной системы обработки текста состоит из набора взаимосвязанных информационных технологий, процессов, алгоритмов, набора методов обработки текста, инструментов, интерфейсов и набора процессов. Предлагаемая модель представляет собой цифровое представление таджикского языка.

Последовательность обработки текста состоит из процессов поиска, обработки, анализа и понимания текстовых элементов. Для относительно доступного анализа модели информационной системы предлагается следующая схема, включающая этапы и рабочие процессы обработки текстовых данных (рис. 3.1).

Рисунок 3.1. - Процессы обработки текстовых данных

На основе созданной модели разработан комплекс компьютерных проектов и система автоматической обработки текстовых данных. Краткая информация о проектах представлена в следующих главах.

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

Унифицированный язык моделирования UML (Unified Modeling Language) используется для графического описания и моделирования объектов информационной системы. В рамках визуального моделирования язык UML широко использует стандарты объектно-ориентированной методологии. UML определяет четыре основных типа моделей информационной системы:

- статическаю модель;
- динамическаю модель;
- модель взаимодействия объектов;
- физическая модель.

Язык UML – это преемник объектно-ориентированных методов анализа и проектирования, появившихся в конце 1990-х годов. Истоки UML можно проследить до конца 1994 года, когда Грэди Буч [99] и Джеймс Рембо [100] под руководством Rational Software начали объединять Буча и OMT (технику объектного моделирования). В конце 1995 года они создали первую единую классификацию методов, получившую название «*Unified Method*». Также в 1995 году к рабочей группе присоединился Ивар Якобсон, изобретатель метода OOSE (объектно-ориентированная разработка программного обеспечения) [101]. В этом смысле UML представляет собой составной продукт методов Буха, Рембо и Якобсона.

Язык UML был включен в процесс стандартизации OMG (Object Management Group) (Группы управления объектами) и в настоящее время считается мировым стандартным языком моделирования программного обеспечения.

Все крупные компании, являющиеся производителями CASE-продуктов, помимо Rational Software (Rational Rose), такие как Microsoft, IBM, Hewlett-Packard, Oracle, Sybase, Paradigm Plus (CA), System Architect (Popkin Software), Microsoft Visual Modeler используют UML для своих продуктов.

Средства UML. UML, как считают авторы, это язык определения, визуализации, проектирования, конструирования и документирования программного обеспечения и организационных, экономических и технических систем. UML содержит ряд стандартных диаграмм и различных символов. Стандартная интерпретация UML, принятая OMG, предлагает следующие общие типы диаграмм для моделирования:

1. Схема рабочего процесса используется для моделирования бизнес-процессов и требований к создаваемой информационной системе.

2. Диаграммы классов используются для моделирования статистической структуры классов и связей между ними.

3. Поведенческие диаграммы используются для моделирования процесса обмена информацией между объектами информационной системы.

4. Диаграммы взаимодействия делятся на две группы: диаграммы последовательности и диаграммы кооперации. Они используются для моделирования взаимодействия объектов информационной системы в конкретный момент времени и структуры взаимодействия.

5. Диаграммы состояний используются для моделирования поведения объектов информационной системы при переходе из одного состояния в другое.

6. В каждом варианте использования диаграммы деятельности используются для моделирования работы информационной системы или деятельности субъектов.

7. Диаграмма компонентов используется для моделирования иерархии компонентов информационной системы, таких как рабочие файлы, процедуры и базы данных**.**

8. Диаграмма развертывания используется для моделирования физической архитектуры информационной системы.

В настоящее время разработаны и предложены другие типы UML-диаграмм для расширения возможностей моделирования информационных систем. Но для моделирования информационных систем разного профиля достаточно разработок, в приведенном выше списке.

§3.2. Моделирование поведения информационной системы

В рамках моделирования информационных систем в первую очередь учитывается то, как они работают. Работа информационной системы в основном определяется на основе составления диаграммы состояний использования и деятельности.

*Диаграмма вариантов использования***.** Понятие вариантов использования (use case) впервые было введено Иваром Джейкобсоном в процессе моделирования UML. К настоящему времени она используется для анализа состояния работы информационной системы и стала основным компонентом разработки и планирования проекта.

Ситуация использования представляет собой последовательность действий, которые реализует информационная система в ответ на события внешнего субъекта, то есть исполнителя.

Пользовательская ситуация обеспечивает взаимодействие пользователя и информационной системы. В простейшем случае ситуация использования определяется путем анализа всех функций пользователей.

Действующее лицо (Actor) – это образ, который пользователь играет свою роль в информационной системе с субъективным отношением.

Исполнитель играет только свою роль, а не конкретных лиц или набор должностей. Кроме того, в диаграмме ситуации использования исполняющие лица могут представлять себя как внешний субъект, для доступа к некоторой информации в информационной системе. Отметка исполнителей в таблице будет известна только тогда, когда им обязательно понадобятся некоторые из варианта использования.

Пользователи делятся на три типа: пользователи информационной системы, неотъемлемая часть внешней информационной системы, взаимодействующая с проектом, и оборудование. Устройство считается активным лицом, если от него зависит запуск какого-либо события информационной системы.

Для наглядного показа работы как основных элементов процесса обработки текстовых данных в информационной системе, можно использовать следующую диаграмму вариантов использования (рис. 3.2).

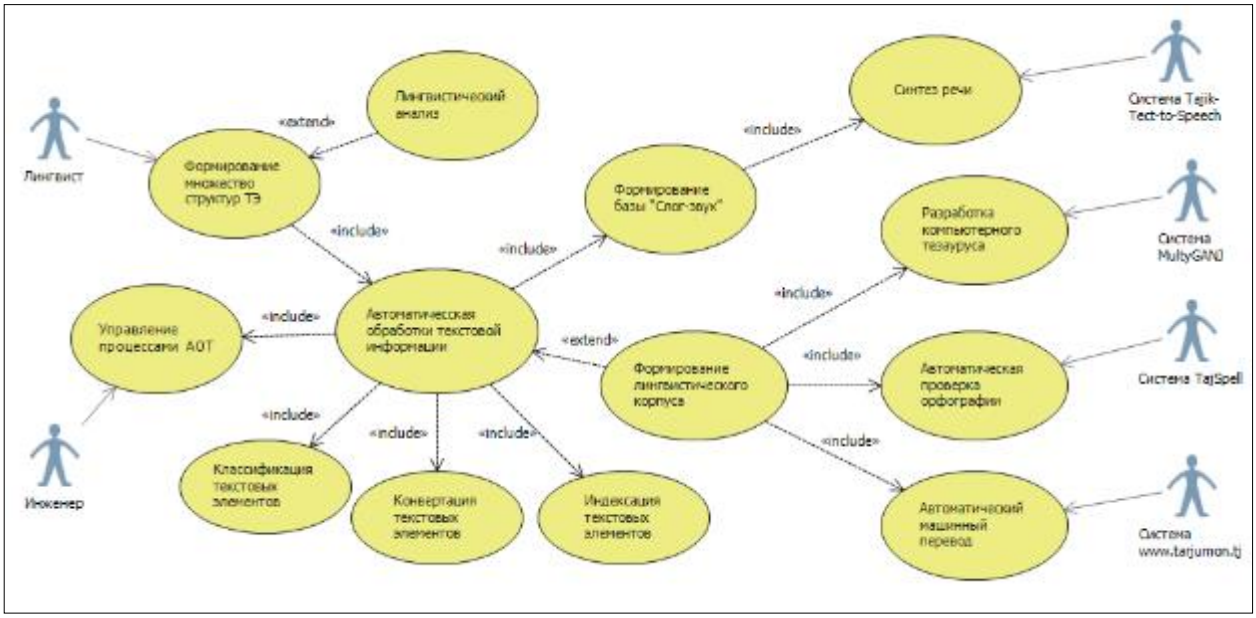

Рисунок 3.2. - Диаграмма варианта использования для информационной системы

В приведенной схеме основные роли выполняют лица исполнители и составные части информационной системы. Знаки в форме эллипса обозначают вариант использования, основанные на различных отношениях, то есть знаки в форме стрелок указывают на разные внешние и внутренние связи исполнителей информационной системы.

На диаграмме показаны два действующих лица: инженер и специалист-языковед в роли лингвиста. Кроме того, свою роль в ней играют четыре неотъемлемые части информационной системы: Tajik text-to-speech, MultiGANJ, TajSpell, tarjumon.tj. Также на схеме определены основные действия, выполняемые информационной системой: комплексная организация текстовых элементов, управление процессами обработки данных, автоматическая обработка текстовых данных, синтез голоса, обработка тезауруса, проверка орфографии и машинный перевод.

Взаимодействие варианта использования и исполнителей представлено на диаграмме, приведены основные требования к информационной системе с точки зрения поведения пользователей. Таким образом, функция, выполняемая информационной системой – это вариант использования. Исполнитель – это заинтересованное лицо в рамках создаваемой информационной системы. На приведенной выше диаграмме показаны усилия, предпринятые исполнителями в отношении варианта использования, чтобы у него была необходимая информация.

На примере предложенной схемы лингвист задействован в большом количестве различных вариантов использования: «Комплексная организация элементов текста», «Лингвистический анализ», «Обработка текстовых данных», «Синтез голоса», «Обработка тезауруса», «Проверка орфографии» и «Машинный перевод».

Из варианта использования «Обработка текстовых данных» стрелка направлена в сторону «Синтез голоса». В этом случае между ними возникает зависящая от них ситуация «Образование слого-звуковой основы».

Исполнители могут быть неотъемлемой частью внешней информационной системы, поэтому в данном случае часть «Tajik text-to-speech» показана в качестве исполнителя. Он направлен на обеспечение голосового синтеза, то есть компьютерного произношения текста, предоставленного лингвистом. Стрелка от варианта использования к исполнителю показывает ожидание информационной системы некоторой информации от исполнителей. В этом случае в варианте использования «Организация слого-звуковой базы» для информационной системы отображается информация об общем запасе произнесенных слогов для произношения.

Тем не менее, все варианты использования связаны с требованиями внешних сторон для реализации установленной цели внутри информационной системы.

Чтобы создать реальные пользовательские задачи и рассмотреть альтернативные варианты решения проблемы, вариант использования всегда должен анализироваться совместно с исполнителем. Исполнитель может

совершать любые действия относительно варианта использования. Важность различных ролей исполнителя зависит от подхода к ним.

Основным назначением схемы варианта использования является документирование информационной системы, то есть всей информации, входящей в среду внедрения информационной системы и исполняющих лиц как внутри, так и вне среды, с указанием отношения между ними.

Для правильного моделирования карты варианта использования необходимо соблюдать следующие правила:

1. Между исполнителями не должно быть никаких родственных связей. С этой точки зрения исполнители находятся вне операций информационной системы. Это указывает на то, что отношения между ними не зависят от их способностей.

2. Взаимосвязь между двумя варианта использования не может быть определена одним пунктом. Этот тип диаграммы описывает, какой вариант использования может получить доступ к информационной системе. Диаграмма деятельности используется для отображения последовательности действий в каждом варианте использования.

3. Вариант использования должен содержать инициативу исполнителей.

То есть стрела целиком должна быть направлена от направления исполняющих лиц к направлению варианта использования.

Лучшим источником определения варианта использования является анализ событий, осуществляемый вне информационной системы. Начать следует с перечисления всех событий, произошедших во внешней среде. В этом случае информационная система должна быть проинформирована о порядке событий. Определение событий – это впечатление о возможности внедрения в информационную систему. Для функционирования информационной системы необходимы базовые компоненты. Эти элементы перечислены в документе «фактический процесс». Документирование обработки данных, выполняемой в каждом варианте использования, является целью фактического процесса. Подход пользователей информационной системы подробно описывается в этом документе.

Реальный процесс не должен зависеть от способа его выполнения. Основная цель – описание работы информационной системы, а не то, как выполняются рабочие процессы. Фактически процесс состоит из следующих частей:

- краткого описания;
- начального условия;
- процесса основных показателей;
- реального альтернативного процесса;
- заключительного условия.

Краткое описание. Каждый вариант использования должен иметь краткое описание, в рамках которого поясняются любые действия, совершаемые как пользователем, так и информационной системой. Например, в случае «Обработка текстовых данных» система обработки текстовых данных включает следующие комментарии:

Вариант использования «Комплексная организация элементов текста» и «Управление процессами обработки» для специалиста-лингвиста и инженера соответственно дают возможность проводить анализ и контролировать процесс обработки текста.

Начальное условие. К начальному условию варианта использования относятся такие условия, которые должны быть выполнены до начала самой основной ситуации. Примером такого условия является выполнение другого условия или определение доступа пользователя на его выполнение. Не все варианты использования имеют предпосылки. Поэтому можно сказать, что диаграмма варианта использования не отображает порядок внедрения информационной системы.

Через начальное условие можно создать документ об исходных данных. Исходное состояние одной ситуации может быть состоянием варианта использования. Например, реализация условия «Комплексная организация текстовых элементов» может быть выполнена в качестве предварительного условия для «Голосового синтеза», «Обработки тезауруса», «Проверки орфографии» и «Машинного перевода».

Основной и альтернативный реальный процесс. Истинное выражение варианта использования представляется в реальном основном и альтернативном процессах. Фактический процесс шаг за шагом объясняет, как это происходит, когда возможности ситуации реализуются. К основному и альтернативному фактическому процессу относятся следующие случаи:

- способ начать работу;
- разные способы исполнения;
- настоящий умеренный или основной процесс;
- возврат из основного реального процесса;
- процесс возникновения ошибок или препятствий;
- способ закончить.

Например, реальный процесс варианта использования «Обработка текстовых данных» интерпретируется следующим образом:

Основной процесс.

1. Вариант использования начинается с момента входа лингвиста в информационную систему;

2. Информационная система выдает список возможных операций;

- синтез речи;
- обработка тезауруса;
- проверка орфографии;
- машинный перевод.

3. Информационная система «приветствует» пользователя и требует выбора операций.

4. Пользователь выбирает необходимую операцию.

5. Пользователь выбирает раздел «Синтез речи».

6. Информационная система просит ввести текст для произношения.

7. Пользователь вводит текст.

8. Информационная система определяет наличие орфографических ошибок во введенном тексте. При обнаружении орфографической ошибки выполняется

набор альтернативных операций А1. Если во время проверки орфографии возникает ошибка, выполняется набор действий Е1, ориентированный на ошибку.

9. Информационная система делит слова в тексте на слоги. Если слог не соответствует слоговой структуре, т.е. не соответствует таджикским правилам правописания, выполняется альтернативная операция А2.

10. Информационная система осуществляет поиск звукового файла слогов из источника данных «слог-звук» и направляет его на произношение.

11. Информационная система представляет пользователю голосовой файл введенного текста.

12. Статус занятости заканчивается.

Набор альтернативных операций А1. Проверка правильности введенного текста:

1. Информационная система сообщает об ошибке в слове, в тексте;

2. Для пользователя предлагается реализация варианта использования «Проверка орфографии»;

3. Вариант использования заканчивается.

Набор альтернативных операций А2. Разделение слова на слоги:

1. Информационная система сообщает пользователю, что в источнике данных «слог-звук» нет звукового варианта слога.

2. Пользователю предлагается продолжить произнесение текста, не рассматривая слово с ошибкой.

3. Вариант использования заканчивается.

Ошибка процесса Е1. Ошибка проверки орфографии:

1. Информационная система сообщает пользователю, что при вводе текста произошла орфографическая ошибка, и предлагает выбрать правильное слово из списка, подготовленного в источнике данных.

2. Информационная система записывает ошибку в книгу ошибок.

3. Пользователю предлагается продолжить произнесение текста, не рассматривая слово с ошибкой.

4. Вариант использования заканчивается.

Заключительные условия – это условия, которые всегда выполняются после окончания варианта использования.

По окончании варианта использования на какой-либо строке ставится отметка, которая может быть использована в дальнейшем как начальное условие через конечное условие, информация о порядке выполнения варианта использования. Например, после проверки правописания текста производится его произношение. Проверка написания слов на ошибки выражается как заключительное условие процесса произношения.

Отношения между вариантом использования и исполнителями. В языке UML в диаграмме состояний процесса определены несколько типов отношений между элементами диаграммы. Это линейные отношения, отношения включения, отношения расширения и соединения.

Отношение линии связи – это отношения между состоянием процесса и исполнителя. На языке UML взаимосвязь линии связи представлена одной линией со стрелкой. Направление стрелки указывает на инициатора линии связи.

Отношение взаимозависимости содержания используется в том же случае, если какая-либо часть работы информационной системы повторяется несколько раз. Благодаря этой взаимосвязи можно идентифицировать несколько процессов. Пример, как показано на рисунке 3.3., включает состояние «Синтеза речи» и «Формирования слого-звукового базы», а произношение слова зависит от подготовки речевого варианта слога.

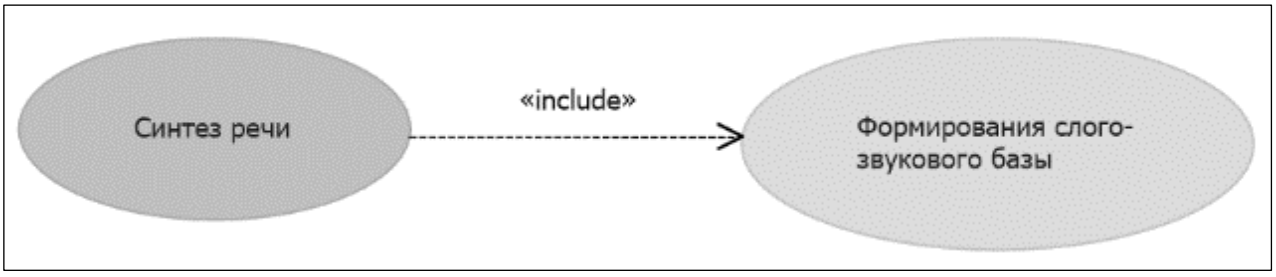

Рисунок 3.3. - Отношение зависимости вкючения

Отношение расширяющей зависимости используется, когда любая часть работы информационной системы повторяется один раз при возникновении определенного условия. Благодаря этой взаимосвязи можно определить условные процессы. Например, как показано на рисунке 3.4., выполнение условия «Комплексная организация текстовых элементов» зависит от условия состояния «Лингвистический анализ».

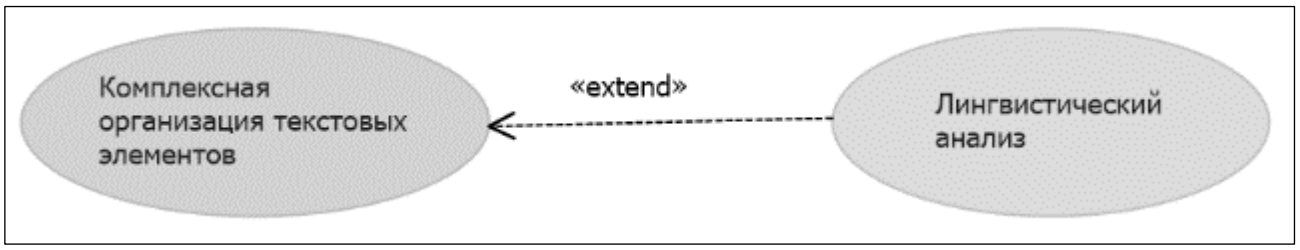

Рисунок 3.4. - Отношение зависимости расширения

Посредством отношений объединения можно показать общие особенности исполнителей. Например, пользователи могут быть двух типов – публичные и индивидуальные. Эти отношения можно использовать при моделировании группировки пользователей (рис. 3.5.).

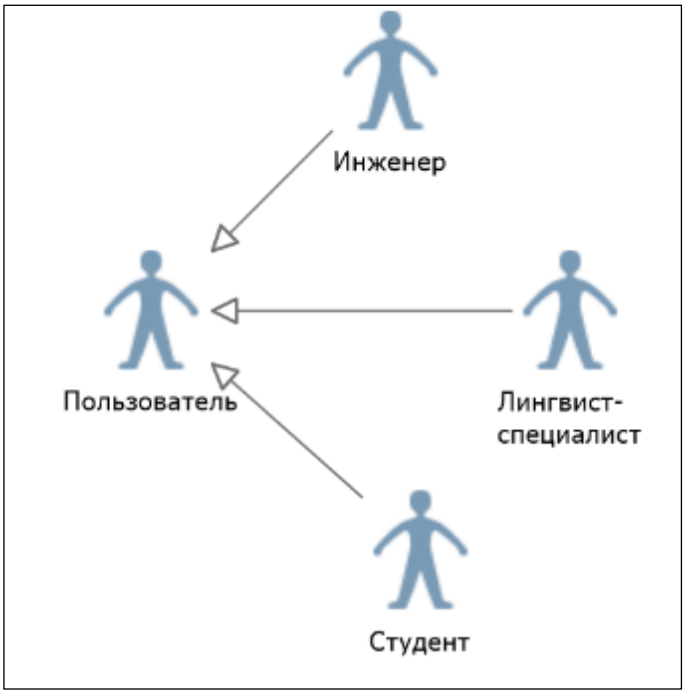

Рисунок 3.5. - Объединение исполнителей

Как показано на рисунке 3.5., устанавливать такую связь нужно только в том случае, когда способ действия людей, выполняющих один процесс, отличается от способа действия другого процесса. Если определенная группа пользователей совместно выполняет один и тот же процесс, нет необходимости указывать совместность исполняющих лиц.

Диаграмма варианта использования является основным инструментом организации требований к информационной системе. Каждый вариант использования имеет свои особые требования к обрабатываемому проекту.

§3.3. Модель взаимодействия объектов информационной системы

Диаграмма взаимодействия описывает процесс взаимодействия между группой объектов информационной системы. Обычно диаграмма взаимодействия моделирует поведение объекта только в одном состоянии существования, который состоит из серий объектов и различных общих для них сообщений.

Сообщение – это средство, с помощью которого отправляемый объект запрашивает выполнения действия у принимающего объекта.

Информационные сообщения – это сообщения, в которых принимающий объект предоставляет некоторую информацию для изменения своего состояния.

Сообщение запроса – это сообщение, которое предоставляет запрос какой-либо информации о принимающем объекте.

Строгое сообщение – это сообщение, требующее определенных действий от принимающего объекта.

В случае моделирования информационной системы обрабатываются два типа диаграмм сотрудничества: диаграмма последовательности и диаграмма кооперации.

Диаграмма последовательности отражает процесс событий, происходящих в варианте использования в зависимости от определенного времени. Например, ситуация «Синтез речи» показывает несколько этапов, таких как проверка написания слова, деление слова на слоговую структуру; деление слова на слоги;

поиск слоговых звуков в базе данных «слог-звук». В случае отсутствия его в источнике данных, предоставляет другой вариант слова и его произношение по слогам. Успешный синтез речи показан на рисунке 3.6.

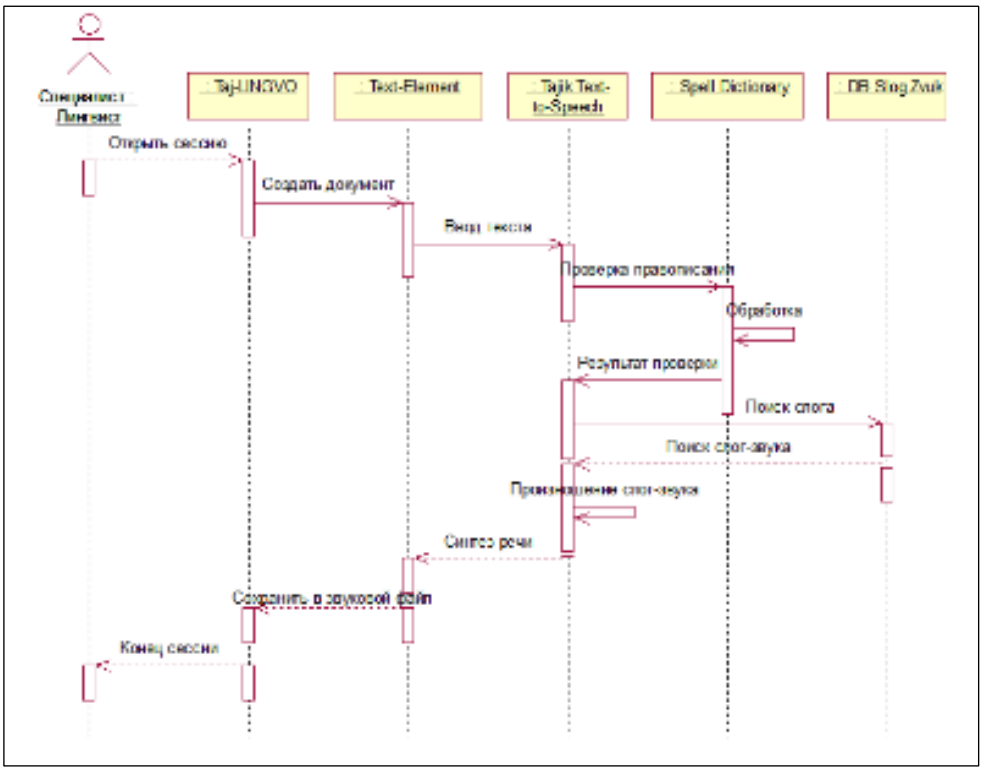

Рисунок 3.6. - Диаграмма последовательности процесса «Синтез речи»

Эта диаграмма описывает последовательность фактического процесса в варианте использования «Синтез речи». Все исполнители в верхней части схемы показаны как объекты информационной системы. В представленном примере исполнитель – специалист-лингвист. Объекты, активные в информационной системе в процессе «Синтез речи», также представлены в верхней части схемы.

Стрелка на диаграмме используется в направлении между исполнителем и объектами. На диаграмме последовательности объект представлен в виде прямоугольника, через который проходит вертикальная линия. Эта линия называется линией жизни объекта. Он представляет собой часть жизни объекта в процессе взаимодействия внутри процесса.

Сообщения между собой представляются в виде стрелок между линиями жизни двух объектов.

Лучший способ получить доступ к некоторым объектам – это узнать названия имен в реальном процессе. Также возможно прочитать документ, написанный в виде настоящего сценария.

Под понятием сценария рассматривается истинный тип реального процесса. Реальный процесс варианта использования «Синтез речи» предоставляет информацию о человеке, который получает из информационной системы произношение введенного текста посредством таджикского подпрограммы Tajik Text-to-speech. Не все объекты проявляются в реальном процессе. Существует возможность наличия в процессе объектов управления, которые контролируют последовательность выполнения процесса в заданном состоянии.

Диаграмма кооперацій. Второй тип диаграммы взаимодействия – это диаграмма кооперацій. Как и диаграмма последовательности, она показывает процесс событий конкретного сценария в рамках данного варианта использования. Если диаграмма последовательности выполняется во времени, диаграмма кооперацій организует отношения между объектами с точки зрения структуры. На рисунке 3.7 показана диаграмма кооперацій, отражающая процесс синтеза речи.

Из структуры диаграммы следует, что она содержит всю информацию диаграммы последовательности, но диаграмма кооперацій объясняет реальный процесс по-другому. Взаимодействие между объектами легче понять из диаграммы кооперацій, чем из диаграммы последовательности.

В этих целях для любого сценария изображаются оба типа диаграмм. Хотя обе диаграммы служат одной и той же цели и содержат одну и ту же информацию, они представлены с разных точек зрения.

В кооперативной диаграмме, как и в диаграмме последовательности, стрелка представляет сообщение и реализует обмен информацией в рамках варианта использования. Их временная последовательность обозначается нумерацией сообщения.

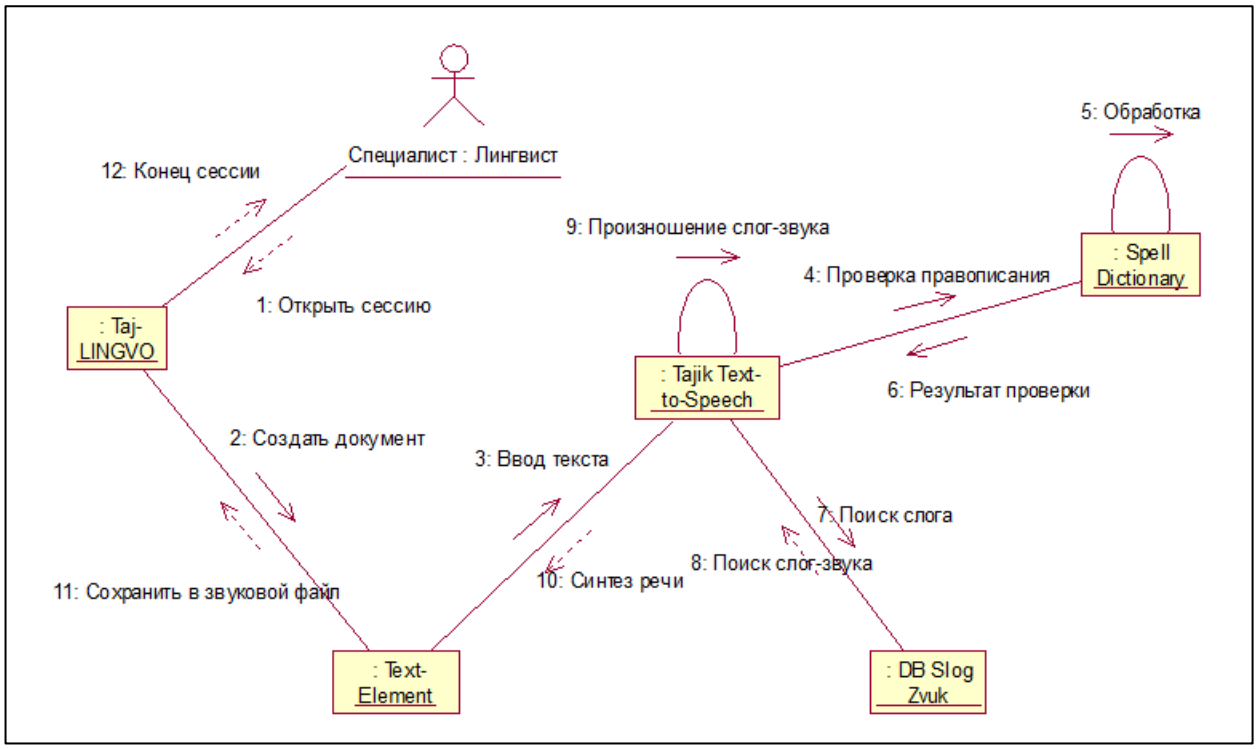

Рисунок 3.7. - Кооперативная диаграмма для синтеза речи

В целом, для моделирования взаимодействия объектов для каждой ситуации занятости диаграмма последовательности и взаимодействия разрабатывается отдельно. Следует отметить, что может наблюдаться возможность реализации и проявления одной или нескольких новостей в двух и более вариантах использования. В этом случае последовательно-кооперативная диаграмма может быть средством решения задачи для моделирования взаимодействия объектов информационной системы. Набор объектов, полученный в процессе сотрудничества, может стать основой для моделирования статистической структуры информационной системы.

§3.4. Концептуальная модель и логическая структура информационной системы

Диаграмма классов определяет набор различных классов информационной системы и статистические отношения между ними. На диаграмме показаны

характеристики класса, его операции и правила, используемые для отношений классов.

Диаграмма классов подсистемы «Обработка текстовых данных» показана на рисунке 3.8.

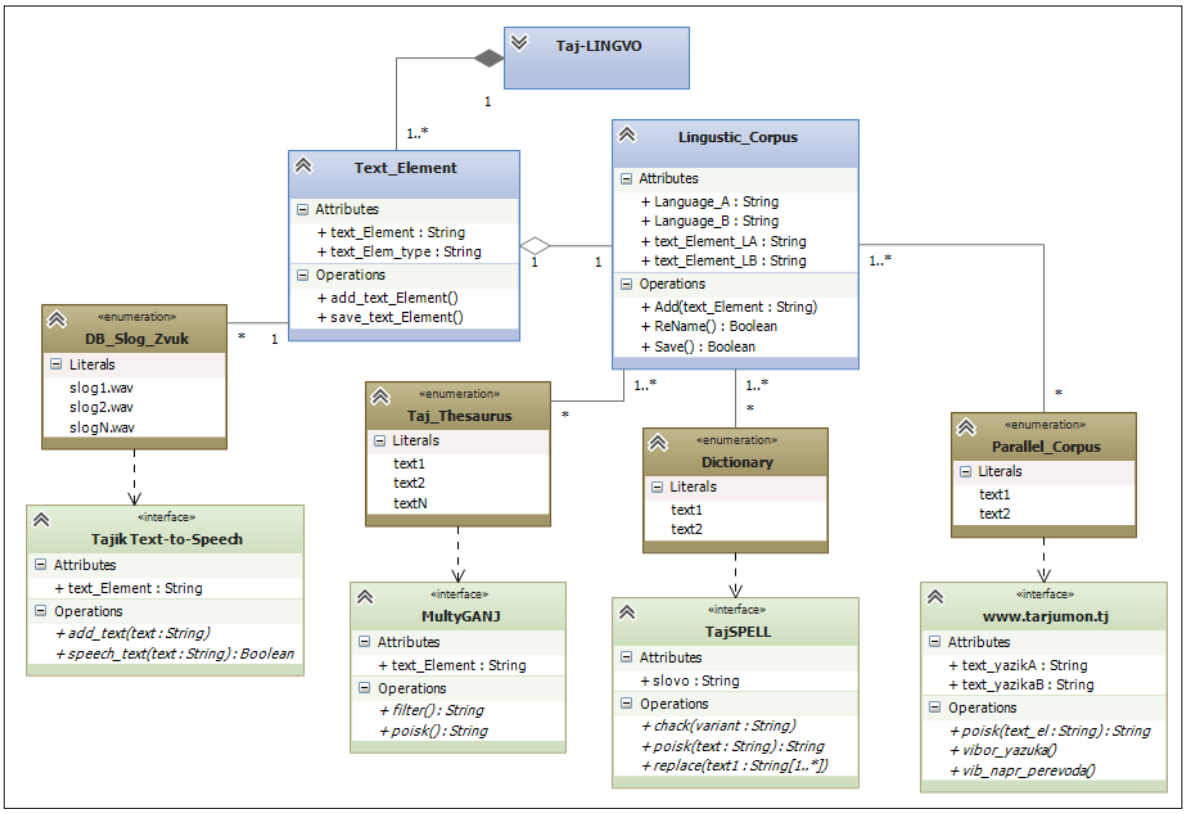

Рисунок 3.8. - Диаграмма классов подсистемы “Обработка текстовых данных”

На этой диаграмме показаны логические взаимоотношения между классами, в которых реализован вариант использования «синтез речи». В этом процессе предусмотрены 4 класса реализации: TajLINGVO (основной инструмент обработки данных на таджикском языке), Text_Element (текстовый элемент), DB_Slog_Zvuk (источник слоговых речевых данных) и Tajik Text-to-Speech (рабочая среда для произношение текста). Прямоугольник, разделенный на три части – это представленный вид класса. Первая часть содержит имя класса, вторая – наименование его функций, третья – поведение класса.

На диаграммах классов, да и вообще во всех типах диаграмм, для именования используется латинский алфавит, поскольку он поддерживается языками

программирования. Использование другого алфавита для обозначения объектов, операций и атрибутов (символов). приводит к ошибкам, поскольку инструменты CASE это не поддерживают.

Отношения классов представляют линиями взаимодействия между классами. Класс Text_Element (текстовый элемент) родственен классу DB_Slog_Zvuk (источник данных слоговых звуков), поскольку между ними происходит обмен информацией, и они работают вместе. Класс Text_Element не имеет отношения к таджикскому классу преобразования текста в речь, поскольку они не получают информацию друг от друга.

Стереотипы классов – это механизм, который делит классы на типы. В языке UML существует в основном три типа классов: Boundary (пограничный); Entity (сущностный); Control (управляющий).

Пограничные классы – это классы, расположенные на границе информационной системы или вообще в среде вокруг нее.

Примеры классов этого типа включают дисплеи, отчеты, интерфейсы для специальных устройств (например, динамиков, сканеров или принтеров) и интерфейсы для других информационных систем. Чтобы найти широкий класс, вам нужно обратить внимание на диаграмму вариантов использования.

Для каждого взаимодействия между исполнителем и вариантом использования должно существовать хотя бы одно отношение граничного класса. Именно этот класс позволяет исполнителю взаимодействовать с информационной системой.

Классы объектов содержат базовую информацию. Они не очень полезны пользователю, поэтому в их именах обычно используются доменные термины. Обычно для каждого класса объектов в базе данных создается одна или несколько таблиц.

Класс управления отвечает за согласованность действий других классов. Обычно каждый вариант использования должен содержать класс-обработчик, который управляет последовательностью событий в этом случае. Класс контроллера отвечает на игру, но сам ничего не делает, поскольку другие классы

не отправляют ему пакет сообщений. В ответ на это действие класс этого типа отправляет несколько сообщений. Следовательно, класс менеджера отвечает только за другие классы и иначе называется классом менеджера.

В информационной системе также могут работать руководители других классов, общих для нескольких вариантов использования. Например, класс Lingustic_Corpus отвечает за обработку событий правописания и перевода. Класс Text_Element управляет взаимодействием сообщений с операциями базы данных. Существуют и другие классы управления для работы с другими элементами информационной системы, такими как распределение ресурсов, распределенные вычисления или обработка ошибок.

На основе методов классификации вышеуказанных классов также можно сформулировать собственный метод в зависимости от отчета о проблеме.

Механизм пакетов. Пакеты используются для группировки нескольких классов, имеющих некоторые общие свойства. Существуют разные методы классификации. Они сгруппированы по особому шаблону. В этом случае пакетом будет пакет с классом объекта, пакет с граничным классом, пакет с управляющим классом и т.д. Такой режим работы эффективен с точки зрения предоставления готовой информационной системы, поскольку все граничные доступные в устройстве классы помещены в один пакет.

Иную точку зрения на объединение классов представляет основа их деятельности. Например, пакет безопасности содержит все классы, отвечающие за безопасность. В данном случае другими пакетами являются «Обслуживание пользователей», «Отчетность» и «Обработка ошибок». Преимуществом этого подхода является его возможность повторного использования.

Применить механизм привязки можно не только к классам, но и к любому элементу модели. Если нет требований к группировке классов, это необязательно. Один из них, в основном используемый в UML, – это зависимость. Связь между двумя пакетами может существовать при условии существовании логической связи между двумя классами пакетов. Итак, диаграмма пакета, приведенная на рисунке 3.9 состоит из набора классов и зависимостей между ними.

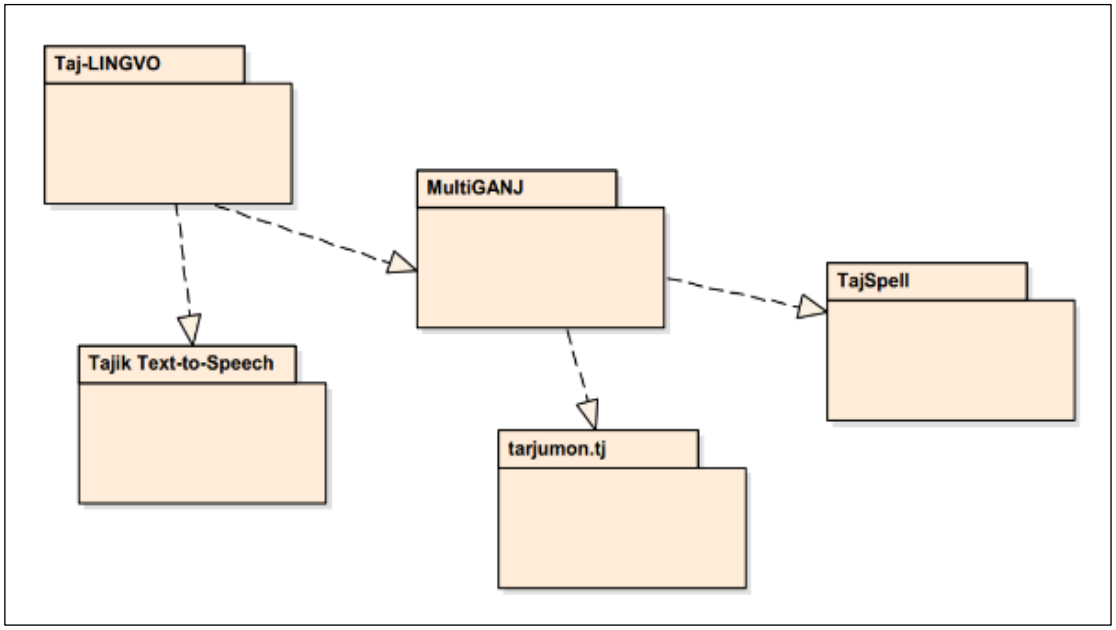

Рисунок 3.9. - Схема пакета информационной системы TajLINGVO

Пакеты и зависимости являются элементами диаграммы классов, поэтому диаграмма пакетов – это тип диаграммы классов.

Связь между двумя элементами существует, когда изменение определения одного элемента приводит к изменению другого элемента. Причины зависимости в классах могут быть разными: один класс отправляет сообщения другому классу, один класс содержит часть данных другого, один класс использует другой класс в качестве меры действия. Если класс меняет свою среду, любое отправляемое им сообщение может стать недействительным.

Как можно уменьшить количество зависимостей в информационной системе? Пакеты не отвечают на эти вопросы. Но они помогают в выборе зависимостей. После того как все это будет сделано, останется только сократить рабочие процессы. Диаграмму пакета также можно назвать основным инструментом управления общей структурой информационной системы. Пакеты являются важным инструментом для крупных проектов.

Атрибут (принадлежность, особенность) – это информационный элемент, связанный с классами. Например, в классе Text_Element (текстовый элемент) есть свойства text_Element_Name (имя), text_Element_Type (тип). Поскольку свойства содержатся внутри класса, они остаются закрытыми для других классов. Связывая

их, необходимо указывать, какие классы имеют доступ к чтению и изменению свойств. Этот тип отношений называется «видимым». В отношениях можно определить четыре типа доступа к значениям. Каждый из них можно проанализировать на следующих примерах (рис. 3.10).

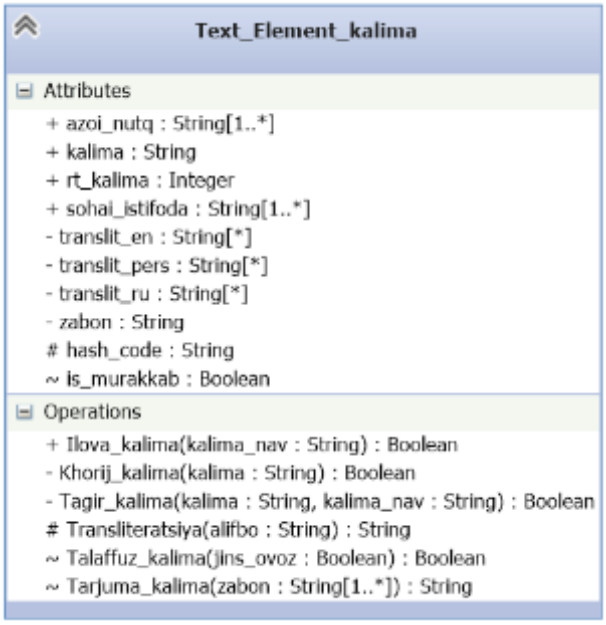

Рисунок 3.10. - Вид свойств класса слов

Public (общий, общедоступный) – это значение указывает на то, что значение атрибута доступно другим классам. Каждый класс может просматривать или изменять значения атрибутов. В этом случае класс Lingustic_Corpus может изменить значение атрибута text_Element_Name класса text_Element_Name. Согласно нотации UML атрибут идентифицируется знаком «+».

Private (частное, секретное) – этот тип членства невидим и доступен не для всех классов. Класс Lingustic_Corpus знает значение свойства text_Element_Type и может его изменить, но его класс не может его увидеть или изменить. При необходимости он может попросить класс Text_Element (текстовый компонент) отобразить или изменить значение этой функции. Скрытое свойство обозначается символом «-» в соответствии с нотацией UML.

Protected (защищенный) – данный тип свойств доступен для класса и его прототипов. Предположим, что существует два разных типа текстового элемента с

именем и типом. Возьмем еще два класса: Hijo_class и Kalima_class, которые являются генерацией класса Text_Element. Мы можем увидеть или изменить защитные свойства text_Element_Type из класса Text_Element, но из класса Lingustic_Corpus это действие невозможно. Символом UML для защитных отношений является символ «#».

Package or Implementation (пакет, группа) – свойства этого типа являются общедоступными и доступны только в отдельном пакете. Допустим, что атрибут text_Element_Name содержит существующий пакет. При этом, если он есть в классе этого пакета, он будет изменен в классе Lingustic_Corpus. Этот рабочий процесс обозначается символом «~».

Во всех случаях классовые свойства должны быть частными и защищенными. Этот вопрос контролируется самими свойствами и текстом программы. Путем приватизации или защиты можно выйти из ситуации изменения значения характеристик классов информационной системы. В зависимости от визуального оформления значения атрибута определяются специальные условия обработки текста программы класса.

Действия выполняются во взаимосвязи с действием класса. Операция состоит из трех частей: имени, размера и типа возвращаемого значения. Измерения – это значения, полученные в результате операции при вводе. Тип возвращаемого значения принадлежит результату операции. Диаграмма классов показывает имя операции, ее измерения и тип возвращаемого значения.

В языке UML операции определяются следующим образом:

Название операции (аргумент-1: тип данных аргумента-1, аргумент-2: тип данных аргумента-2,…) тип возвращаемого значения.

При определении операции возможны четыре различных их типа.

Операция реализации выполняет некоторые базовые операции. Эту операцию можно обнаружить, проанализировав диаграмму взаимодействия. Каждое действие исполнителя должно с легкостью обнаруживаться до соответствующего требования. Это мы можем видеть в различных типах дизайна. Действие вводится из сообщений в диаграмму взаимодействия, сообщение выводится из реального

списка процессов. Это позволяет включать все требования в текст программы, и каждая часть текста программы содержит требования определенного типа.

Операция управления используется для создания и уничтожения объектов. К этому типу операций относятся классы «конструктор» и «деструктор».

Операции доступа обычно являются конфиденциальными и защитными. Независимо от этого, другие классы могут видеть или изменять его значение в особых случаях. Для этого разрабатываются доступные операции.

Например, атрибут text_Element_Type задан для класса Text_Element. Чтобы остальные классы могли изменить эту возможность, в класс Text_Element добавляем две доступные операции GetElementType и SetElementType.

В этом случае он просто принимает значение свойства ElementType и возвращает его запрашивающему классу. Операция SetElementType является универсальной и обеспечивает поддержку классов, которые ее запрашивают, а также сброс значения свойства ElementType. Эта операция содержит все правила и условия контроля, которые должны быть выполнены, а также может быть изменен внешний вид текстового элемента. Этот метод обеспечивает уникальные ограничения безопасности для свойств внутри класса, защищая их от других классов, но при этом обеспечивая контроль. По умолчанию для каждого атрибута класса создаются операции get и set (получение и изменение информации).

Вспомогательные операции – это операции класса, которые необходимы для выполнения ответственности, но другим классам не обязательно знать об операции. Это конфиденциальный и защищенный действие класса.

В случае моделирования статистической структуры информационной системы эффективны следующие шаги:

1. Проанализировав последовательную и кооперативную диаграмму, преобразовать большую часть сообщений этой диаграммы в исполняемые операции. Рефлексивные сообщения преобразовать в вспомогательные действия.

2. Добавить операцию управления для создания и изменения других классов в качестве «конструкторов» и «деструкторов».

3. Операция доступа для каждой функции классов должна быть создана и определена с помощью операций Get и Set.

4. Для получения доступа к возвращаемому значению операции, следует учитывать значение атрибута класса. Коммуникация включает в себя логические взаимодействия между классами. Он предоставляет классам информацию о том, как получить доступ к свойствам, операциям и связям других классов. Чтобы класс мог отправить сообщение другому классу на основе диаграммы последовательности или взаимодействия, между ними должна быть связь.

При моделировании диаграмм классов существует четыре типа отношений для установления логических отношений между классами:

- ассоциация;
- зависимость;
- агрегация;
- реализация.

Ассоциация – это логическая связь между классами, на диаграмме классов они изображаются простыми линиями (рис. 3.11).

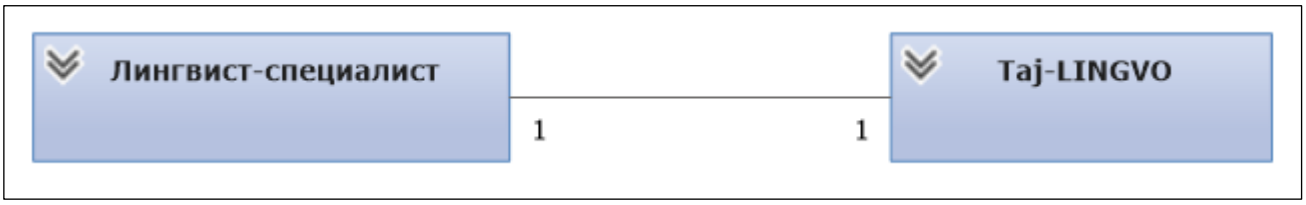

Рисунок 3.11. - Ассоциативные отношения в диаграмме классов UML

Объеденение между объектами может быть односторонним или двусторонним. В UML двунаправленная ассоциация представляется в виде простой линии без стрелок или линии со стрелками с обеих сторон. Для однонаправленного соединения отображается только однонаправленная стрелка.

Направление объединения определяется путем анализа последовательности и кооперативной схемы. Если все сообщения отправляются ему от одного класса и принимаются другими классами, то между этими классами нет места для

односторонней связи. Связь является двусторонней, когда контрагенту отправляется сообщение. Рефлексивная ассоциация определяется таким же образом. В этом случае экземпляр класса взаимодействует с другим экземпляром того же класса.

Отношения *зависимости* также представляют отношения между классами и всегда указывают направление зависимости. В этом случае класс зависит от определения другого созданного класса. Зависимости представлены пунктирными линиями (рис. 3.12).

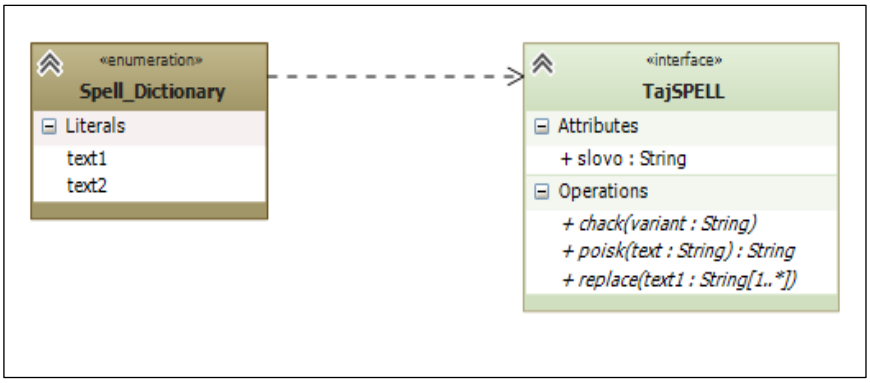

Рисунок 3.12. - Связь зависимостей в диаграмме классов UML

При составлении программных текстов классов в них добавляются новые свойства. В некоторых случаях для защиты связи разрабатывается специальный регулятор языка программирования. Например, в языке C++ в текст программы вводится обязательный модификатор #include.

Агрегация – ограниченная форма объединения. Она представляет собой связь между всеми частями или частью понятия. Например, наряду с определением класса Lingustic_Corpus известен также класс Text_Element для определения других частей текстового корпуса. В результате объект, определенный на основе класса Text_Element, состоя из объекта класса Lingustic_Corpus, имеет два объекта текстового языка и т.д. Агрегаты обозначены ромбовидными линиями.

При организации отношений агрегирования в языке UML используется более крупная агрегация, т. е. композиция. Согласовать периодический жизненный цикл основного объекта информационной системы с другими вспомогательными

объектами можно с помощью связи композиции. Эти две группы объектов работают вместе и в определенных случаях исчезают. Требуемое полное удаление затрагивает все части объекта (рис. 3.13).

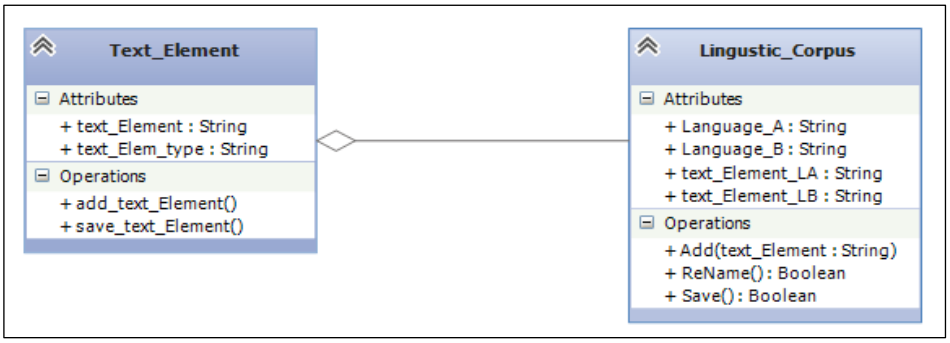

Рисунок 3.13. - Агрегационная связь в диаграмме классов UML

Такое условное удаление часто является частью определения агрегации, но только при наличии взаимно однозначной агрегации операций. Например, если удалить слог на основе класса Text_Element, действие должно повлиять на записанный звук слога в классе DB_Slog_Sound.

Объединение определяет отношения наследования между двумя классами. В большинстве случаев для обработки таких информационных систем используются методы объектно-ориентированного программирования. Класс может наследовать все свойства, операции и отношения другого класса. В UML отношения наследования называются агрегатами и представляются в виде стрелки от класса-потомка к классу-предку (рис. 3.14).

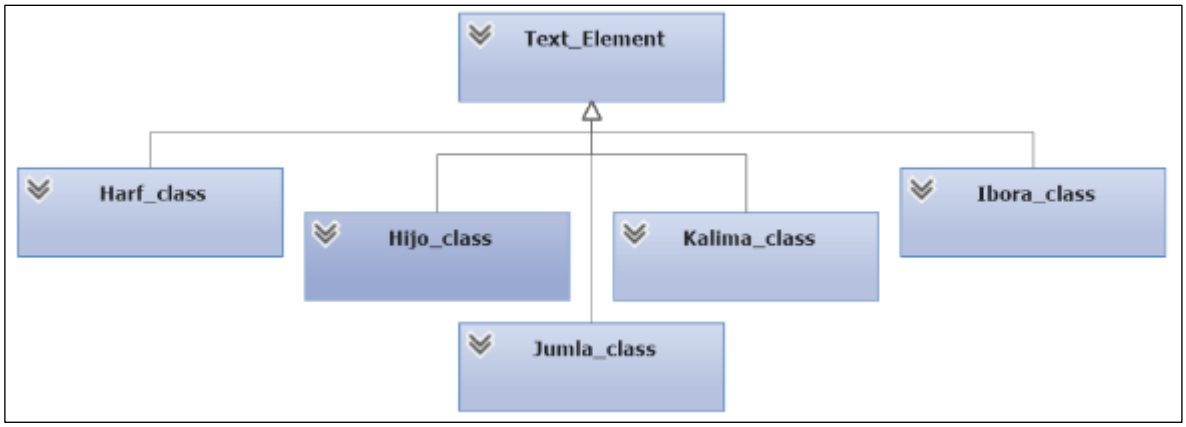

Рисунок 3.14. - Соединительные отношения на диаграмме классов UML

Классы-преемники, кроме свойств, действий и отношений класса-предка, могут быть определены со своим собственным набором свойств и операций.

*Множественность у*казывает на тип взаимодействия одного экземпляра класса с другим экземпляром класса, использующим это соединение.

Например, при обработке средства произношения слов в информационной системе необходимо определить классы DB_Slog_Zvuk (источник данных о звучании слога) и Text_Element_Hijo (текстовый элемент, слог). Между ними устанавливаются следующие отношения: каждый слог должен быть звонким и каждый голос должен представлять один слог. Эта процедура позволит ответить на следующие вопросы: «К какому слогу относится этот звук и из скольких записанных звуков он состоит?»

На приведенные выше вопросы отвечает множественное общение. Его индикаторы размещены на обоих концах линии связи. Запись слога установлено, что один голос может принадлежать одному и только одному слогу, но один слог может быть записан двумя лицами – мужским и женским голосами (рис. 3.15.).

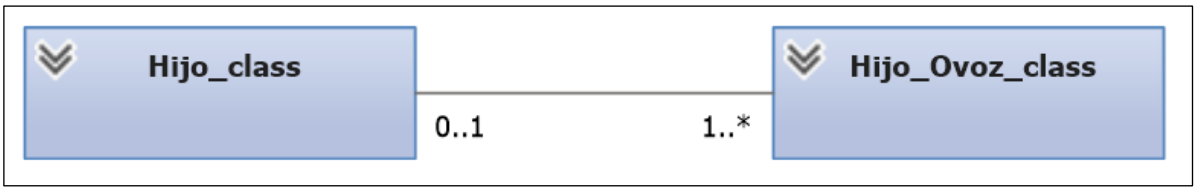

Рисунок 3.15. - Множественные отношения в диаграмме классов UML

В языке UML множественные отношения можно определить как:

- ноль и более (0..*);
- один или несколько (1..*);
- ноль или единица (0..1);
- сокращенная запись одна, только одна (1..1).

Название отношений – это глагол или форма глагола, определяющая цель, для которой оно нужно. Название связи используется для идентификации ее содержания и цели. Например, существует зависимость между классами Hijo_class и Hijo_Ovoz_class. Проблема произношения слога, основанная на объекте класса

DB_Slog_Zvuk, представляет собой звук слога или сам вновь идентифицированный слог. Для определения эту ассоциацию можно назвать «Выбор» (рис. 3.16).

Рисунок 3.16. - Название соединений в диаграмме классов UML

В случае моделирования статистической структуры информационной системы на диаграмме классов определение названий в связях не является обязательным. Можно указать имена контактов с соответствующей строкой контакта.

Название ролей. Вместо имен, для которых нужны эти отношения, пишутся имена ролей в отношениях ассоциации или агрегации. Например, глядя на классы DB_Slog_Zvuk и Text_Element_Hijo, можно сказать, что класс DB_Slog_Zvuk играет роль произношения класса Text_Element_Hijo. Название ролей – это существительные или специфичные для них фразы, которые появляются в таблице с классом. При моделировании структуры информационной системы на диаграмме классов рекомендуется использовать только название роли или название контакта. Что касается обработки информационной системы, то на первом месте стоит моделирование диаграммы классов. Это объясняется тем, что вопрос определения статистической структуры и логических связей между классами способствует общей организации объектов, функциональных возможностей и компонентов информационной системы (рис. 3.17.).

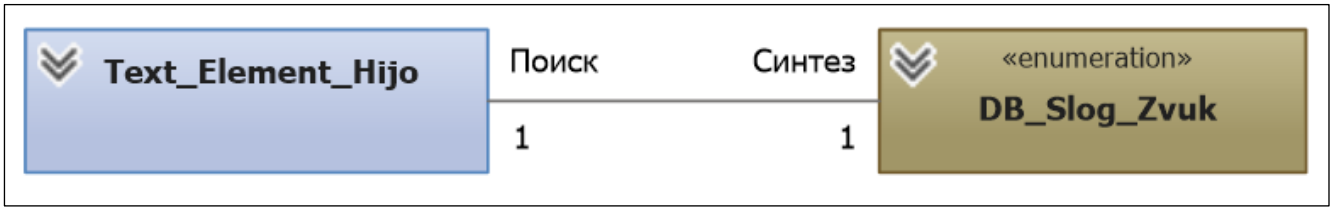

Рисунок 3.17. - Название ролей в диаграмме классов UML

В процессе моделирования статистической структуры информационной системы разрабатывается *диаграмма состояний* в зависимости от подверженности объектов различным ситуациям. Одним из наиболее распространенных типов диаграмм состояний – это использование объектно-ориентированных методов.

На рисунке 3.18. показан пример диаграммы состояний для текстового элемента. Из этой диаграммы можно понять, с какими ситуациями может столкнуться текстовый элемент. На этой диаграмме также можно наблюдать процесс перехода текстового элемента из одного состояния в другое.

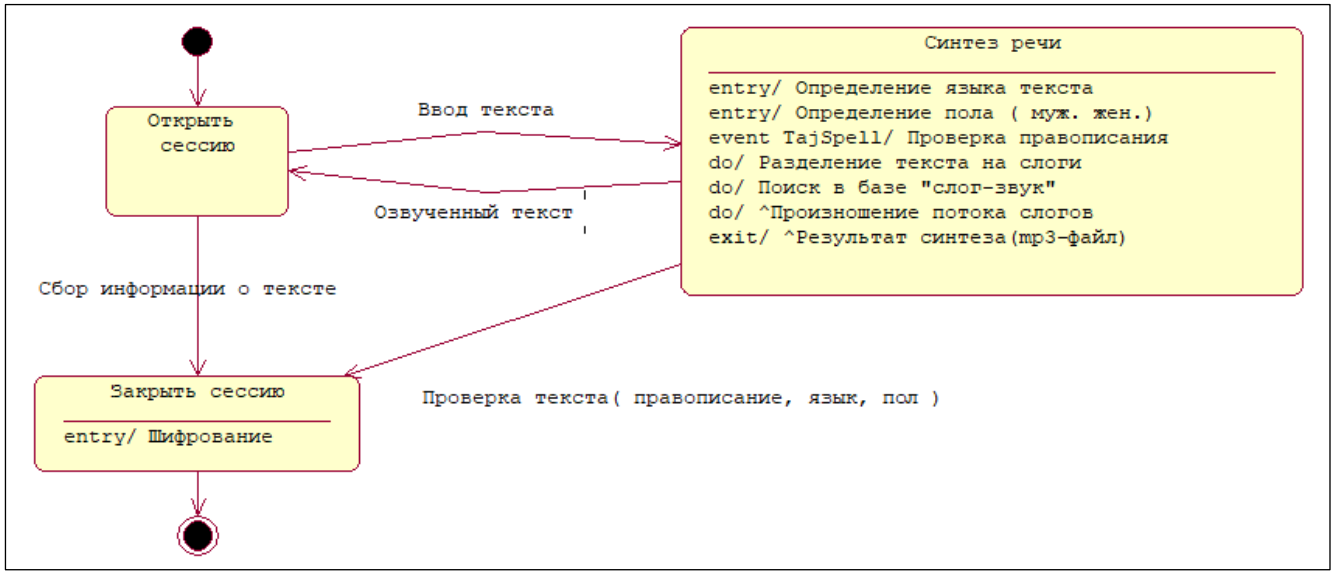

Рисунок 3.18. - Диаграмма состояний для класса Text_Element

Например, если информационная система требует произнесения текстового элемента, он перейдет из состояния «Закрытий», то есть зашифрованного состояния, в состояние «Открытий». В этом случае требование информационной системы называется событием (event), и только такое событие обеспечивает переход объекта из одного состояния в другое.

Если информационная система произносит введенный текстовый элемент, то она переходит в режим «Произношение текста». Такая ситуация наблюдается только в случае встречи звуков слогов, отделенных от текстового элемента в источнике данных «слог-звук». На диаграмме состояний определенное состояние управляет переходом объекта из одного состояния в другое.

На графике есть два основных состояния: старт и стоп. Исходное состояние отмечено черной точкой и соответствует состоянию объекта на момент его создания. Конечное состояние обозначается черной точкой в белом круге, который непосредственно содержит состояние завершения. Диаграмма состояний может иметь только один объект ресурса. В зависимости от направления воздействия объекта на различные ситуации можно определить одно или несколько конечных состояний.

Если объект находится в каком-либо реальном состоянии, могут осуществляться различные процессы. В примере на рисунке 3.18., произнесение текстового элемента отправляет соответствующие сообщения в информационную систему. Эти запущенные процессы на диаграмме состояний называются действиями. В процессе объектно-ориентированного моделирования диаграмма состояний может быть связана с пятью типами данных: *деятельностью, операцией ввода, операцией вывода, событием и историей состояний.* Рассмотрим каждые из этих данных на диаграмме состояний класса Text_Element информационной системы TajLINGVO.

Активность означает работу объекта в конкретном случае. Например, если текстовый элемент находится в состоянии «Закрытий», в информационную систему возвращается зашифрованная версия текста. В процессе моделирования действие рассматривается как остановленное и выполняется до завершения. Только если объект существует в этом состоянии, активность продолжится. В случае перехода объекта в другое состояние деятельность прекращается.

Входное действие – это метод указанного действия объекта, который входит в него сам. В примере произнесения текстового элемента при его переключении в режим «Произношение текста» выполняется операция «Определить пол диктора». Независимо от того, произносится текст или нет, обязательно определяется голос говорящего. Отличие операции ввода от активности заключается в том, что обработка входных данных представлена как непрерываемая.

Действие выхода определяется как и действие входа. Это важная часть процесса восстановления. В примере произношения текстового элемента, когда

объект Text_Element выходит из состояния «Произношение текста», независимо от направления его движения, выполняется действие «Подтверждение произношения». Это часть такого перехода. На диаграмме состояний операции вывода столь же непрерывны, как и операции ввода.

Поведение объекта во время операций ввода и вывода по существу относится к отправке события другому объекту. Например, объект Text_Element (текстовый элемент) может отправлять событие объекту DB_Slog_Zvuk (источник данных слоговых звуков) как «поиск элемента в источнике данных». Для более точного выявления события определяются цель, новость и ее характеристики.

Переход – это метод изменения объекта из одного состояния в другое. Серия переходов на диаграмме состояний (рис. 3.18) показывает, как объект переходит между своими состояниями. На схеме все переходы из начального состояния в конечное отмечены стрелками. Переходы также могут иметь рефлексивные свойства. Объект может вернуться в то состояние, в котором он находился. Рефлекторные переходы представлены стрелками, которые начинаются в одном и том же положении и заканчиваются в одном и том же положении.

Существуют различные классификации переходов: аргументы, случаи, условия окружающей среды и отправленные события. Событие вызывает переход из одного состояния в другое. В примере произношения текстового элемента событие расстановки переносов приводит к изменению состояния текстового элемента со скрытого на открытый. На диаграмме состояний для описания события можно использовать как названия действий, так и простые выражения. Для определения событий можно использовать аргументы. Например, событие «Разбить слово на слог», которое переключает произношение слога из состояния «Ввод» в состояние «Произнести», содержит аргумент Miqdor_hijo (количество) слога в произносимом слове.

В большинстве случаев переходы должны представлять собой событие, поскольку только они создают условия для реализации переходов. Существуют также автоматические переходы, не имеющие ни одного события. При этом сам

объект быстро переходит из одного состояния в другое, представляя собой выполнение операций ввода, операций и операций вывода.

Условия среды определяют, переход выполняется или не выполняется.

В примере произношения текстового элемента событие Syllabify Word изменяет текстовый элемент с «Разбить слово на слоги» на «Открыть» только в случае выполнения условия ввода для таджикских слов. В противном случае переход не состоится. Ограничительные условия на схеме обозначаются дефисом после названия события в круглых скобках. Ограничительные условия следует задавать только при задании автоматического перехода объекта из одного состояния в другое. Он определяет требования к правильному и логичному развитию диаграммы состояний и через какой переход она выполняется.

Действие – это непрерывная операция, которое выполняется в рамках перехода. Указывает операции входа и выхода внутри состояния и определяет, какое действие выполняется при входе или выходе объекта. В нем описывается большинство действий в строке перехода, поскольку их не нужно выполнять при входе в состояние или выходе из него.

Например, когда текстовый элемент переходит из открытого состояния в скрытое, выполняется процесс шифрования. Эта непрерывная активность происходит только при переключении из открытого режима в скрытый режим. Событие или действие может представлять внутреннее поведение объекта, либо объект может хранить сообщение, отправленное другим объектом. Прототипирование не требует создания диаграммы состояний для каждого класса и используется только при управлении сложными классами.

Диаграмма деятельности – это представление способа реализации ряда параллельных рабочих процессов в информационной системе. Диаграмма деятельности является поддержкой представления параллельных операций за счет моделирования процессов в двух и более рабочих процессах, т.е. это мощный инструмент как для моделирования рабочих процессов, так и для параллельного программирования. Основным недостатком схемы является то, что связь действия и предмета не четко видна и не распознаваема (рис. 3.19.).

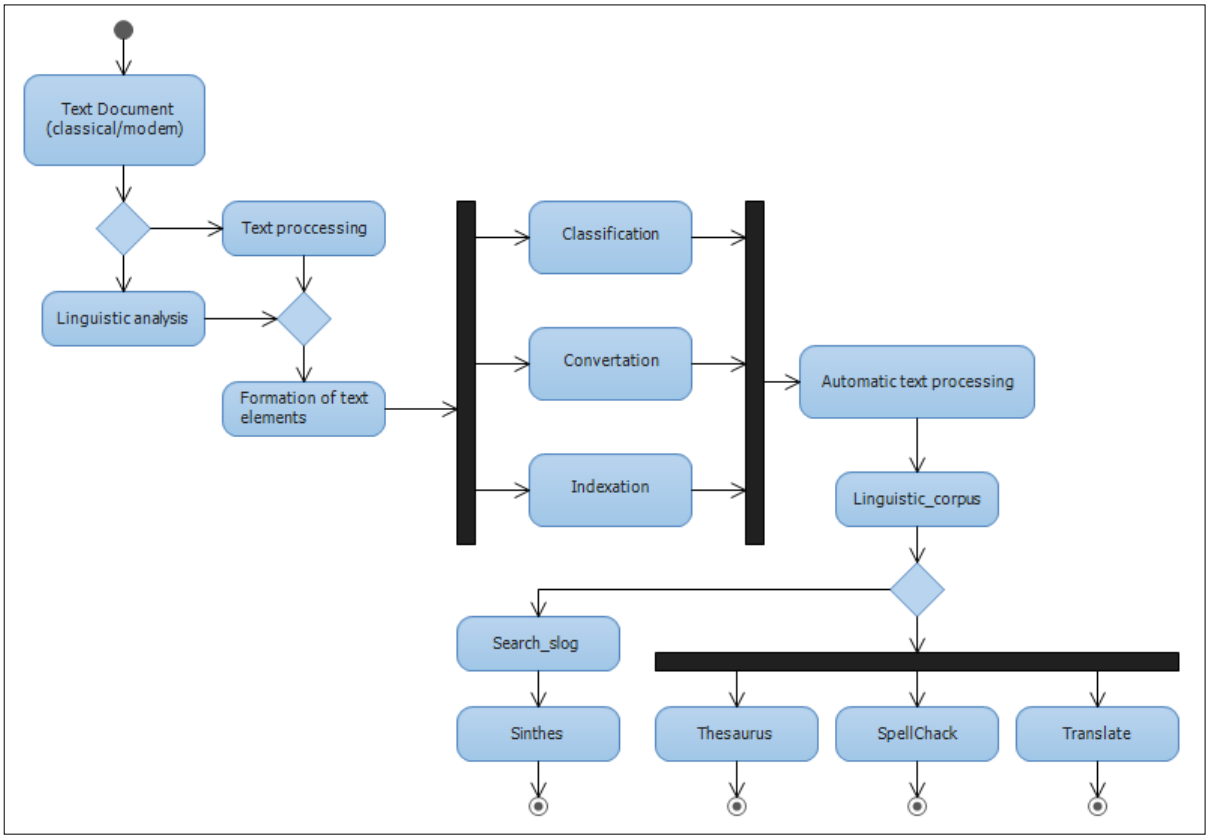

Рисунок 3.19. - Диаграмма деятельности в информационной системе TajLINGVO

Эта ссылка обозначается галочкой в названии объекта. Диаграмму деятельности можно использовать в двух ситуациях: при анализе вариантов использования и при анализа рабочего процесса в различных вариантах использования. Диаграмма деятельности считается одним из средств моделирования способа действия, представляющего выполнение рабочего процесса подобно алгоритмам.

В случае моделирования информационной системы с использованием графика деятельности способ работы информационной системы определяется на основе анализа процесса обмена данными и этапов обработки процессов управления. Представленная диаграмма определяет логический алгоритм жизненного цикла информационной системы, но не может отражать основные этапы описания алгоритма. Для конкретного объяснения каждого алгоритма таких процессов, как «проверка орфографии», «синтез речи», «машинный перевод» полная информация приводится в других главах исследовния.

§3.5. Моделирование физической модели информационной системы

Завершающим этапом моделирования информационной системы является определение физической части проекта, которая определяется в двух типах UML-диаграмм: компоненты и размещения.

Диаграмма компонентов разрабатывается в процессе моделирования реального уровня информационной системы. На графике показаны компоненты программного обеспечения и связи между ними. Для разработки диаграмм следует использовать несколько типов компонентов: рабочий файл, файлы включения, библиотека процедур и текстовый пакет компьютерных программ.

В процессе моделирования информационной системы каждый класс моделей формируется в тексте компьютерной программы. В процессе обработки они добавляются в диаграмму компонентов. На рисунке 3.20 показана диаграмма компонентов информационной системы TajLINGVO.

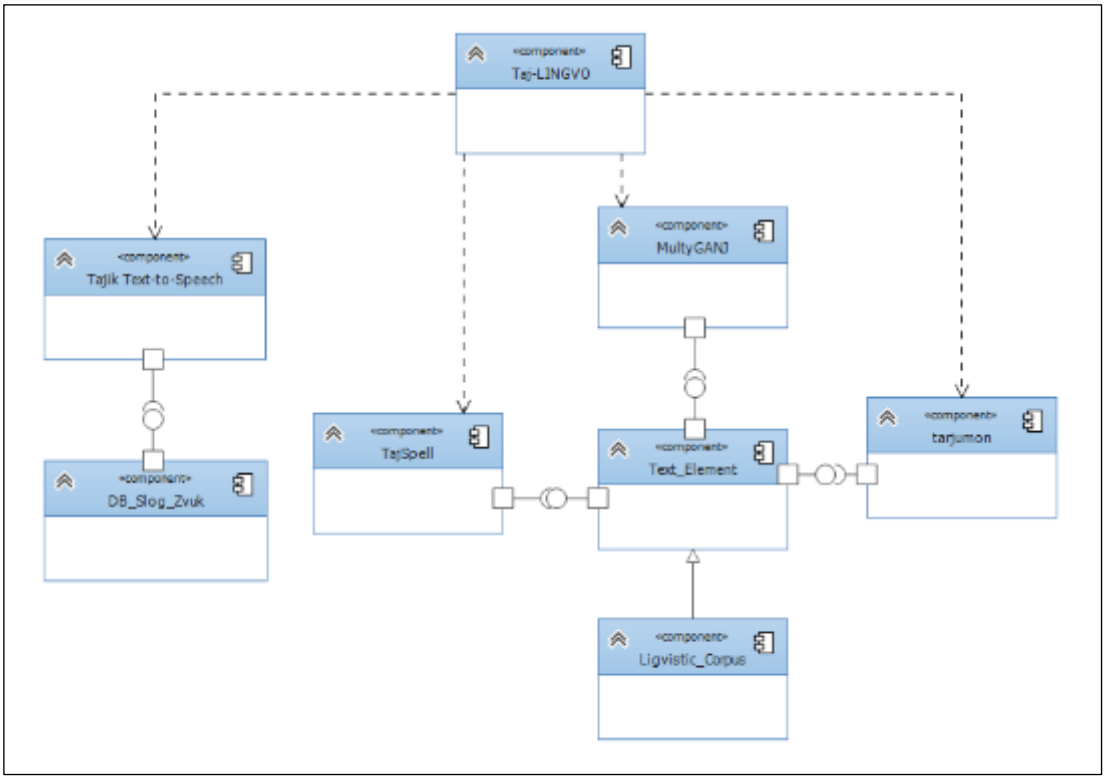

Рисунок 3.20. - Диаграмма компонентов информационной системы TajLINGVO

На диаграмме показаны компоненты информационной системы обработки текстовых данных. В данном случае требования к обработке частей программного обеспечения основаны на языке C++ и компьютерной среде обработки проектов MS Visual Studio .Net. Каждый класс будет иметь свой рабочий файл и файл рекомендаций. Каждый элемент размещается на диаграмме в зависимости от своего класса.

В процессе моделирования программного обеспечения класс Text_Element становится элементом Text_Element. Это также происходит во втором компоненте Lingustic_Corpus. Вместе эти два компонента представляют размер и заголовок класса TajLINGVO. Компоненты названы на основе классификации написанной на языке программирования класса TajLINGVO. Скрытый компонент соответствует рабочему файлу, написанному преимущественно на языке C++. Компонент TajLINGVO считаясь классификацией действий, показывает процесс обработки текстовых данных. В этом случае процесс обработки рассматривается как выполнение компьютерной программы.

Компоненты соединены пунктирной линией, которая указывает на их взаимосвязь. Например, класс Text_Element зависит от класса Lingustic_Corpus. Данная связь проявляется как зависимость текстового элемента от текстового корпуса таджикского языка. После обработки всех классов создается основной исполняемый файл информационной системы.

Информационная система TajLINGVO состоит из четырех процессов обработки текста: TajSPELL, Tajik Text-to-Speech, MultiGanj и Tarjumon. Первый компонент используется для автоматической проверки правописания текста, введенного на таджикском языке. Второй файл – компонент синтеза речи. Третий компонент MultiGanj предназначен для организации и обработки тезауруса таджикского языка. Четвертый элемент Tarjumon используется для активации процесса автоматического машинного перевода текста с таджикского языка на другой язык, например, русский или английский. Диаграмма компонентов сервера TajSpell показана на рисунке 3.21.

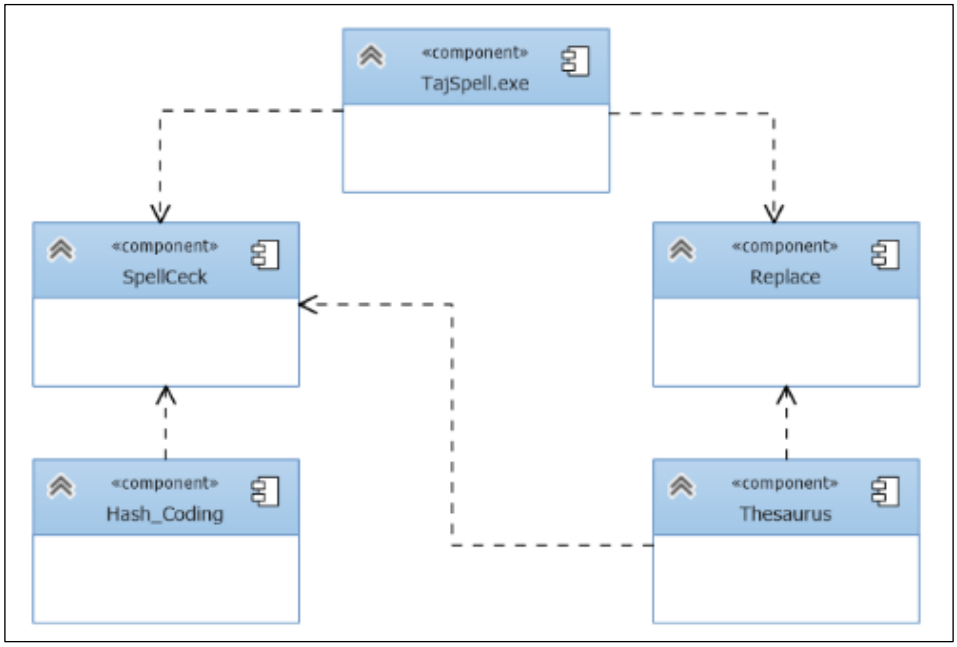

Рисунок 3.21. - Диаграмма компонентов сервера TajSpell

Диаграмма компонентов предусмотрена для тех участников проекта информационной системы, которые участвуют в процессе ее разработки и создания. Отсюда становится очевидным, в каком порядке обрабатываются компоненты и какие исполнительные компоненты создаются информационной системой. На рисунке 3.21 показаны зависимости диаграммы классов с рабочими компонентами.

Из рисунка видно, что диаграмма компонентов в информационной системе может быть в некоторой зависимости от количества компонентов и исполняемых файлов. Каждая деталь может состоять из набора отдельных компонентов. В общем, набор компьютерных программ – это набор специальных компонентов.

Диаграмма размещения. Моделирование фактического взаимодействия между программными и аппаратными компонентами входит в функцию диаграммы размещения. Это хороший инструмент, который контролирует интеграцию объектов и компонентов в информационную систему. Схема развертывания состоит из нескольких узлов, каждый из которых содержит специализированное оборудование, такое как процессор, простой вычислительный блок или сам хост-сервер.

Диаграмма размещения показывает моделирование физической взаимосвязи между программными и аппаратными компонентами. Это хороший инструмент для контроля интеграции объектов и компонентов в информационную систему. Схема развертывания состоит из ряда узлов, каждый из которых содержит специализированное оборудование, такое как процессор, простой вычислительный блок или хост-сервер. Схема компоновки показывает реальную структуру сети и расположение компонентов информационной системы. В примере информационной системы TajLINGVO использовано несколько частей, которые выполняются на отдельном реальном оборудовании, например сервере или узле. Схема развертывания информационной системы TajLINGVO представлена на рисунке 3.22.

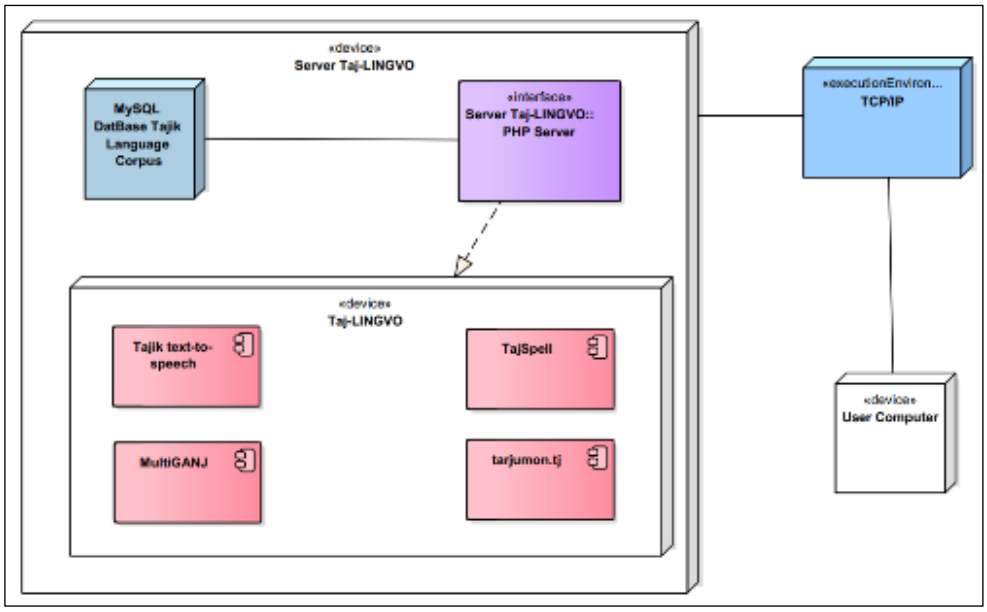

Рисунок 3.22. - Схема расположения информационной системы TajLINGVO

Из представленной диаграммы можно получить информацию о расположении реальных частей информационной системы. Программное обеспечение для автоматической обработки текста работает на нескольких разных веб-сайтах. В сетях интернет обработка данных осуществляется на индивидуальном сервере TajLINGVO. В свою очередь, в ограниченной сети информационная система взаимодействует с серверами источников данных,

которые работают под управлением MySQL. Наконец, компьютер пользователя с дополнительным оборудованием, таким как динамик, микрофон и принтер, подключается к основному серверу TajLINGVO. Диаграмма размещения может быть использована как инструмент управления проектом информационной системы программистами, разработчиками информационных систем и всем персоналом, так как она отображает фактическое расположение оборудования и отдельных его частей.

Выводы по третьей главе

Изучено моделирование информационных процессов и обработка информационных систем обработки текстовых данных на основе возможностей языка моделирования UML и CASE-инструментов, таких как MS Visual Studio .Net, IBM Rational Rose, Enterprise Architect.

Проанализирована проблема моделирования работы информационной системы с учетом варианта использования и пользователей процесса обработки текстовых данных. Также проанализировано моделирование взаимодействия объектов с учетом жизненного цикла и структуры диаграммы последовательности и взаимодействия. На основе построения диаграммы классов проведено исследование статической структуры информационной системы. Решена задача выявления общих объектов информационной системы и логической связи между ними. Определены особенности возможностей связи между объектами информационной системы с целью опережения обработки текстовых данных. Кроме того, исследованы возможные случаи появления текстового элемента в процессе обработки. Фактическая конструкция информационной системы была определена с целью обобщения практических возможностей программного обеспечения и оборудования. Установлена связь между статической структурой, компонентами и узлами технических средств при реализации информационной системы. Разработан UML-проект системы обработки текстовых данных с учетом модели поведения, взаимодействия, структуры и физических аппаратных средств.

ГЛАВА 4. ПРОЕКТИРОВАНИЕ, РАЗРАБОТКА И ВНЕДРЕНИЕ АВТОМАТИЧЕСКОЙ ПРОВЕРКИ ПРАВОПИСАНИЯ ТАДЖИКСКОГО ЯЗЫКА

§4.1. Проектирование и разработка электронных словарей

С появлением компьютеров и глобальной сети интернета возникла проблема обработки и использования электронных словарей. Существуют разные формы электронных словарей, такие как windows-приложение, мобильное приложение и веб-приложение, но каждое из них обрабатывается уникальными методами в зависимости от предоставления информации по определенной теме. В данном разделе рассматриваются вопросы подготовки проекта электронного словаря и разработки систем автоматического перевода на таджикский язык для развития изучения таджикского компьютерного языка, проблемы разработки программного проекта на базе CASE-технологий и этапы разработки электронного словаря.

Изучение одного языка общения на основе другого языка с появлением использования возможностей электронных словарей дало основу для его развития. Вопросы обработки и использования электронного словаря рассматриваются автором на основе традиционных методов изучения языка и их сочетания с возможностями информационных технологий. На основе новых алгоритмов и методов обработки текстовых данных и хранения информации стали возможны современные инструменты реализации проектов с возможностью перевода, в частности, эффективная функция поиска таких лексических единиц, как ключевые слова в больших текстах и их реализация в электронный словарь. Словарный источник данных обеспечивает возможность морфологического анализа слов, классификации, семантики и теории автоматического перевода, что решает одну из задач компьютерной лингвистики.

Современные компьютерные технологии значительно упростили не только процесс анализа текстовых элементов, но и обработку программного обеспечения электронного словаря. При обработке электронного словаря учитывается

несколько важных особенностей: возможность поиска подходящего текста, сортировка результата, группировка его по определенным символам и обработка потенциально неограниченного объема информации. Следует отметить, что в настоящее время растет спрос на электронные словари с названием «активный тип». Электронный словарь нужен не только переводчикам, но и пользователям компьютерных технологий, желающим выучить иностранный язык.

В настоящее время во всемирной сети международных компаний доступно большое количество электронных словарей, как бесплатных, так и платных. Такие как: ABBYY Lingvo, Multitran, PROMT, Cambridge Dictionary, Merriam-Webster, Longman Dictionary, Translate Google, Macmillan Dictionary, Dictionary.com, Oxford Dictionary, Oxford Learner’s Dictionaries, Linguee, Dict.com, Gramota.ru, dic.academic.ru, ruscorpora.ru, Reverso, BRKS.info, Zhonga.ru.

На основе анализа и использования перечисленных электронных словарей установлено, что не все из них поддерживают таджикский язык. По этой причине возникает необходимость спроектировать и разработать электронный словарь на таджикском языке. В зависимости от результатов теоретических исследований обсуждается процесс проектирования электронных словарей и этапы его разработки.

Виды электронного словаря. Электронные словари по функциям делятся на два типа, классификация их частей представлена на рисунке 4.1.

Первый тип словарей для изучения языка общения, то есть они создаются из набора слов на одном или нескольких языках с объяснением, описанием и определением каждого слова, с предложением использовать их в словосочетании или в предложении. В свою очередь, существуют два подтипа общих и специальных словарей, которые отражают цель и содержание словаря. Каждый тип электронного словаря решает задачу изучения языка в зависимости от поставленных задач, то есть представляет интерпретацию, перевод, структуру, тип и этимологию каждого слова. Следует отметить, что двуязычные или многоязычные словари создаются для организации и развития процесса перевода в

делопроизводства. Также предлагаются специальные словари для отдельного изучения фразеологии или частей речи языка в зависимости от выбранного слова.

Рисунок 4.1. - Классификация электронных словарей

Второй тип – создание списка слов внутри языка общения с пояснением, конкретным определением, историей их происхождения и сферой употребления в виде языковой культуры. В зависимости от области обучения электронные словари создаются и используются в виде специальных электронных книг или специализированных электронных книг.

*Проектирование логической структуры словаря***.** Важным шагом в разработке программного обеспечения является разработка соответствующего подхода и структуры программного обеспечения. Поскольку предметом исследования является основной фактор при создании электронных словарей, необходимо учитывать логическую структуру доставки программного обеспечения: веб-приложение, мобильную или десктопную версию. Каждая из перечисленных структур имеет как преимущества, так и недостатки и осуществляет соответствующую обработку данных в своей рабочей среде.

Разработка программного обеспечения состоит из анализа требований и этапа предварительной диагностики. Содержание проекта обычно состоит из двух

частей: поведение программного обеспечения; логическая структура. CASE-средства часто используются для обеспечения успешной реализации проектов компьютерного программного обеспечения. Обычно описывается поведение программного обеспечения, то есть анализ требований к диаграммам варианта использования и деятельности.

Чтобы объяснить возможные варианты использования и направления действий пользователей, была разработана диаграмма вариантов использования. Взаимодействия между группами объектов определяют процесс реализации и описание различных вариантов использования. Кроме того, на диаграмме также определяются альтернативные ситуации, т.е. виды из указанной последовательности событий.

Возможности пользователя электронного словаря. Практические возможности, выявленные на основе анализа поведения и структуры проекта развертывания пользовательской и системы электронного словарья представлены на рисунке 4.2.

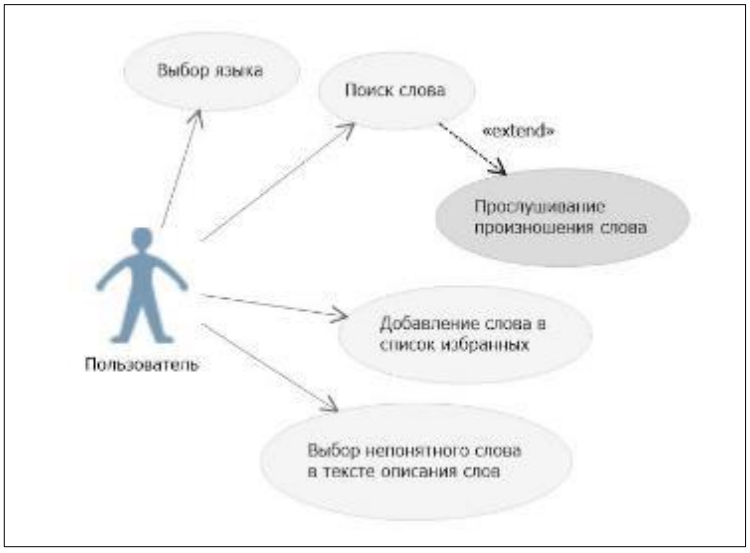

Рисунок 4.2. - Диаграмма вариантов использования электронного словаря

На основе разработанной диаграммы варианта использования определяем возможности пользователей:

- выбор языка словаря;

- поиск слова;

- прослушивание произношения слова;

- возможность нажать на непонятное слово в дефиниции слова;

- добавление слова в свой любимый словарь.

Особенности программы «Электронный словарь». Диаграммы классов создаются на основе анализа объектов, определения их описания и подготовки логической структуры программного обеспечения. После анализа требований и условий работодателя проекта перед регуляторами встает проблема подготовки концептуального представления о поставке программного обеспечения. Минимальными возможности программного обеспечения электронного словаря могут быть следующими:

- специальные правила правильного произношения слова;

- предъявление имени участника соответствующей речи;

- представление истории и происхождения слова;

- представление содержания слова;

- тезаурус (синонимы, антонимы);

- представление отраслевы/профессиональнх терминов, понятий;

- примеры употребления слов во фразах и предложениях;

- аудио и видеопроизношение слова;

- доступ к картинке или фото в виде словесного изображения;

- представление перевода слов на несколько языков;

- объяснение слова примерами из статей, книг и т.п.;

- примеры разных форм одного и того же слова.

Диаграмма классов представляет концептуальный дизайн проекта, представляя понятия, свойства и возможные отношения между ними. На рисунке 4.3. показана диаграмма классов проекта электронного словаря, которая иллюстрирует логическую структуру программного обеспечения.

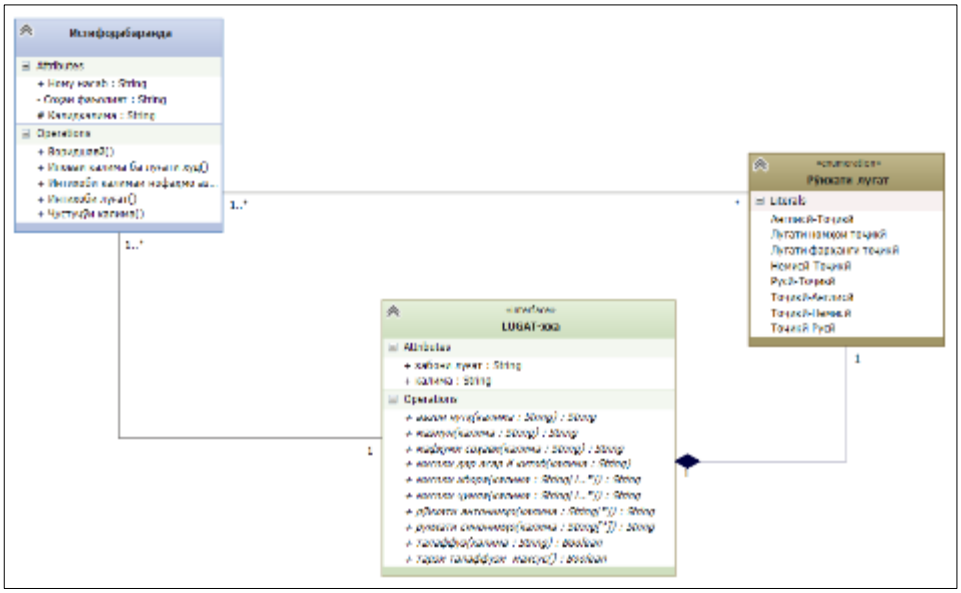

Рисунок 4.3. - Логическая структура программы электронного словаря

Обработка источника данных. Источник данных – это набор информаций или данных, собранных из определенной структуры, имеющие логическую связь друг с другом и выражающие одну сложную тему. Для обработки источника данных необходимо выполнить следующие шаги:

- собрать данные;
- идентифицировать пользователей;
- идентифицировать предметы и их свойства;
- определить логические связи между объектами;
- исключить возможности дублирования данных;
- обработать таблицы;
- определить связи между таблицами.

В настоящее время реляционные структуры данных эффективно используются для организации источников данных и управления ими. Реляционная структура данных – это способ просмотра данных через таблицы и возможные операции над ними. То есть должны быть решены три основных вопроса – определение структуры таблицы, целостность данных в таблице и управление данными.

*Правило***.** Отношение R выражается на основе множества областей D1, D2, ..., Dn и результата произведения десятикратного умножения подмножества этих областей, т.е. R ⊆ D1 × D2 × ... × *D*n.

Пусть для источника данных электронного словаря определены следующие домены: D1 – набор слов, D2 – набор фраз, D3 – набор предложений, D6 – набор частей речи, D5 –набор сфер употребления слов, *D*4 – перевод слова на другой язык.

Отношение R1 - для слов и области их употребления определяется порядком R1 ⊆ D1 × D4.

Отношение R2 – для слов, какой части речи соответствует слово, словосочетаний с использованием этих слов и перевода слова на другой язык определяется порядком R2 ⊆ D1 × D4 × D2 × D6.

Выбор данных из отношения R осуществляется при наличии строк, удовлетворяющих условию P. То есть операция отбора – эта единая операция, она выполняется таким образом, чтобы соответствовала логическому условию P.

$$R = \sigma_{\text{p}}(R_1)\ DataBase$$

Например, для выбора перевода слова « зардолу » с таджикского языка на русский в источнике данных $LUGAT$ выполняется следующее реляционное действие:

$$\sigma_{\text{самт="тоҷикӣ−русӣ"}}(\sigma_{\text{калима='зардолу'}})LUGAT$$

На основе анализа сферы электронных словарей была получена следующая группа данных, которая зависит от предпочтений автора программы: слово, способ произнесения (транскрипт), направление перевода, язык, история появления, сфера употребления, часть речи, голосовой файл с произношением, картинка (при необходимости), видеофайл с произношением, словосочетания и предложения с использованием этого слова, смысловое значение слова. На рисунке 4.4. показана логическая структура источника данных электронного словаря, основанная на реляционной структуре данных.

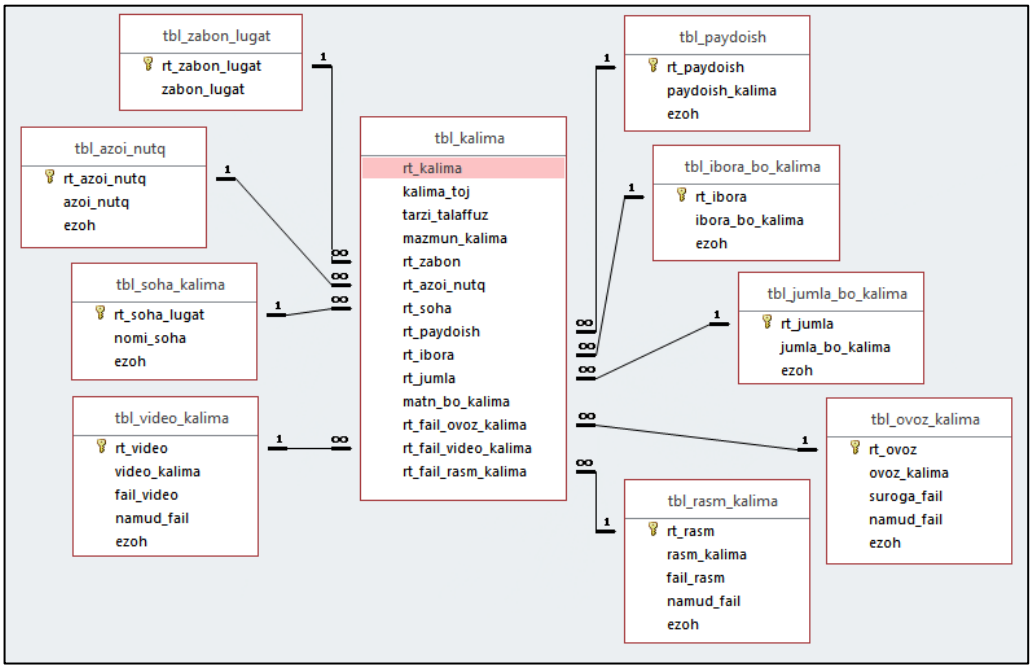

Рисунок 4.4. - Логическая структура источника данных электронного словаря

Из всех возможных операций с данными широко используется операция выбора. Характеристики выбора определяются в зависимости от требований или состояния набора данных. Критерий выбора – это определенный критерий выбора данных, при котором используются логическое содержание данных, соответствующее значение и логическая связь между данными. Язык запросов SQL используется для исполнения реляционных операций в компьютерных программах обработки источников данных [40].

Например, для выбора перевода слова «зардолу» с таджикского языка на русский в источнике данных *LUGAT* формируется SQL-запрос следующим образом (см. SELECT lugat.*

FROM lugat.tbl_kalima, lugat.tbl_zabon_lugat

WHERE lugat.tbl_kalima.rt_zabon = lugat.tbl_zabon_lugat.rt_zabon AND

lugat.tbl_zabon_lugat. zabon_lugat = “тоҷикӣ-русӣ” AND

lugat.tbl_kalima = “зардолу”

Разработка условий пользователя. Функциональность, внешний вид и удобство использования – это особенности программного обеспечения, которые затрагивает система разработки условий пользователя. Главные требования пользователя – это удобный функционал программы. Внешний вид рабочей среды непременно разрабатывается с учетом функциональных возможностей программы с использованием взаимно совместимых цветов и элементов управления. Независимо от личных потребностей пользователя, рабочая среда должна обеспечивать все цели использования программного обеспечения. Вышеперечисленные особенности обеспечивают эффективность, а также надежность и перспективность использования программы [22; 32].

В зависимости от возможностей рабочую среду электронного словаря можно разделить на две группы: настольной программы (рис. 4.5.) и веб-приложения (рис. 4.6.).

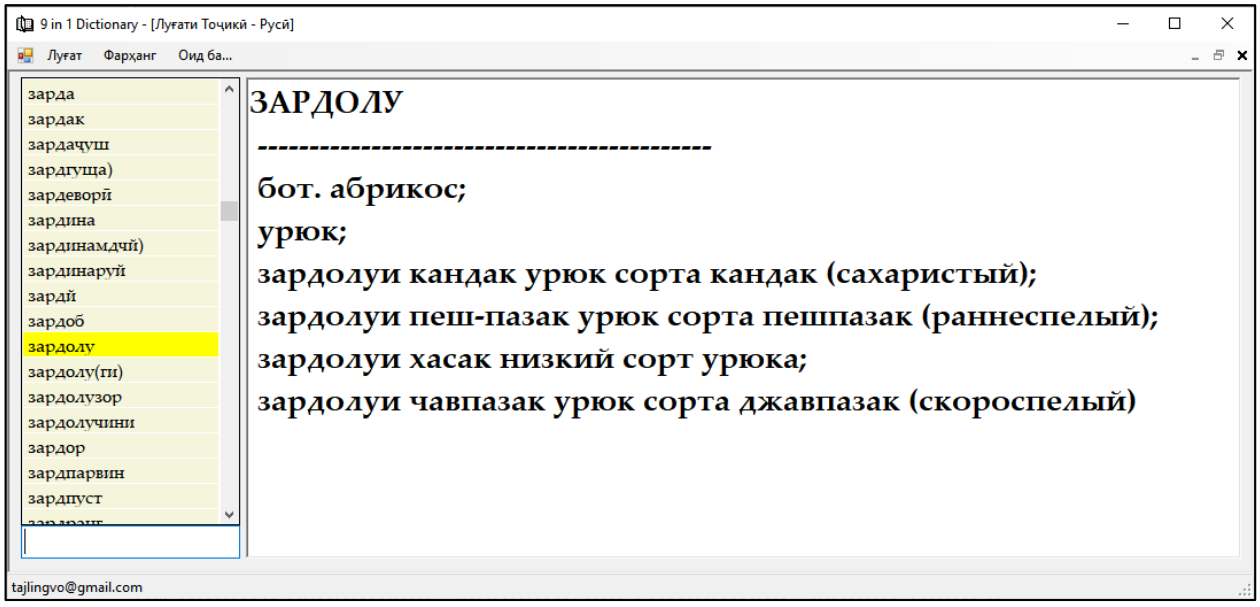

Рисунок 4.5. - Электронный словарь в виде настольной программы

На основе практических исследований разработана программа электронного словаря в виде настольной программы. Программа находится в свободном доступе в интернете.

Преимущества электронного словаря в виде настольной программы:

- работает в автономном режиме;

- база данных хранится в оборудовании пользователя;

- безопасность программы гарантируется ее разработчиками;

- работает без использования интернета, что сокращает расходы.

Недостатки электронного словаря в виде настольной программы:

- невозможность улучшения внешнего вида программы;

-необходимость скачивания и установления ее заново при появлении нового типа программы;

-использование постоянной памяти оборудования пользователя, что подразумевает возможность замедления его работы;

- появление конкуренции с электронными словарями типа веб-приложения.

На рисунке 4.6 показан внешний вид рабочей среды электронного словаря в виде веб-приложения.

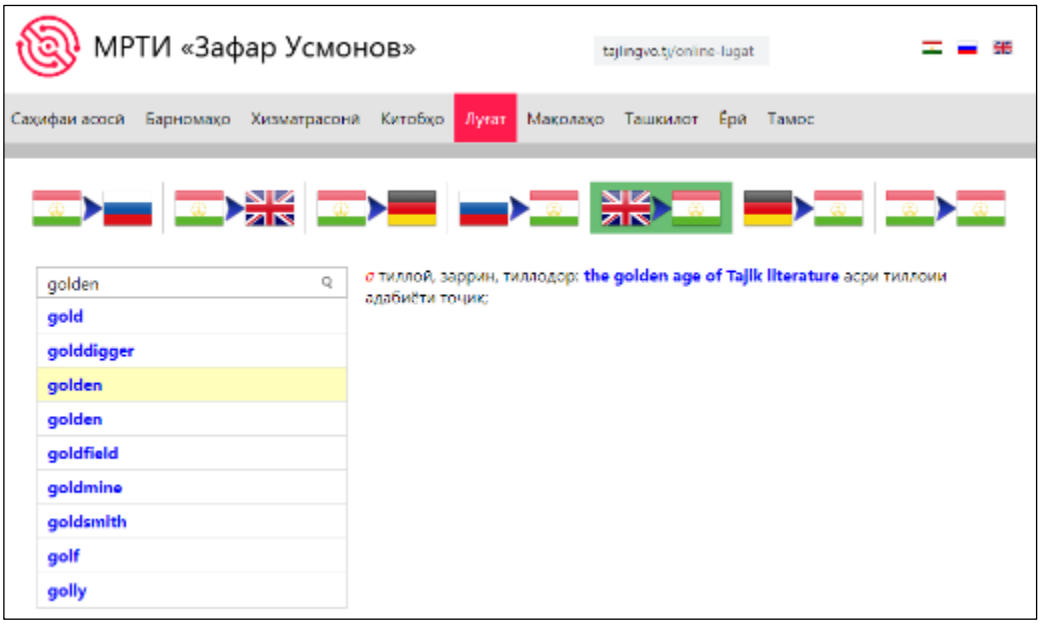

Рисунок 4.6 - Электронный словарь с веб-приложением

Преимущества электронного словаря перед веб-приложениями, заключаются в следующем:

- доступность программы электронного словаря каждому пользователю с веб-сервера в любое время;

- сохранность на сервере и доступность по сети источник данных словаря;

- пользователь не зависит от типа операционной системы;

- веб-приложение не требует дополнительной памяти на любом типе оборудования.

Недостатками электронного словаря в виде веб-приложения являются:

- обязателный доступ пользователя к сети интернет или локальной сети;

- невозможность использования словаря в автономном режиме без сети;

- ограниченные возможности у незарегистрированных пользователей;

- большой объем постоянной памяти на сервере.

В рамках научных исследований можно отметить, что разработка систем автоматического перевода текстов с таджикского языка пока не достигла уровня востребованности. Основная причина этого – недостаточное развитие современной компьютерной лингвистики. Однако следует отметить, что направления по изучению и подготовке подобных проектов в настоящее время рассматриваются.

По крайней мере, исходя из нынешнего уровня развития компьютерной индустрии и использования программ электронных словарей, перевод текста в ближайшем будущем станет более успешным. Итак, когда мы говорим о лучших способах разработки систем автоматического перевода, мы должны в первую очередь обратить внимание на разработку различных типов электронных словарей.

§4.2. Разработка компьютерного тезауруса таджикского языка

Тезаурус (от греческого языка – хранилище), как общее понятие – специальный термин, точнее, это словарь, совокупность информации, ресурс или сборник, включающий понятия и термины особой области или сферы деятельности в целях установления лексической или корпоративной связи. В языкознании под тезаурусом понимают словарь особого типа, в котором проявляются смысловые отношения (синонимы, антонимы, паронимы, гипонимы, гиперонимы и т.д.) между лексическими единицами. Тезаурус является основным инструментом анализа предметных областей.

В отличие от обычного словаря тезаурус дает возможность определить значение слова не только с помощью понятий, но и путем соотнесения слова с другими понятиями или их группой. По этой причине тезаурус следует использовать для пополнения смысловой базы системы искусственного интеллекта.

При использовании электронной документации, поиске информации, определении содержания текста и изучении языка тезаурус рассматривается как основной инструмент. Хотя стандарт тезауруса был принят в 2001 году (ГОСТ 7.25-2001), исследований в этой области, охватывающих таджикский язык, до 2016 года очень мало. В частности, созданы различные компьютерные словари, отражающие отдельные компоненты тезауруса.

Структура тезауруса Microsoft Word. В офисной программе Microsoft Word тезаурус используется с целью изучения языка и при необходимости использования тех или иных лексических единиц в зависимости от цели. Тезаурус Microsoft Word, как и упомянутый выше тезаурус, представляет собой словарь, в котором каждая лексическая единица объясняется отношением слова, понятия или группы понятий. В эту программу уже добавлен тезаурус для языков, которые использует для общения большинство людей (русский, немецкий, английский и т.д.). Для того, чтобы добавить в офисную программу тезаурус таджикского языка, следует подготовить словарь со специальной структурой и для каждой лексической единицы определить следующие свойства: pos - часть речи; idiom – устойчивое сочетание; antonyms – антоним; relatedword - связанное слово; relatedinfo – комментарий отношений; synonym – синоним, hyponym – гипоним; hypernym - гипероним; slang – отраслевое слово (сленг); colloquial – диалектное слово; archaic – устаревшее слово (архаизм); offensive – оскорбительное слово; alternate – альтернативное слово.

К сожалению, такого словаря на таджикском языке пока не существует. Независимо от этого, часть тезауруса можно составить, используя существующие словари.

Тезаурус WordNet – это электронный/сетевой (графовая структура) семантический тезаурус. Словарь состоит из 4 сеток (для 4 основных частей речи: существительного, глагола, прилагательного и наречия). Основной единицей словарного запаса в WordNet является не слово, а синсет, который объединяет слова одинакового содержания и сам является узлом сети.

Каждый синсет дополнен пояснением и примером употребления слова в контексте. Слово или словосочетание может встречаться в нескольких синсетах со сменой части речи.

Каждый синсет содержит список синонимов, синонимических оборотов или символов, объясняющих связь между ним и другим синсетом. Слова, имеющие несколько значений, входят в разные синсеты и могут принадлежать к разным синтаксическим и лексическим классам.

Синсеты в WordNet связаны с несколькими семантическими зависимостями:

- гипероним (фрукт → яблоко);
- гипоним (земля → планета);
- членство (факультет → профессор);
- член (профессор → факультет);
- мероним: (дерево → ветка);
- слово антоним (белое → черное).

Тезаурус – это база данных, в которой хранится информация о различных связях (отношениях) между словами (словарными статьями) и фразами (идиоматическими статьями).

Понятно, что технический прогресс невозможен без межкультурной коммуникации. Чтобы преодолеть словарный барьер, лингвисты составили множество словарей, но, к сожалению, немногие из них содержат необходимую информацию для перевода и понимания текстов, связанных с изучением языка. Кроме того, несмотря на все свои неоспоримые преимущества, печатные словари имеют и некоторые недостатки. Например, чем больше информации содержится в

словаре, тем ценнее тезаурус, тем богаче научный аппарат, но в то же время им труднее пользоваться.

Структура и состав тезауруса таджикского языка. В связи с тем, что тезаурус для таджикского языка готовится впервые, при определении дескрипторов следует действовать таким образом, чтобы тезаурус подходил как для офисной программы Microsoft Word, так и для WordNet. На основе изучения существующих словарей других языков был составлен перечень возможностей для создания тезауруса таджикского языка, как показано в таблице 4.1.

Таблица 4.1. - Основные возможности компьютерного тезауруса

№	Описание	Пример	Применение
1	Часть речи	Существительное, прилагательное	MS Word Thesaurus
2	Идиома	Рӯйи сурх	Тезаурус MS Word
3	Антоним	Сафед-сиёҳ	Тезаурус MS Word
4	Словосочетание (оборот)	Духтари зебо	Тезаурус MS Word
5	Комментарий связи	Духтаре, ки сирати зебо дорад	Тезаурус MS Word
6	Синоним	Офтоб, шамс, хуршед	Тезаурус MS Word
7	Гипоним	Замин сайёра	Тезаурус MS Word и WordNet
8	Гипероним	Мева-себ	Тезаурус MS Word и WordNet
9	Термин	Алгоритм	Тезаурус MS Word
10	Диалектизм	Шарик	Тезаурус MS Word
11	Архаизм	Миршаб	Тезаурус MS Word
12	Уничижительное слово	Аблаҳ	Тезаурус MS Word
13	Альтернативные слова	Бӯйи латиф, бӯйи хуш	Тезаурус MS Word
14	Членство (семантическое отношение)	факультет → профессор	WordNet
15	Членство (семантическое отношение)	профессор → факултет	WordNet
16	Мероним (семантическое отношение)	дарахт → шоха	WordNet
17	Смысл	Абармард-марди фавқулода бузург	Дополнительный
18	Перевод	Китоб-книга, book	Дополнительный
19	Постфикс	Коргар-ҳо,он	Дополнительный
20	Заимствование	Машина – англисӣ	Дополнительный
21	Транскрипция	Машина-mashina	Дополнительный

Лингвистический тезаурус таджикского языка. Создан интерактивный электронный словарь MultiGANJ на основе разработанного лингвистического тезауруса таджикского языка, состоящего более чем из 150000 лексических единиц.

В целях экономии времени пользователя возникла необходимость разработки электронного словаря, который бы имел достаточное количество словарных единиц для приема и перевода текстов на таджикский язык и в то же время имел удобное компьютерное приложение. Для решения данной проблемы была поставлена задача создать лингвистический тезаурус таджикского языка, а также интерактивный компьютерный словарь MultiGANJ на его основе. Одновременно в рамках исследовательского процесса был разработан большой интерактивный многоязычный электронный тезаурус.

Варианты тезауруса и словаря. Тезаурус таджикского языка создавался в течение нескольких лет с использованием местных и зарубежных словарей, статей в зарубежных изданиях и личного опыта общения со специалистами по таджикскому языку.

Разработанный тезаурус и словарь MultiGANJ предназначены для перевода слов с таджикского языка на русский и английский и наоборот. Он предназначен также для углубленного изучения таджикского, русского и английского языков.

В тезаурусе 65 000 таджикских слов, 70 000 русских слов и 35 000 английских слов для перевода. Кроме того, он содержит более 3500 синонимов, 1600 антонимов и 780 омонимов таджикских слов.

Основанный на стандартной графике и имеющий единый пользовательский интерфейс, программный продукт работает автономно. Тезаурус имеет ряд мощных функций, таких как «Поиск по шаблону», «Сканирование выбранных слов», «Нечеткий запрос».

Программа имеет понятный интерфейс, что позволяет выполнить перевод даже пользователю с самым низким уровнем владения компьютером.

Для корректного отображения таджикских символов используются шрифты, поддерживающие Unicode (рис. 4.7).

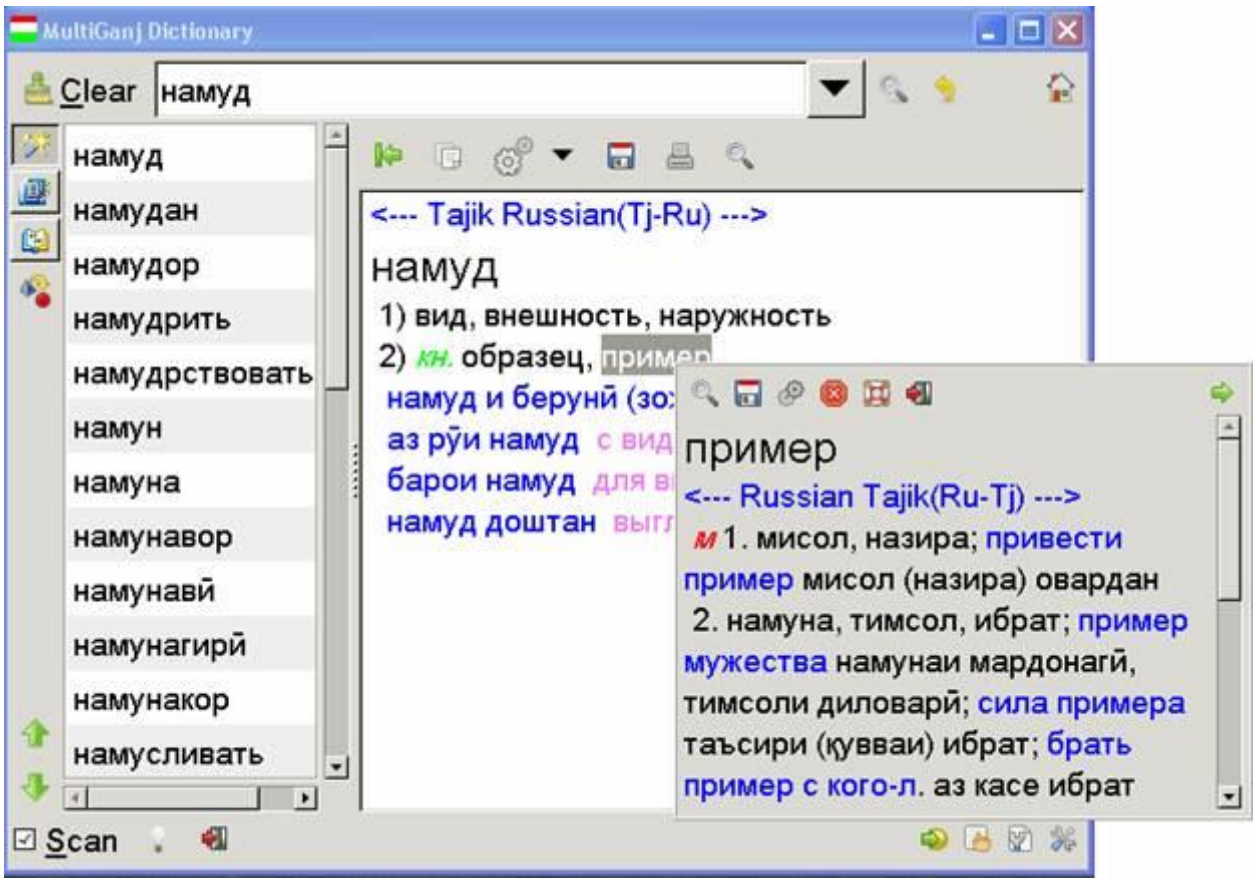

Рисунок 4.7. - Главное окно таджикского тезауруса - MultiGanj

Ведущие термины в словаре расположены строго в алфавитном порядке, что позволяет быстро найти необходимое слово или оборот. «Открывающиеся окна» дают пользователю уверенность в выборе правильного слова для выражения определенного технического понятия.

Словарь не требует доступа к интернету, этим он отличается от большинства словарей таджикского языка, его можно использовать в любое время и в любом месте без подключения к сети. Словарь прост в скачивании на мобильные устройства, удобен для использования профессионалами в автономном режиме непосредственно на работе.

Словарь программы можно пополнить двумя способами. Так, начинающие пользователи могут добавть новые слова, используя интерфейс программы.

Для каждой языковой единицы может быть предусмотрено одно слово или несколько вариантов перевода. Определение термина можно изменить или

дополнить в любой момент, поэтому на его основе очень легко обновлять тезаурус и словарь.

Когда вы выделяете в тексте слово, которое необходимо перевести, словарь автоматически предлагает все варианты, хранящиеся в памяти программы, по мере необходимости.

Поиск слов в словаре осуществляется с помощью командной строки. В то же время тезаурус «подсказывает» пользователю похожие слова по мере его ввода, что еще больше увеличивает скорость поиска и повышает интерактивность. Структура программной среды позволяет добавлять новые словари к существующим, что дает возможность создать единую многоязычную базу лингвистических терминов на одном уровне.

Область использования. Тезаурус и словарь могут быть полезны и интересны не только студентам, магистрантам, докторантам и начинающим переводчикам в области науки и техники, в том числе переводчикам иностранных источников, но и опытным переводчикам, поскольку ИКТ в настоящее время развиваются очень быстро, особенно новые непереводчики – традиционно связанные с появлением конструкций и их практическим применением.Тезаурус и словарь могут быть востребованы среди переводчиков, работающих с новейшими информационными технологиями, поскольку тезаурус и словарь позволяют существенно увеличить скорость перевода текстов на таджикский язык.

Необходимо организовать лингвистический тезаурус для решения проблемы использования качественного контента на таджикском языке и получения текстовой информации в сети интернет.

Следует отметить, что тезаурус, в отличие от толкового словаря, позволяет определять значение не только через определения, но и через связь слова с другими понятиями и их группами. Разработка лингвистического тезауруса таджикского языка поможет создать основу для создания сложных компьютерных программ автоматической проверки правописания текстов и улучшить качество таджикского языкового контента в интернете. В рамках разработки компьютерного тезауруса таджикского языка были решены ряд важных задач, а именно:

1. Нормы и структуры русских и английских тезаурусов изучались в компьютерной лингвистике.

2. Разработана структура компьютерного тезауруса таджикского языка.

3. Создан лингвистический тезаурус таджикского языка.

4. Разработаны программные модули для поиска текстовой информации на базе таджикского тезауруса.

5. Таджикский тезаурус создан как компьютерный проект MultiGANJ.

Научное и практическое применение разработанного тезауруса позволяет разработать новое программное обеспечение, например, для проверки грамматики таджикского языка, перевода текстовых данных с таджикского языка на другие языки, разработки поисковых систем на таджикском языке.

Достигнутые результаты послужат научному прогрессу в области компьютерной лингвистики таджикского языка и новым достижениям в этой области в будущем.

§4.3. Особенности автоматической системы проверки правописания на таджикском языке

Обзор текстовых элементов на таджикском языке. Таджикский язык получил мировое признание как язык древней культуры и многовековых литературных традиций. Несмотря на неоднократные нападения на территорию нашей страны со стороны иноземных захватчиков, таджикский язык сохранил свой грамматический строй и основную часть словарного запаса. Таджикский язык является национальным литературным и разговорным языком Республики Таджикистан. Таджикский язык широко распространен в некоторых регионах Средней Азии и Афганистане. Таджикский язык входит в иранскую группу индоевропейской языковой семьи.

Алфавит таджикского языка. С 1940 года в таджикском литературном языке начали использовать русский алфавит с добавлением шести специальных букв: «ғ», «ӣ», «қ», «ӯ», «ҳ», «ҷ». В 1998 году буквы «ц», «щ», «ь» и «ы» были исключены из

алфавита. Современный таджикский алфавит состоит из 35 букв, которые расположены в порядке, аналогичном русскому алфавиту. Рядом с русскими буквами находятся специальные таджикские символы. Таким образом, алфавит включает следующие буквы: Аа, Бб, Вв, Гг, Ғғ, Дд, Ее, Ёё, Жж, Зз, Ии, Йй, Ӣӣ, Кк, Ққ, Лл, Мм, Нн, Оо, Пп, Рр, Сс, Тт, Уу, Ӯӯ, Фф, Хх, Ҳҳ, Чч, Ҷҷ, Шш, Ъъ, Ээ, Юю, Яя.

В таблице 4.2 представлена кодовая страница строчных и прописных букв таджикского алфавита.

Таблица 4.2. - Кодовая страница UNICODE для букв таджикского алфавита

№	Символ	Символ Unicode	№	Символ	Символ Unicode	№	Символ	Символ Unicode
Маленькая буква			26.	Ф	\x0424	16.	л	\x043b
1.	А	\x0410	27.	Х	\x0425	17.	м	\x043c
2.	Б	\x0411	28.	Ҳ	\x04b2	18.	н	\x043d
3.	В	\x0412	29.	Ч	\x0427	19.	о	\x043e
4.	Г	\x0413	30.	Ҷ	\x04b6	20.	п	\x043f
5.	Ғ	\x0492	31.	Ш	\x0428	21.	р	\x0440
6.	Д	\x0414	32.	Ъ	\x042a	22.	с	\x0441
7.	Е	\x0415	33.	Э	\x042d	23.	т	\x0442
8.	Ё	\x0401	34.	Ю	\x042e	24.	у	\x0443
9.	Ж	\x0416	35.	Я	\x042f	25.	ӯ	\x04ef
10.	З	\x0417	*Болшая буква*			26.	ф	\x0444
11.	И	\x0418	1.	а	\x0430	27.	х	\x0445
12.	Ӣ	\x04e2	2.	б	\x0411	28.	ҳ	\x04b3
13.	Й	\x0419	3.	в	\x0432	29.	ч	\x0447
14.	К	\x041a	4.	г	\x0433	30.	ҷ	\x04b7
15.	Қ	\x049a	5.	ғ	\x0493	31.	ш	\x0448
16.	Л	\x041b	6.	д	\x0434	32.	ъ	\x044a
17.	М	\x041c	7.	е	\x0435	33.	э	\x044d
18.	Н	\x041d	8.	ё	\x0451	34.	ю	\x044e
19.	О	\x041e	9.	ж	\x0436	35.	я	\x044f
20.	П	\x041f	10.	з	\x0437			
21.	Р	\x0420	11.	и	\x0438			
22.	С	\x0421	12.	ӣ	\x04e3			
23.	Т	\x0422	13.	й	\x0439			
24.	У	\x0423	14.	к	\x043a			
25.	Ӯ	\x04ee	15.	қ	\x049b			

Предлагаемый алгоритм проверки правописания на таджикском языке должен соответствовать кодировке символов таджикского алфавита в стандарте

UNICODE. Поскольку основной процедурой алгоритма является обработка текстовой информации, главную роль играет таблица символов.

Текст с использованием специальных символов таджикского алфавита «ғ», «ӣ», «қ», «ӯ», «ҳ», «ҷ» на основе нестандартных шрифтов, таких как «Arial TJ», «Arial Tajik», «Tajikan», «Times New Roman TJ», «Times New Roman Tajik», «Academy Tajik» считаются некорректными текстами.

Для проверки таких текстов предлагается преобразовать коды в государственный стандарт UNICODE. Решение этой проблемы требует от всех пользователей компьютерных программ, работающих с таджикским текстом, перехода на стандарт UNICODE.

Нестандартный шрифт. В Таджикистане в 90-е годы прошлого века для ввода таджикского текста было разработано более 100 различных компьютерных реализаций шрифтов на основе таджикского алфавита. Следует отметить, что использование таких шрифтов в процесссе документирования приводят к значительной проблемее автоматической проверки правописания таджикского языка. Некоторые из них перечислены в таблице 4.3.

Таблица 4.3. - Список нестандартных шрифтов на компьютере

1	Academy Tajik	15	FreeSet Tojik	29	Tajik Jiharev Ejod
2	Arial Tj	16	Gothik Tojik	30	Tajik Souvenir Ejod
3	Adver Tojik	17	Impact Tojik	31	Tajik FuturaPress
4	Alterna Tojik	18	Journal Tojik	32	Tajikan
5	Antiqua Tojik	19	Rodeo Tojik	33	TajikTimesET
6	Arial Tj Bold	20	Tajik Baltic	34	Times New Roman Tajik
7	Arial Tj Italic	21	Tajik Ribbon	35	Times New Roman Tj
8	Arial Tj Bold Italic	22	Tajik InformC	36	Times Tojik
9	Arial Black Tojik	23	Tajik_Bengaly	37	Tadjik Normal
10	Book Man Tojik	24	Tajik_Art_Script	38	X Tajik Monaco Cyr
11	Courier New Tj	25	Taurus Tojik	39	X Tajik Times Cyr
12	Cooper Tojik	26	Taurus Tj	40	Vanta Tojik
13	Décor Tojik	27	Tajik Courier Ejod		
14	Diser Tojik	28	Tajik Helvetica Ejod		

В целях активного внедрения возможностей информационно-коммуникационных технологий с использованием таджикского языка и средств

ввода данных в компьютерные устройства предложен стандарт кодирования и компоновки таджикского алфавита по таблице UNICODE. В соответствии с Законами Республики Таджикистан «О государственном языке» [2] и «Об информации» [3] Постановлением Правительства Республики Таджикистан [10] от 2 августа 2004 г. утвержден стандарт № 330 для компьютерных приложений и для использования на территории Республики Таджикистан.

Об автоматическом преобразовании таджикского текста в стандартные шрифты. В данном разделе диссертации рассматривается проблема автоматического преобразования текстов на таджикском языке с нестандартной кодировкой в стандартную кодировку. Предлагается алгоритм автоматического изменения текста, на основе которого создается программный модуль поддержки офисных приложений.

Несмотря на то, что государственный стандарт таджикской компьютерной графики утвержден, все еще находятся пользователи, проявляющие индифферентность к работе по принятому стандарту. При этом драйвер раскладки таджикских букв для клавиатуры компьютера и руководство по его установке для использования в повседневной деятельности (с правильным объяснением положения на клавиатуре и в таблице символов шести специальных таджикских букв) остаются доступными через Интернет. Государственный стандарт таджикской компьютерной графики в стандарте UNICODE выглядит следующим образом.

Конвертация текстов с нестандартными шрифтами. В настоящее время существование больших нестандартных таджикских текстов стало серьезным препятствием для дальнейшего развития компьютерных технологий в Таджикистане. Возникающая проблема ощущается в расширении сферы использования уже разработанных программных комплексов, таких как автоматическое преобразование таджикских текстов в кириллической графике в персидские графические тексты, автоматическая проверка орфографии таджикских текстов, автоматический перевод с таджикского языка и т.д. Реальный путь решения этой проблемы – создание компьютеризированной системы

автоматического преобразования нестандартных текстов в тексты, поддерживаемые государственным стандартом. В процессе работы конкретные решения данной проблемы были предложены из отдельных проектов по обработки текстовых файлов. Однако проблема построения текста и структура исходных файлов до сих пор не решена.

В связи с необходимостью преодоления этих недостатков были разработаны алгоритм и программный модуль с поддержкой офисных программ, таких как Microsoft Office, OpenOffice, LibreOffice и др. Модуль расширяет возможности офисных программ и позволяет пользователю конвертировать весь текст целиком или с нужной позиции (рис. 4.8).

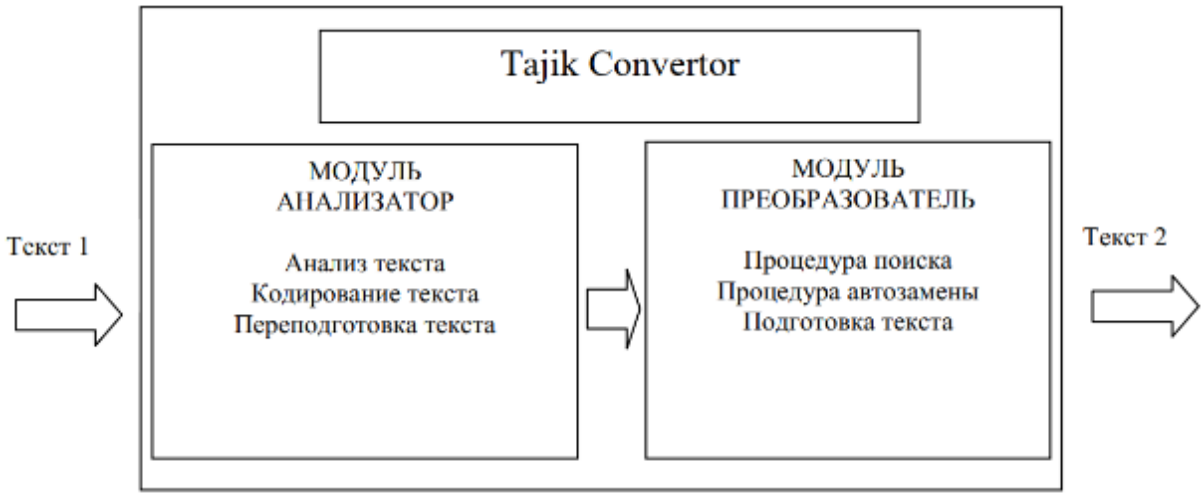

Рисунок 4.8. - Логическая структура проекта Tajik Convertor

Модуль преобразователя показан в виде концептуальной схемы на рис. Схема состоит из двух частей – «Анализатор» и «Модификатор». В первую часть входят следующие субмодули: «Анализ текста», «Кодирование текста», «Реконструкция текста». В этом разделе определены названия текстовых шрифтов, не соответствующих государственному стандарту Таджикистана. Результаты будут отправлены во вторую часть. В данный раздел входят следующие подмодули – «Режим поиска», «Режим автозамены» и «Подготовка текста», преобразующие символы с нестандартной кодировкой в стандартную форму.

Процесс преобразования зависит от шести специальных букв таджикского языка и соответствующей им кодировки разными шрифтами. В таблице 4.4 приведен перечень шрифтов с возможностью поддержки специальных букв таджикского языка, которые, согласно проведенному исследованию, чаще всего используются в деловой сфере.

Таблица 4.4. - Специальные таджикские буквы в нестандартных шрифтах

№	Стандартные шрифты Palatino Linotype	Ғ	ғ	Й	й	Қ	қ	Ҷ	ҷ	Ҳ	ҳ	Ӯ	ӯ
1	Academy Tajik	U	u	B	b	R	r	X	X	{	[	E	E
2	Arial TAJIK	U	u	B	B	R	r	X	X	{	[	E	E
3	Arial Tj	Ґ	ґ	Ї	ї	Ќ	ќ	Љ	Љ	Њ	њ	Ў	Ў
4	TAJIKAN	Щ	щ	Ц	ц	Ы	ы	Ж	Ж	Ь	ь	Ў	Ў
5	Times New Roman TAJIK	U	u	B	b	R	r	X	X	{	[	E	E
6	Times New Roman Tj	Ґ	ґ	Ї	ї	Ќ	ќ	Љ	Љ	Њ	њ	Ў	Ў

В таблице установлено соответствие между специальными символами стандартного шрифта Palatino Linotype и символами шести нестандартных шрифтов, используемых на практике. Именно на основе этой таблицы символы шести нестандартных шрифтов преобразуются в специальные буквы стандарта Palatino Linotype, соответствующие им по вертикали.

Алгоритм конвертации. Опишем порядок преобразования некоторых W слов, которые представляют собой определенную последовательность из n букв таджикского алфавита.

1. Вставьте счетчик букв p в слово W и установите p:= 0.

2. Пусть p := p + 1, что означает переход значения от буквы с номером p к следующей букве в слове W.

3. Проверьте, что если p> n, то переходите к b. 4. В противном случае сравните букву r в слове W на совместимость с символами шести букв нестандартного шрифта.

Если совпадений нет (это происходит, когда рассматриваемая буква является одной из 29 распространенных букв таджикской и русской графики), то

возвращайтесь к б. 2. Если совпадение есть, то замените этот символ соответствующей вертикальной буквой в таблице 3 шрифтом Palatino Linotype. Вернитесь к Б. 2.

4. Конец. Преобразование букв в слове W завершено.

Что касается 29 букв, общих для русского и таджикского алфавитов, то все решения были схожими в том смысле, что их положение оставалось неизменным как на клавиатуре компьютера, так и на кодовой странице. Ключевая разница между реализованными решениями просматривалась лишь в размещении конкретных таджикских букв.

Использование различных таджикских компьютерных шрифтов предприятиями, учреждениями и частными лицами привело к появлению больших, но ограниченных ресурсов текстовой информации, совершенно непригодных для приема, передачи и автоматической обработки информации во всех других местах, кроме точек их создания.

Компьютерная программа Tajik Convertor, созданная на основе этого алгоритма, использует подпрограммы быстрого автоматического поиска и замены для получения текста в стандартной раскладке. После завершения конвертации всего текста программа выводит окончательный текст одним шрифтом, поддерживающим стандарт Palatino Linotype.

§4.4. Алгоритм проверки правописания на примере таджикского языка

Автоматические системы обработки естественного языка и текстовой информации поддерживают многочисленные пакеты программного обеспечения и компьютерные программы. В области компьютерной лингвистики одной из важных и востребованных задач является разработка системы автоматической проверки правописания и ее редактирования на основе существующих правил определенного языка, пакетов автоматического синтеза и распознавания речи, модулей голосового управления конечного устройства, а также системы автоматического перевода. Более распространенной проблемой является проверка

орфографии. Более того, миллионы людей во всем мире каждый день используют компьютерные средства проверки орфографии.

Компьютерные программы поддерживают два типа систем: средства проверки орфографии и редакторы орфографии. Проверка орфографии обнаруживает орфографические ошибки в данном тексте. Программа проверки правописания обнаруживает орфографические ошибки и ищет наиболее вероятные правильные слова. Автоматическую коррекцию ошибок можно использовать в интерактивном диалоговом режиме, когда пользователь выбирает нужное ему слово из списка правильных слов. В настоящее время в офисных программах и текстовых устройствах используется второй тип проверки и исправления. Проверка орфографии теперь доступна практически во всех настольных приложениях, требующих ввода данных пользователем. Слова с ошибками обычно отмечаются красной линией внизу, чтобы предупредить пользователя об ошибке. Существует несколько типов ошибок, которые можно классифицировать по-разному. Незнание правил правописания языка приводит к ошибкам. При меньшем проценте ошибки при наборе и вводе текста также могут вызывать ошибки передачи информации от одного пользователя к другому. Проверка орфографии последовательно выполняет две основные задачи: сначала находит орфографические ошибки, а затем исправляет их. В целом можно выделить три этапа поиска и исправления ошибок:

- в слове пропущена одна буква, например: вместо «хуршед» написанот «хуршд»;

- в слове две соседние буквы поменялись местами, например: вместо «хуршед» пишется «хушред»;

- в слово входит лишняя буква, например: вместо ««хуршед» пишется «хурршед»

Алгоритмы обнаружения орфографических ошибок. Решение проблемы выявления и исправления орфографических ошибок в текстовой информации считается сложной задачей компьютерной лингвистики. Для решения задачи используются различные методы и алгоритмы. Задача поиска и выявления ошибок

решается путем поиска неправильных слов по правилам языка. Неверное слово, указанное в тексте, отмечается ниже красной волнистой линией. Для решения проблемы многие учёные-лингвисты провели ряд исследований. На последующих этапах были разработаны и реализованы в программных обеспечениях и текстовых процессорах многочисленные методы и алгоритмы. Одним из наиболее эффективных способов выявления орфографических ошибок в тексте является анализ n-грамм символов и поиск существующего словаря. Такие методы анализируют каждый символ слова, выполняют смещение пары соседних символов. Далее идет проверка правильности слова по словарю. Набор правильных слов организован в виде базы данных на основе языкового резерва.

Для создания словарной базы используются инструменты хранения и обработки структуры данных: массив строк; связанный список; древовидная структура; хеш-таблица; реляционная таблица. Для сокращения времени обработки и качественного поиска слова в словаре наиболее часто используемой структурой является хеш-таблица. Хеширование обеспечивает эффективный доступ к поиску слов, что значительно сокращает время обработки словаря. Используя существующие ключи и индексы, функция поиска передается хеш-функции. Использование словаря с древовидными информационными структурами эффективно, когда искомое слово соответствует большому набору неправильных слов. Решить задачу поиска можно, используя набор связанных бинарных деревьев. Исследования показали, что алгоритм Ахо-Корасика является одним из самых популярных и эффективных алгоритмов проверки орфографии. Предлагаемый алгоритм использует словарь. Функция поиска выполняется путем смещения в абстрактной структуре данных. Алгоритм использует эффективную структуру данных – дерево, обеспечивающее учет каждого символа как последовательности и определяющей функции. Проведя сравнительный анализ двух структур данных, древовидной и хеш-таблицы, можно отметить, что, используя n-граммы морфем лингвистического ресурса, можно реализовать более производительный алгоритм выявления орфографических ошибок.

С целью создания словаря на таджикском языке был проведен сравнительный анализ четырех структур данных: бинарного дерева, троичного дерева, многостороннего дерева и дерева с уменьшенной памятью. Образец структуры показан на рисунке 4.9.

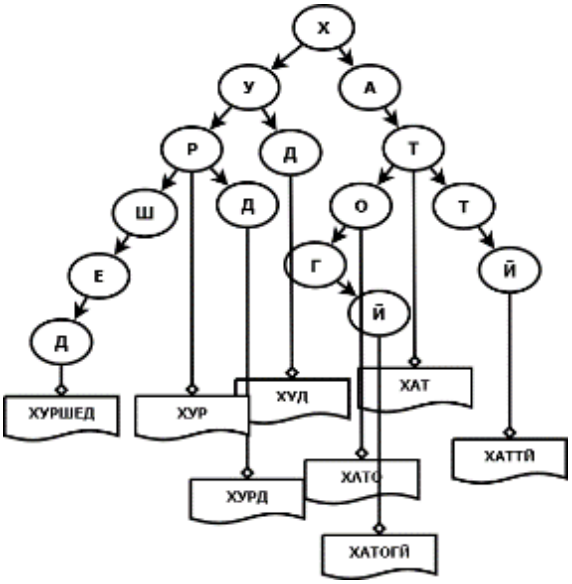

Рисунок 4.9. - Пример древовидной структуры данных для построения словаря

С точки зрения хранения данных и времени поиска слов в словаре было определено, что наиболее подходящей структурой данных является поиск в бинарном дереве.

Следует отметить, что список слов в словаре каждый раз меняется, и в результате мы можем найти большой список. Наблюдаются некоторые ограничения в выявлении ошибок и выполнении поиска правильного слова. Чтобы преодолеть ограничения, необходимо реплицировать предварительное дерево во временное дерево аналогичного уровня сложности. Исследования показали, что хеш-таблицы и древовидные структуры данных могут использоваться для создания словарей в системах автоматической проверки орфографии.

Алгоритмы исправления ошибок. Исправление обнаруженных ошибок производится путем проверки орфографии с возможностью исправления ошибок. Для реализации коррекции осуществляется поиск схожих слов с ошибками в данном слове. Задача редактора состоит из трех этапов: поиск ошибок, создание

списка возможных исправлений, предложение слов для изменения ошибки. Исправление ошибок основано на различных алгоритмах, таких как: исправление ошибок без использования заранее подготовленного списка правильных слов, исправление ошибок с использованием актуального списка слов в словаре.

Рассмотрим основные виды орфографических ошибок. Ошибки могут быть типографскими, когнитивными или фонетическими. Опечатка возникает, когда пользователь нажимает ошибочно другую клавишу компьютера. Когнитивные ошибки возникают из-за незнания правильного написания слова. Фонетические ошибки можно рассматривать как частные случаи когнитивных ошибок, которые относятся к ошибочным словам, которые произносятся так же, как правильное слово.

Относительно распространенным и распространенным алгоритмом, используемым для исправления ошибок, является алгоритм «минимального расстояния редактирования» или само «расстояние редактирования». Этот алгоритм, предложенный Левенштейном, используется во всех функциях проверки орфографии в текстовых редакторах. Расстояние редактирования определяется как минимальное количество операций редактирования, необходимое для преобразования одного слова в другое. В большинстве случаев исправление орфографической ошибки требует вставки, удаления или замены одного символа или перемещения двух символов. Когда неправильное слово преобразуется в словарное слово с помощью одной из этих операций, словарное слово считается приемлемым и реальным исправлением. Алгоритм Левенштейна [68] относится к области динамического программирования и, по-видимому, наиболее широко используется при удаленном редактировании вычислений. При рассмотрении словаря из n слов алгоритмы исправления, основанные на расстоянии редактирования, обычно требуют n сравнений для каждого ошибочного слова. Для сокращения времени поиска используется метод обратного редактирования. Для уменьшения количества сравнений можно использовать отсортированный, разделенный на виды словарь или разделить словарь по алфавитному порядку и длине слов.

Алгоритм проверки орфографии. Сначала в качестве входных данных вводим условное слово с ошибкой – W.

Формируем список таджикских слов, указывая их частотность – S[I]. Для более эффективного поиска слов в словаре создаем базу данных, используя структуры данных хэш-таблицы.

Частоту слов следует определять в зависимости от того, в какой мере слово употребляется в таджикском языке. Затем повторяем поиск по хеш-таблице для каждого найденного в хеш-таблице слова проверяем, совпадает ли слово с ошибкой.

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. Теперь рассмотрим структуру словарной базы.

Структура словарной базы данных представляет собой хеш-таблицу. Словарная база данных реализована на основе двух воображаемых массивов с возможностью хранения ключевых и смысловых пар. В качестве ключей хеш-таблицы массив – корень и элементы являются постфиксами таджикского языка соответственно.

Базовая реализация алгоритма проверки орфографии показана на рисунке 4.10.

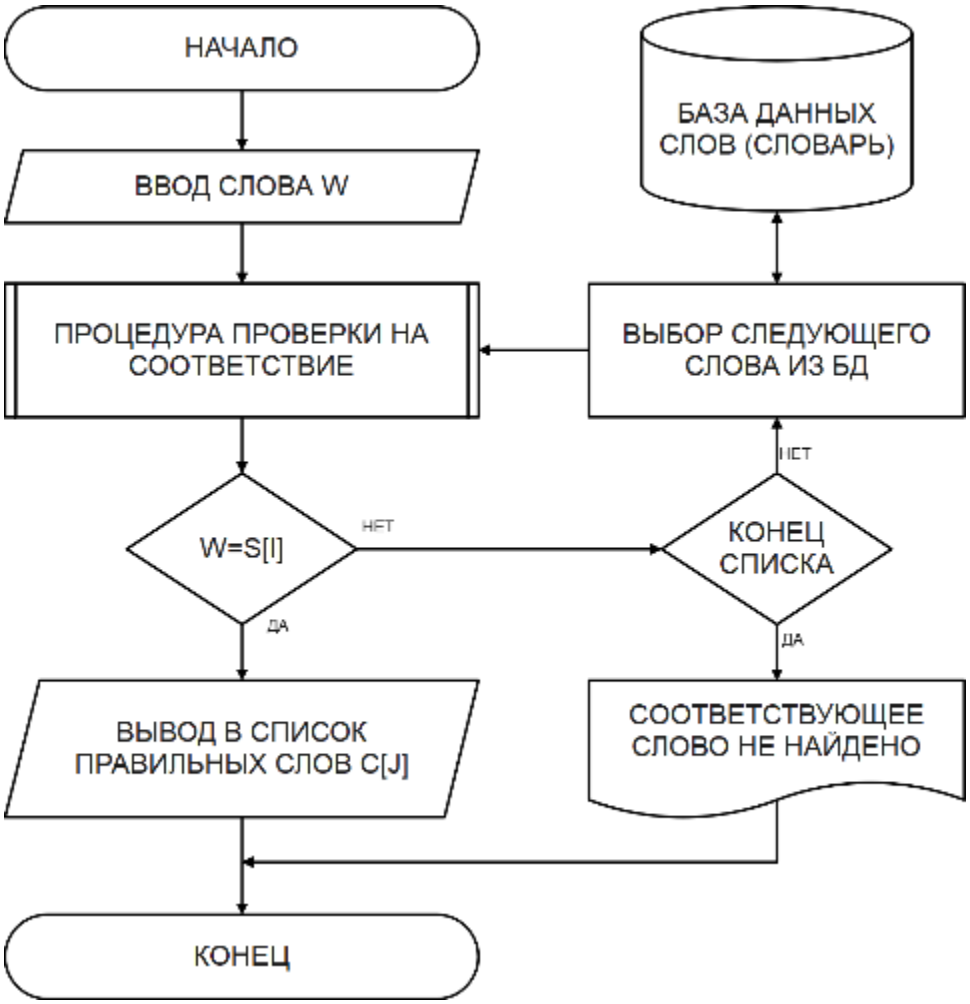

Рисунок 4.10. - Базовая структура алгоритма проверки орфографии

После выполнения специальной хэш-функции получаем словоупотребление таджикского языка. После завершения предварительных вычислений программа проверки орфографии обнаружит слово с ошибкой. На практике хеш-таблица использует вспомогательные индексы вместе с другими хеш-функциями для обработки ошибок, например определения правильной последовательности постфиксов, сопоставления правильных слов с пунктуацией: « . »; « , »; « ; »; « : »; « ? »; « ! »; « () »; « " »; « - »; « … »; « – » ; « * ». С помощью хэш-функций производится поиск орфографических ошибок в каждом вспомогательном индексе и перед ранжированием объединяется список конфликтов.

Наглядный пример формирования списка словоупотреблений в таджикском языке показан на рисунке 4.11.

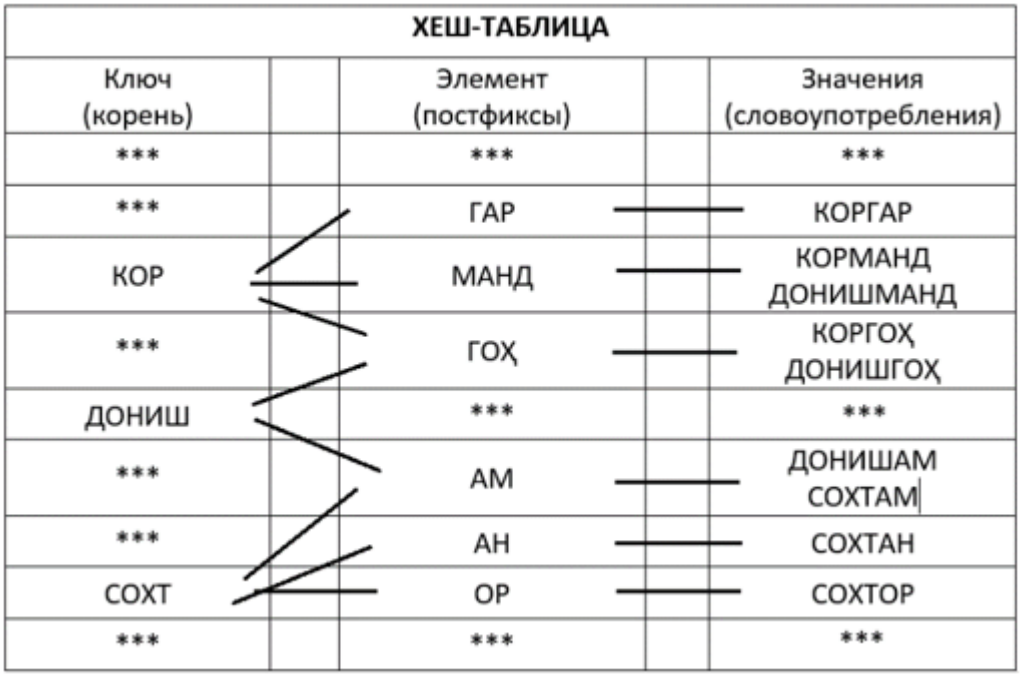

Рисунок 4.11. - Структура хэш-таблицы для генерации использования слов

Важно отметить, что список конфликтов хеш-таблиц содержит только слова, которые есть в словаре, для чего требуется полный список всех корней и постфиксов таджикского языка.

В данной работе анализируются алгоритмы проверки орфографии и исправления естественного языка. Алгоритм обработки представлен в виде блок-схемы, на основе которой с помощью текстового редактора MS Word была подготовлена диагностическая гипотеза программы. При этом рассматривается текст, поддерживающий только буквы таджикского алфавита в стандарте UNICODE. Полученные результаты показывают, что можно разработать реальный инструмент для автоматической обработки таджикского языка с помощью проверенных методов и языковых ресурсов, т.е. данные о корнях и постфиксах. Кроме того, результаты исследования подтверждают, что структура данных хеш-таблицы обеспечивает наилучшую производительность при хранении словаря. Разработанный алгоритм автоматической проверки правописания полностью

соответствует требованиям Закона Республики Таджикистан «Об информации» [4], который включает правоотношения в процессе создания и использования документаций и информационных ресурсов, а также права и обязанности субъектов, участвующих в информационных процессах, то есть пользователей компьютерных программ.

§4.5. Автоматическая система проверки правописания на таджикском языке - TajSpell

Системы автоматической проверки орфографии состоят из программных модулей, выполняющих проверку орфографии заданного текста. Проверка осуществляется на основании правил орфографии языка текста. Ошибки отмечаются особым образом, например, подчеркиванием. В большинстве случаев пользователю предлагается возможность выбрать правильную гипотезу стиля письма из открывшегося списка.

Систему проверки орфографии можно разместить в виде отдельного модуля в программной системе, например, текстового редактора. Он также может работать как отдельная программа. При этом у программы есть возможность интеграции с другими приложениями.

Следует отметить, что первые системы проверки орфографии появились еще в конце 70-х годов прошлого века. Подобную систему для IBM разработала рабочая группа из шести лингвистов из Джорджстонского университета. В 1980 году эти программы были установлены на персональные компьютеры CP/M и TRS-80, а в 1981 году были подготовлены первые пакеты для персональных компьютеров IBM. Эти системы проверки представляли собой независимые программы, доступ к которым осуществлялся из текстовых программ. В насстоящее время во всех программах обработки текста имеются или установлены автоматические языковые системы, поддерживающие более 100 языков. Например, в пакете MS Office поддерживается более 100 языков. Каждая система проверки орфографии основана на базе данных слов и элементов слов – префиксов, суффиксов и основ. К

сожалению, в этой системе нет возможности проверять орфографию текстов на таджикском языке. В России широко используется знаменитый пакет «ОРФО», имеющий возможность проверки правописания на шести языках: русском, английском, немецком, французском, испанском и украинском. Для всех этих языков существует удобная форма сложения слов во всех их формах. Пакет «ОРФО» поддерживает все продукты Microsoft, его можно подключить к программам PageMaker, WordPerfect, WordPro и Quark Xpress. Во всех случаях используется только один пользовательский словарь.

Для реализации автоматической системы проверки орфографии текстов на таджикском языке впервые разработан комплекс программных модулей TajSpell, совместимый с пакетами Microsoft Office.

Система проверки орфографии TajSpell состоит из нескольких программных модулей, которые позволяют пользователю проверять орфографию в режиме реального времени, т.е. непосредственно в момент написания текста (например, в Microsoft Word слова с ошибками отмечаются красной волнистой линией). Нажав правую кнопку мыши, пользователь может исправить неправильное слово, выбрав из предложенных гипотез или внести это слово как правильное в словарь.

Проверка реализована в интерфейсах программного пакета MS Office. Используя стандартные методы работы, с пакетом MS Office можно подготовить свои документы, например в MS Word. Выполнив последовательность команд «*Рецензироание – Язык – Язык проверки правописания* …» и выбрав «*Тоҷикӣ*», пользователь активирует процесс проверки орфографии текста на таджикском языке.

Структура системы автоматической проверки орфографии таджикского языка представлена на рисунке 4.12.

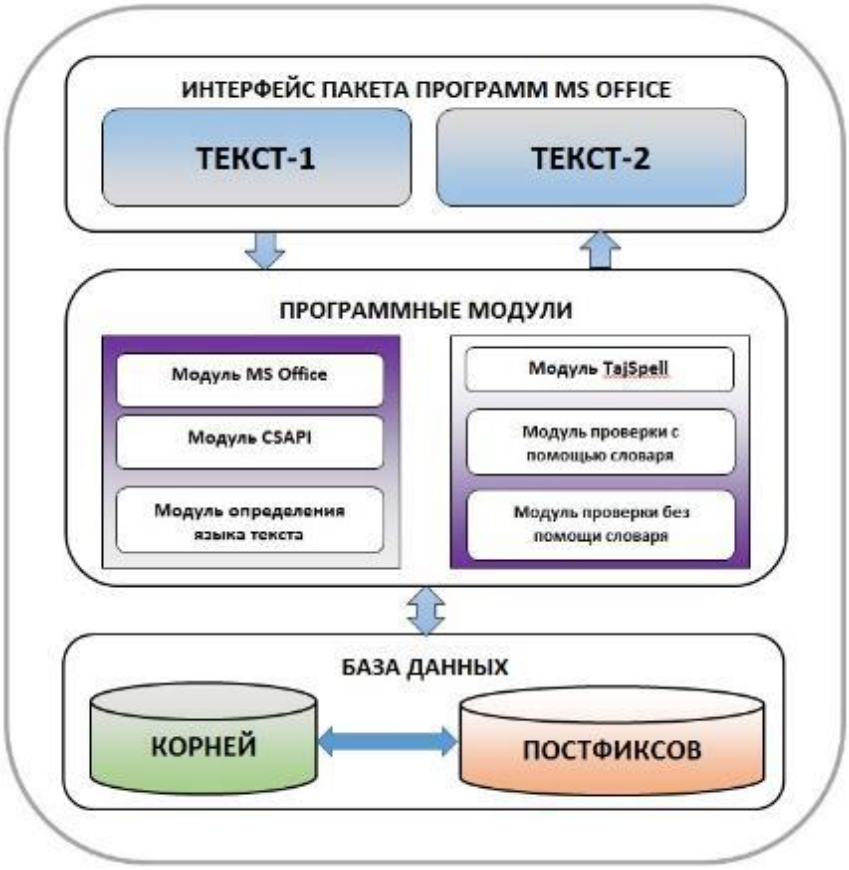

Рисунок 4.12. - Структура системы автоматической проверки орфографии

Модуль проверки правописания соответствует стандарту третьей версии Common Spelling Application Programming Interface (CSAPI).

В приложениях проверка орфографии организована по технологии Microsoft CSAPI, которая поддерживает автоматическую проверку орфографии текстов на таджикском языке.

Модуль TajSpell обладает возможностями исправления таджикского текста, проверки орфографии, переноса со строки на строку и имеет тезаурус таджикского языка.

При проверке модуль орфографии предоставляет следующие возможности:

- для каждого слова, признанного ошибкой, предоставляется список слов, подходящих для исправления;

- помещает выбранную гипотезу в текст;

- предоставляет возможность исправить выбранное слово или добавить его как правильное в словарь пользователя;

- позволяет выбирать из существующих словарей или готовить новый словарь.

Модуль построчного перехода реализует растановка переносов слов в тексте документа по правилам перехода слов таджикского языка.

Модуль «Тезаурус» предоставляет возможность использовать более десяти тысяч синонимов и антонимов таджикского языка.

В целях подготовки основной информации автоматической системы проверки правописания текстов на таджикском языке на первом этапе был обработан набор текстов на таджикском языке, состоящий из 59344883 слов (из них 273734 уникальных слов).

Затем была проведена комплексная обработка текстов, состоящих из 1159344883 слов, 81 постфиксов (простого, двухсоставного и трехсоставного), 128760 различных суффиксов и 70487 основ.

Для поддержки таджикского языка в программном модуле TajSpell подготовлены файлы-словари и суффиксы. В качестве базовых значений используются 70487 регулярных форм (корень и основа), определенных в исходном файле.

Второй файл состоит из 128760 суффиксов, сгруппированных в 865 классов. Показатели классов (флаги) классов обозначаются арабскими цифрами от 1 до 865. Каждый класс состоит из набора суффиксов, образованных из «общей основы».

В связи с тем, что 90,01% таджикских слов не содержат приставок, префиксы не добавляются в файл суффиксов, но полностью присутствуют в основе слов файла словаря.

Приблизительные расчеты показывают, что, используя 70 487 правильных форм (корень и основа) и 128 760 суффиксов, можно охватить более 120 миллионов слов таджикского языка.

Проверка орфографии реализуется с помощью стандартных инструментов, которые предоставляют широкие возможность проверки в любом приложении с данной функцией проверки (например, в MS Word, MS Excel, MS PowerPoint, MS Outlook, Outlook Express и т.п.).

Каждый пользователь может допустить ошибку при наборе текста и его проверке. Эти ошибки – смещение букв, отсутствие пробелов, слившиеся слова или забытые знаки препинания – сложно заметить при наборе или чтении текста. Используя систему TajSpell, можно существенно сократить количество ошибок в тексте или сэкономить время на его исправление.

На рисунке 4.13. показан процесс автоматической проверки орфографии таджикского текста в MS Word.

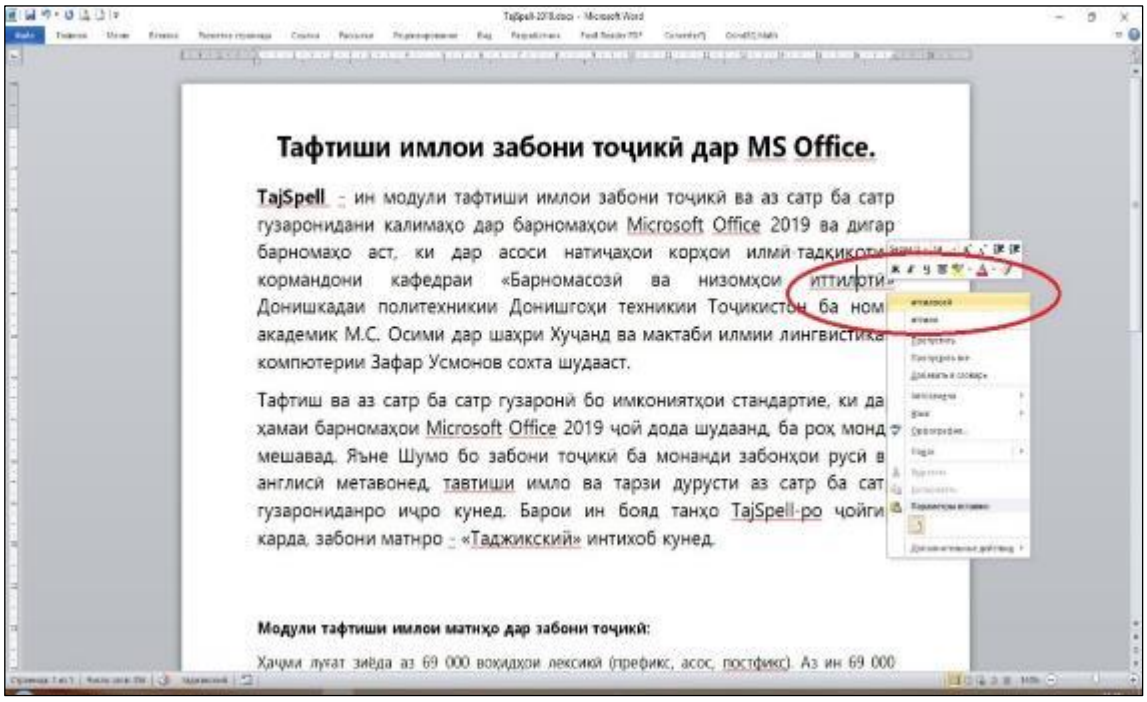

Рисунок 4.13. - Проверка орфографии TajSpell в MS Word

Чтобы активировать проверку орфографии текста в MS Word 2010-2019, на вкладке «Рецензирование» (Review) в группе «Правописание» (Proofing) следует нажать клавишу «Орфография и грамматика» (Spelling and Grammar) или кнопку F7. Для активации проверки правописания, использования тезауруса или перехода от строки к строке необходимо активировать в тексте индикатор языка «Таджикский». При выборе языка кнопок "TG" - таджикский тоже имеет тот же эффект. В этом случае при наборе текста его языком будет «Таджикский», а проверка правописания и переход со строки на строку будут активированы автоматически.

Таким образом можно проверять таджикские тексты, написанные в стандартной кодировке UNICODE. Приложения MS Office 2010-2019 полностью поддерживают буквы таджикского языка и обмениваются ими с модулем проверки орфографии в кодировке UNICODE. Важность использования системы проверки правописания проявляется в том, что она направлена на реализацию нормативно-правовых актов [1-10], реализуемых в Республике Таджикистан.

Выводы по четвертой главе

В зависимости от целей информационного обеспечения системы автоматической проверки орфографии проанализирован вопрос организации и использования электронных словарей в зависимости от их типов. Спроектирована логическая структура словаря. На основе анализа поведения и структуры проекта были определены возможности пользователя и компьютерной программы электронного словаря.

Разработана структура компьютерного тезауруса таджикского языка. На основе интерактивного электронного словаря MultiGANJ создано описание лингвистического тезауруса таджикского языка, насчитывающего более 150 000 словарных единиц.

Учитывая тот факт, что на практике часто используются нестандартные шрифты, разработан алгоритм преобразования символов текста в государственный стандарт раскладки букв таджикского алфавита. Предлагаемый алгоритм проверки орфографии на таджикском языке должен быть адаптирован к кодировке символов таджикского алфавита в стандарте UNICODE.

На основе моделей и математических методов разработаны следующие алгоритмы для решения задачи выявления и исправления орфографических ошибок в текстовой информации на таджикском языке:

- алгоритм преобразования текста в стандартный алфавит;
- алгоритм обнаружения орфографических ошибок;
- алгоритм исправления ошибок;

- алгоритм проверки орфографии.

На основе полученных результатов был разработан модуль TajSpell с возможностью исправления таджикского текста, проверкой орфографии, расстановка переноса слов и тезаурусом таджикского языка. Таким образом можно проверять таджикские тексты, написанные в стандартной кодировке UNICODE.

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

ГЛАВА 5. ПРОЕКТИРОВАНИЕ, РАЗРАБОТКА И ВНЕДРЕНИЕ ТАДЖИКСКОГО АВТОМАТИЧЕСКОГО ПЕРЕВОДЧИКА

§ 5.1. Проблемы художественного перевода и его связь с машинным переводом в Республике Таджикистан

В настоящее время в Республике Таджикистан возникла необходимость развития сферы художественного перевода и его литературной критики. Это связано с тем, что нынешняя ситуация в сфере художественного перевода в сфере таджикского литературоведения переживает не лучшие времена. Об этом свидетельствует небольшое количество специалистов в данной области и завершенных исследований по рассматриваемым вопросам.

Чтобы пролить свет на процесс художественного перевода, кратко рассмотрим историю развития перевода.

С первой половины тридцатых годов прошлого века художественный перевод в Республике Таджикистан развивался по-разному. В этот период наблюдается буквальный перевод фрагментов текста, относительно свободный дословный перевод и художественный перевод. На этом этапе перевод произведений русских писателей производился непосредственно с русского языка, что требовало от переводчика хорошего знания этого языка. Однако некоторые переводчики того периода не знали должным образом особенностей русского языка. Кроме того, они не знали теории художественного перевода и не следовали его методам. Недаром известный русский переводчик О. Кундзич писал: «Есть также писатели и переводчики, которые не знают, что такое перевод. Они видят задачу переводчика в точном копировании всех слов и грамматических форм и в результате готовят дословные переводы».

Конечно, недостатки перевода вызваны не только невнимательностью переводчиков. Точную ситуацию тех лет о состоянии переводческого дела харатеризуют слова таджикского переводчика Х. Ахрори: «Каждый переводчик чувствует ответственность, обращает внимание, старается, но по каким-то

причинам не может, у него не хватает сил, в некоторых случаях он беспомощен, потому что у него недостаточно знаний и опыта в этой области».

Причины сложившейся ситуации следует искать в уровне знаний и культуры таджикского художественного перевода того времени. На начальном этапе общий уровень переводов с точки зрения креативности был не столь высок, и этот недостаток в большинстве случаев приводил к появлению посредственных и некачественных переводов. В этот же период художественные переводы произведений русских писателей характеризовались сокращением слов и фраз оригинала. В некоторых случаях таджикские переводчики, если в оригинальном тексте встречали непонятные или труднопонятные слова, отбрасывали их и продолжали переводить текст. В результате такого «творчества» исходный текст искажался, художественность произведения снижалась и, наконец, приводило к разрушению стиля писателя в переводе.

Известно, что художественный перевод является одним из средств обогащения словарного запаса таджикского языка. Благодаря переводу в родной язык включаются новые слова, словосочетания и выражения. Действительно, хороший перевод обогащает таджикский язык, но дословный и некачественный перевод разрушает его. В частности, включение синтаксических конструкций и фразеологии иностранного языка разрушает богатую лексико-грамматическую систему родного языка. Это рассматривается как негативное явление, при котором возникает риск утраты связи языка и национальной культуры. В результате плохого знания переводного материала и непонимания тонкостей языка произведения таджикские переводчики в большинстве случаев сокращали предложения исходного текста. Однако исключение слов, фраз и предложений из исходной версии приводит к неисправимым ошибкам. В некоторых случаях писатели-переводчики, которые плохо знали русский язык и испытывали трудности при переводе, игнорировали отдельные трудные русские предложения и выражения.

Однако в переводах прошлых лет отмечается очень много сокращений слов и словосочетаний. Кроме того, некоторые переводчики грубо нарушали методику художественного перевода, вследствие чего допускали немало ошибок. В

некоторых случаях таджикские переводчики, не понимая значения некоторых русских слов и фраз, переводили их на таджикский язык наугад, полагаясь на свою интуицию, которая их часто подводила, в результате чего больше искажали идею оригинального текста. Подобные недостатки можно увидеть практически во всех художественных переводах произведений русских писателей.

Сейчас посредством переводов в таджикский литературный язык внедряются некоторые фразы и предложения, чуждые нормам таджикского литературного языка и его грамматике. На наш взгляд, это тревожное событие началось с переводов предыдущих лет и продолжается и в настоящее время. Эти дословные переводы привели к возникновению в таджикском литературном языке неясных слов, несвязных словосочетаний и предложений.

Большинство корявых, бессвязных и неправильных предложений созданы в результате дословного перевода, они встречаются в художественном переводе практически у всех таджикских переводчиков того периода. Помимо предложений неправильной формы, а также обрывочных фраз и оборотов, которые были дословно переведены с русского языка или созданы методом кальки, они получили широкое распространение в переводных произведениях таджикских писателей.

С первых лет возникновения национальной переводческой школы, наряду с дословными переводами, некоторые таджикские переводчики относительно свободно подходили к национальным особенностям и лексическому составу оригинального текста, чтобы приблизить перевод к идее и общему содержанию художественного произведения. В результате относительно свободного подхода к оригинальному тексту таджикские переводчики добились больших художественных успехов. Но, с другой стороны, такое отношение иногда приводит к несоответствию формы и содержания переводного произведения, к искажению исходного текста (авторского стиля, языка и художественных особенностей произведения). Вместе с тем, относительно вольный перевод понравился таджикскому читателю больше, чем дословный перевод. Творческий перевод требует отражения не слов, а духа произведения.

Переводчики больше внимания уделяли содержанию произведений русских писателей, иногда проявляя безразличие к лексике оригинала. Они пытались содержание произведения перевести на таджикский язык и адаптировать его к вкусу таджикского читателя.

Следует сказать, что деление предложений оригинала на части и его перевод противоречит методу художественного перевода, поскольку в большинстве случаев нарушается связь между содержанием и формой текста. Однако такой подход не всегда дает желаемый результат и часто всего приводит к тому, что предложения значительно длиннее исходного текста значительно увеличивают процесс перевода. В некоторых случаях переводчики включали новые слова и фразы, чтобы сделать перевод более эффективным и понятным. Подобный подход к оригинальному тексту не приемлем, т.к. иногда приводит к противоположному смыслу оригинальной версии.

Другим недостатком, который можно увидеть в большинстве художественных переводов последних лет, является обильное и неуместное использование диалектных слов. Таджикские переводчики сознательно или невольно использовали при переводе много диалектных слов. В предыдущих переводах в основном встречаются диалекты северных регионов страны (поскольку большинство переводчиков были уроженцами городов Самарканд, Бухара, Худжанд и др.), некоторые из этих слов не характерны для таджикского языка. Кстати, в переводах в основном используются диалекты южных районов Таджикистана.

Художественный перевод является одним из важнейших вопросов литературных отношений, и его всестороннее изучение дает ценный материал в области литературного влияния. Сегодня значение художественного перевода возрастает как никогда, поскольку это одно из хороших средств сближения литературы и народов.

В XXI веке практическое использование компьютерных средств и информационно-коммуникационных технологий в национальном и международном масштабе является основным фактором развития человечества в

сфере информации. Машинный перевод – инструмент, который рассматривается как один из способов создания многоотраслевого пространства во всем мире.

На основе научных исследований и анализа мировых проектов в настоящее время определено, что машинные переводчики делятся на следующие группы:

- машинный переводчик на основе правила RBMT (Rule-based Machine Translation);

- статистический машинный переводчик SMT (Statistical Machine Translation);

- машинный переводчик на основе нейронных сетей NMT (Neural Machine Translation);

- трансляция памяти TM (Translation Memory);

- адаптивный машинный переводчик AMT (Адаптивный машинный перевод)

- интерактивный машинный переводчик IMT (Интерактивный машинный перевод);

- гибридный машинный переводчик HMT (Hybrid Machine Translation).

Исходя из классификации систем перевода и их практической значимости, выделяются следующие виды:

1. Полностью автоматический машинный перевод.

2. Автоматический машинный перевод с участием человека.

3. Машинный перевод с участием человека с использованием ограниченного предметного словаря для небольшого круга прикладных предметных областей.

4. Полный перевод человека с помощью компьютера на примере использования Translation Memory.

Конечно, каждый из вышеупомянутых классов имеет свои преимущества и недостатки. В целях продвижения научного проекта в направлении разработки машинного переводчика таджикского текста второго класса, то есть автоматический машинный перевод с участием человека, ставится на первое место.

До сих пор не существует автоматического проекта или системы-переводчика для перевода всего текста с таджикского языка на другой язык. Автоматический перевод текста, вероятно, доступен для ограниченного числа языковых групп. Например, с английского на русский. Разработка систем онлайн-

перевода международными компаниями считается одной из главных современных и актуальных задач. В таблице 5.1 представлена информация о популярных системах перевода, названии компании, стране и возможности перевода таджикского текста.

Таблица 5.1. - Список популярных систем перевода текста

№	Система перевода	Компания	Страна	Перевод таджикского текста
1	Google Translate	Google	США	есть, с 2015 г.
2	SYSTRANet	Systran	Франция	нет
3	Translate.ru	PROMT	Россия	нет
4	translate.yandex.ru	Yandex	Россия	нет
5	Windows Life Translator	Microsoft	США	нет
6	Worldlingo	Systran	Франция	нет
7	Free Translation	SDL	Великобритания	нет
8	ImTransator	Smart Link Corporation	США	нет
9	Babel Fish	Systran	Франция	нет
10	InterTran	Translation Experts Limited	Великобритания	нет

Из таблицы, видно, что возможности перевода таджикского текста доступны только в системе перевода.

В целях анализа работы переводчика были изучены и выявлены различные критерии оценки качества получаемого перевода.

В результате исследования потенциальности онлайн-перевода Google были установлены следующие возможности:

- определение языка текста;
- перевод слов;
- перевод предложений;
- перевод веб-страниц;
- перевод рукописных текстов;
- перевод имеющихся текстов на картинке;
- голосовой перевод;

- голосовое чтение перевода текста;

- ведение истории переводов;

- проверка орфографических ошибок;

- перевод текстов на 107 поддерживаемых языков;

- перевод документов в форматах DOC, DOCX, TXT, RTF, HTML, Android Resource (XML), Application Resource Bundle (ARB) и других;

- добавление переведенного контента определенного языка;

- поддержка памяти транслятора TMX;

- использование трансляторов в других системах;

- наличие Google Translator Toolkit и сервисов Google API.

Наряду с выявлением возможностей был замечен и список недостатков онлайн-переводчика Google:

- дословный перевод текста;

- потеря содержания исходного текста в случае его обратного перевода с текста, переведенного на другой язык;

- использование системы переводчика только в онлайн-режиме;

- перевод текста с грамматическими ошибками;

- неправильный перевод сложных предложений;

- недоступность источника данных переводчика.

Полученный результат послужил основой для разработки таджикского переводчика на базе технологии Google. На этапах реализации проекта проводится анализ стандартов и разработка машинного переводчика. Текущий проект по таджикскому языку реализуется в направлении компьютеризации таджикского языка.

§5.2. Система автоматической транслитерации

Проектирование системы транслитерации на основе статистических методов. В процессе научно-исследовательской работы, опираясь на вопросы проектирования автоматического переводчика, были предложены методы

транслитерации текста, написанного на таджикском языке. На этом разделе рассматриваются задачи реализации оптимальной транслитерации т.е. преобразования шрифтов таджикского алфавита на шрифты других языков. Прежде всего, представлены методы преобразования текста с кириллицы в латиницу. Также рассмотрен набор письменных носителей для преобразования таджикского алфавита в кириллицу русского языка и латиницу английского языка. Представлен порядок и схема работы информационной системы транслитерации текста на таджикском языке с поддержкой международных стандартов.

Таджикистан обрел государственную независимость в 1991 году, и одним из первых шагов нашей страны было вхождение в мировую политику. Более 150 стран признали Таджикистан в международном сообществе, из них 126 стран имеют дипломатические отношения с нашей страной. Благодаря усилиям Основоположника мира и национального единства, Лидера нации, Президента Республики Таджикистан Эмомали Рахмона, Таджикистан занимает достойное положение в сфере политики, экономики, культуры, образования и науки на региональном и мировом уровнях. В настоящее время Таджикистан является членом ряда международных и региональных организаций, в том числе ООН, СНГ и ШОС.

В рамках двустороннего сотрудничества между странами широко реализуется большой процесс документации и ее обмена как на бумажных, так и на электронных носителях. Обмен информацией между представителями стран обеспечивают системы автоматического перевода текста с одного языка на другой. Процесс перевода позволяет перенести смысл информации с иностранного языка на целевой язык. В процессе перевода информации возникает явление буквенной конверсии, основанное не на значении слова, а на его звучности. В некоторых системах транслитерация – это процесс переноса слов из алфавита одного языка в другой с целью произнесения слов на иностранных языках. Вообще в процессе преобразования букв происходит явление замены буквы одного слова исходного алфавита на фонетически сходные буквы другого алфавита.

Несмотря на то, что на протяжении всего периода существования человечества разными цивилизациями и народами использовались различные письмена, сегодня подавляющее большинство жителей планеты используют только пять графических систем (с соответствующими изменениями, иногда очень серьезные изменения, доработки и обновления для каждого конкретного языка). В таблице 5.2. перечислены письменности, используемые в современном мире, и их характеристики.

Таблица 5.2. - Наиболее популярные письменности в мире

№	Вид письма	Направление письма	Возникновение, век	Структура алфавита
1	Китайский	сверху вниз (слева направо)	с VIII века до нашего времени	выше 50 тысяч иероглифов
2	Латинский	сверху вниз (слева направо)	с VII века до нашего времени	26 букв
3	Индийский	сверху вниз (слева направо)	с III века до нашего времени	50 основных знаков
4	Арабский	сверху вниз (справа налево)	V век	28 букв
5	Кириллический	сверху вниз (слева направо)	VIII век	33 буквы (традиционный)

Следует сказать, что *китайские иероглифы* используются в Китае, Японии, Южной Корее и Северной Корее.

Индийская письменность используется в северо-восточных странах Индийского океана: Индии, Шри-Ланке, Бангладеш, Мьянме, Таиланде, Лаосе, Камбодже, Непале, Бутане.

Сегодня *латиница* применяется во всех странах Северной, Центральной и Южной Америки, большинстве стран Европы и Африки, а также в странах Азии: Турции, Азербайджане, Индонезии, Малайзии, Вьетнаме, Узбекистане, Туркменистане.

Арабская письменность используется в странах Азии, таких как Мавритания, Марокко, Западная Сахара, Алжир, Тунис, Ливия, Египет, Судан, Ливан, Сирия, Иордания, Саудовская Аравия, ОАЭ, Кувейт, Катар, Бахрейн, Йемен, Оман, Иран, Афганистан, Пакистан.

По сравнению с вышеупомянутыми системами кириллица является относительно молодой письменностью и используется в качестве основной письменности в России, Белоруссии, Украины, Сербии, Болгарии, Северной Македонии, Казахстане, Киргизии, Таджикистане, Монголии, а также как в некоторых европейских странах - в Боснии и Герцеговин, Чернеогории.

Таджикский язык использует арабскую графику с момента распространения ислама в странах Центральной Азии. В 1929 году на основе реформ таджикской государственности в стране стал использоваться латинский алфавит. С 1940 года и по сегодняшний день в таджикском языке используется кириллица. В 1998 году в таджикском языке была проведена реформа, и сегодня в таджикском языке используются 35 букв кириллицы.

Решение проблемы преобразования шрифтов заключается в представлении символов одного сценария символами другого сценария с сохранением возможности получения обратного симметричного соотношения. Использование строгой последовательности символов или сопоставления символов решает проблему преобразования различного количества символов в двух алфавитных системах. Основная цель – обеспечить автоматическое восстановление и точный смысл исходного текста. Другими словами, преобразование или транслитерация текста должна возвращать исходный текст в контексте утвержденных критериев преобразования, определяющих правила замены одного символа другим символом.

Результат транслитерации не означает, что слово, написанное с использованием набора символов, например, английского латинского алфавита, не является переводом. Если исходное слово не имеет значения в данном языке, то и его преобразованная форма не имеет значения. Основная функция преобразования шрифтов – обработка текстовых данных при поиске, выборе или индексировании текстового контента. На основе процесса преобразования шрифтов можно найти информацию, написанную другим алфавитом, и вернуть ее к сценарию пользователя по умолчанию.

В компьютерной лингвистике для преобразования шрифтов рекомендуются следующие требования:

- использование международных систем транслитерации и совместимости алфавитов;

- трансформация имен собственных, в том числе личных имен, топонимов, наименование предприятий по сути личных имен, периодических изданий, фольклорных героев, названий стран и народов, наименование национально-культурных реалий;

- термины в специальных сферах с традиционно установленными возможными формами, которые требуют возврата именно в той форме, в которой они используются;

- вспомогательный элемент или компонент смешанного перевода текста методом кальки.

Основные правила преобразования шрифтов из кириллицы в латиницу содержатся в существующих стандартах. В результате исследования установлены правила транслитерации специальных таджикских букв в русскую кириллицу и английскую латиницу.

Правила преобразования таджикской кириллицы в российскую кириллицу представлены на рисунке 5.1.

CH-taj		CH-rus		CH-eng
Ғ	→	Г	→	GH
Ӣ		И		I
Қ		К		Q
Ӯ		У		U
Ҳ		Х		H
Ҷ		ДЖ		J
Ъ		фосила		фосила

Рисунок 5.1. - Порядок обмена специальных буквх таджикского алфавита

На основании проведенных исследований были определены следующие требования в процессе транслитерации: использование международных стандартов произношения букв согласно алфавиту; преобразование имен собственных, например, имен людей, топонимов, названий стран и народов, периодических

изданий, литературы; термины в специальных областях; элемент смешанного перевода текста с учетом требований исходного языка. На основе существующих стандартов и исследования требований к транслитерации создана логическая структура информационной системы (рис. 5.2.).

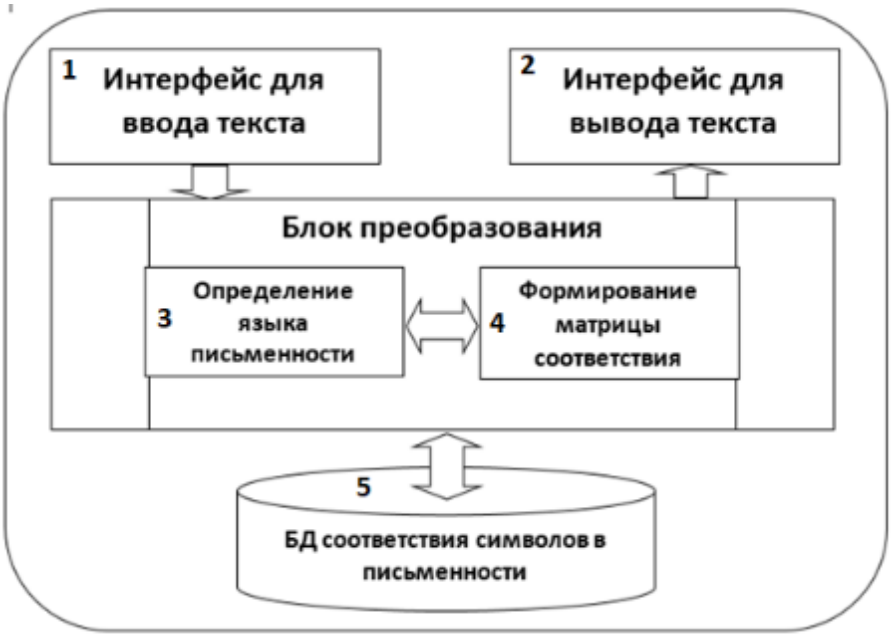

Рисунок 5.2. - Логическая структура системы автоматического транситерации

Структура состоит из трех частей: пользовательского интерфейса, функциональной модели и базы данных. В первой части необходимо создать интерфейсы ввода и вывода текстовой информации, блок 1 и 2 соответственно. Вторая часть определяет основной функциональный блок-системы – блок преобразования текста.

Для формирования преобразования необходимо сформировать два модуля, связанных между собой: модуль определения письменности (блок 3) и модуль формирования матрицы соответствия (блок 4). Третья часть структуры требует создания пар языков для реализации транслитерации и соответствующей матрицы обмена в виде базы данных соответствующих письменных символов (блок 5).

В результате проведенных исследований была разработана система автоматической транслитерации с учетом предложенной логической структуры.

Модели файловых данных реализованы для формирования базы данных линейных символов. В образце модели системы транслитерации показаны следующие языковые пары: таджикская кириллица и английская латиница, таджикская кириллица и русская кириллица. На рисунке 5.3 показан процесс транслитерации названия научной статьи с таджикского языка на английский.

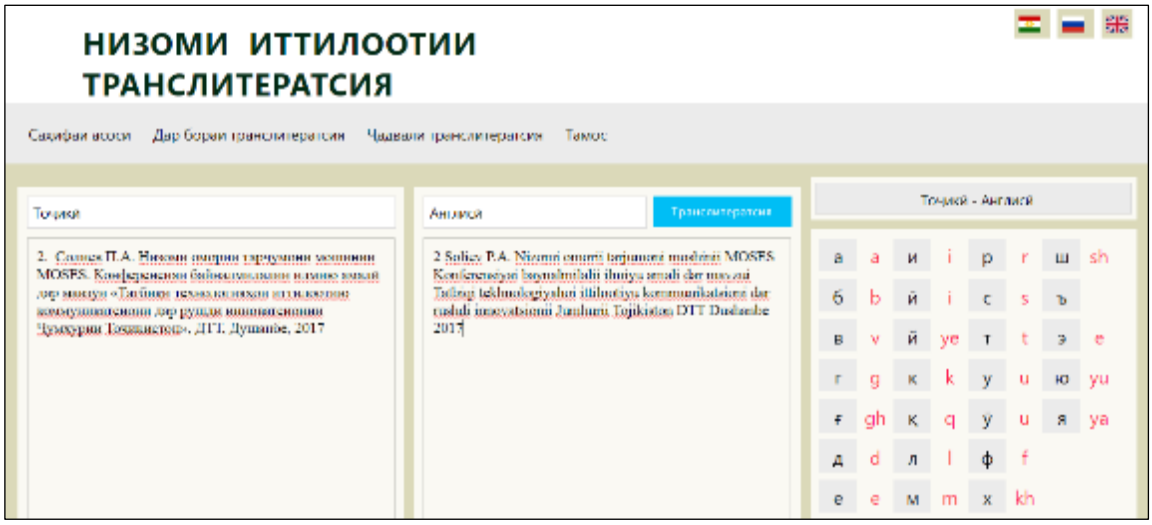

Рисунок 5.3. - Образец формы системы транслитерации (tj-en)

Для детального анализа процесса транслитерации на главной странице системы отображается матрица соответствия. Для расширения практических возможностей системы необходимо добавлять новые пары языков с учетом матрицы совместимости символов (рис. 5.4).

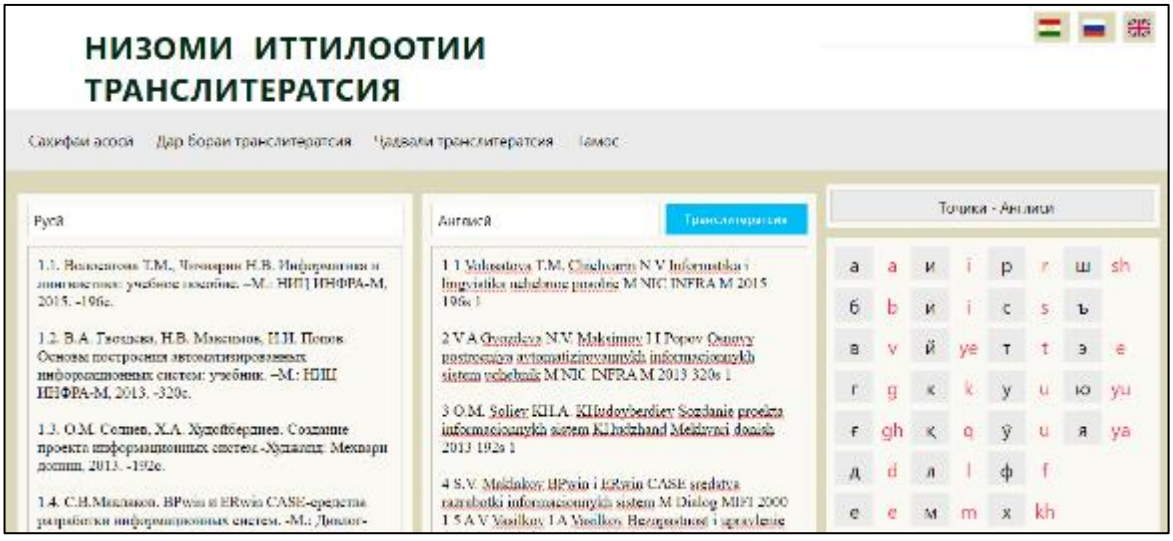

Рисунок 5.4. - Образец формы системы транслитерации (ru-en)

В проекте используется статичная модель транслитерации текста таджикской кириллицы. В настоящее время разработанная модель и система автоматической транслитерации в электронной библиотеке политехнического института ТТУ показывает удовлетворительные результаты. В дальнейших исследованиях предусмотрено проектирование и разработка транслитерации с учетом возможностей нейронных сетей и глубокого машинного обучения, преобразование букв таджикского алфавита в английский и русский алфавит. Разработанный проект доступен в виде веб-приложения во Всемирной паутине по адресу www.tajlingvo.tj\transliteration.

Проектирование системы обмена письмами на основе машинного обучения. Обмен букв осуществляется с использованием понятия транслитерации – преобразования одной или нескольких букв одного алфавита языка (исходного языка) в одну или несколько букв другого алфавита (конечного языка). В целом процесс обмена осуществляется на основе равного фонетического значения букв, соответствия звукового характера конечных букв и особых правил написания основного языка.

Система автоматической транслитерации разработана на основе методов и алгоритмов и реализована с учетом возможностей станков. В настоящее время на основе статистических методов и специальных языковых правил разработаны различные алгоритмы машинной замены букв. На основе алгоритмов разработаны системы анализа текстовых ресурсов, разработки многоязычного текста, поиска и выделения данных из текста, написанного на разных языках. Следует отметить, что система машинного перевода имеет большую потребность в алгоритме машинной замены букв, поскольку в случае машинного перевода имена собственные не могут быть переведены с одного языка на другой.

Значительные результаты в развитии машинной транслитерации были достигнуты учеными дальнего и ближнего зарубежья. Так, ведущие учёные Индии разработали и использовали собственные комплексы систем для транслитерации индийского алфавита, а учёные из арабских стран – для арабо-персидского

алфавита. Свои результаты и достижения также представили ученые-компьютерные лингвисты Российской Федерации, Белоруссии, Казахстана.

С учетом развития таджикского языка и обретения Республикой Таджикистан прочного места на мировой арене, вопрос машинной транслитерации таджикского языка на другие алфавиты мировых языков становится актуальной проблемой и требует своего решения.

Вопрос решения задачи транслитерации наталкивается на ряд сложных проблем. Так, трудности в замене букв представляет сложное и запутанное произношение слов в зависимости от фонетических правил в разговорной речи. Кроме того, несоответствие произношения одной буквы одного алфавита языка паре буквы других алфавитов также является причиной поиска дополнительных научно-технических путей его решения. Поэтому эти типы букв анализируются и обрабатываются отдельно. Бывают случаи, когда некоторые буквы в процессе обмена вообще остаются незамеченными. Есть группа букв, которая преобразуется в другую группу букв в связи с особым своим звукопроизношения.

В настоящее время автоматическая транслитерация широко используется в процессе создания электронных каталогов в библиотеках, обработки личных документов, авиабилетов, банковских карт и в целом при подготовке типовых документов. Следует отметить, что в любом случае процесс транслитерации зависит от правил правописания изучаемого языка. Разумеется, результат замены нацелен на группу заинтересованных людей, чтобы они могли правильно, понятно и привычно произнести полученное слово, используя свой алфавит. На основе проведения научно-практических исследований по замене букв предложены методы сегментации (декомпозиции), то есть разделения слова на символы или пары байтов. Кроме того, вместе с машинным обучением различных структур слов представлены рабочая среда рекуррентных нейронных сетей (RNN – Recurrent Neural Network) и точные инструменты нейронного машинного перевода (NMT – Neural Machine Translation).

Сопоставление двух последовательностей символов (Conv Seq2Seq – Convolutional Sequence to Sequence) – относительно новый инструмент в области

машинного обучения, который успешно применяется для транслитерации таджикского алфавита в буквы алфавита другого языка. Метод сопоставления применяется для предложения транслитерации на основе частоты встречаемости производных слов. Этот метод считается относительно эффективным для транслитерации по сравнению со статистическим методом.

В результате научно-практического исследования можно прийти к выводу, что независимо от языка общения можно предлагать разные структуры транслитерации на основе машинного обучения.

Подготовленные конструкции направляются на проектирование систем автоматической обработки данных на таджикском языке. До сих пор использовалось множество систем для транслитерации из алфавита одного языка в другой на основе статистических структур. Разумеется, эти системы используются и для транслитерации таджикского алфавита.

С учетом возможностей нейронных сетей и глубокого машинного обучения спроектирована и разработана автоматическая система транслитерации таджикского алфавита в английский и русский алфавиты. Разработанные методы в настоящее время показывают удовлетворительные результаты. Отредактированный проект доступен в Интернете по адресу tajlingvo.tj\transliteration.

§5.3. Система машинного перевода на таджикский язык

Классификация и сравнительный анализ системы машинного перевода. С развитием информационного общества возрастает интерес к языкознанию, поскольку благодаря этой науке большая часть интеллектуальной деятельности человека возлагается на электронно-вычислительные машины (ЭВМ), в том числе и на перевод текстов.

Машинный перевод – это процесс перевода текста с одного естественного языка на другой естественный язык с сохранением содержания с помощью

специальной программы ЭВМ. В настоящее время существует три типа систем машинного перевода:

- статистическая система машинного перевода (SMT). Этот вид машинного перевода текста основан на сравнении текстов большого размера попарно (то есть текстов, включающих соответствующий перевод предложений на другой язык). На базе этого типа работают машины-переводчики «Яндекс.Переводчик», Google Translate и ABBYY;

- машинный перевод на основе правил (RBMT). В основе этого типа лежит связь между структурой основного текста и переводимого текста. На базе систем этого типа работают системы машинного перевода PROMT в России, SYSTRAN во Франции и Linguatec в Германии;

- система нейронного машинного перевода (NMT). Эта система, как статистическая система, сравнивает большие тексты попарно и учится находить законы соответствия между ними. В системе нейронного машинного перевода перевод текста осуществляется на основе предложений, а не слов и фраз, в отличие от системы SMT. Системы машинного перевода Яндекс. Переводчик (с 14 сентября 2017 г.) и Google Translate (с 27 сентября 2016 г. в некоторых существующих языковых парах) используют систему NMT наряду с SMT.

Из сравнения систем, представленных в таблице, видно, что использование SMT и NMT требует более мощного оборудования, чем RBMT, а также зависит от языковй базы. Использование системы RBMT считается очень удобным, поскольку не зависит от языкового ресурса, тогда как SMT и NMT полностью зависят от имеющихся ресурсов. По качеству перевода RBMT также превосходит SMT и NMT. Также в случае использования RBMT нет необходимости в мощном аппаратном обеспечении.

Преимущество SMT и NMT перед RBMT состоит в том, что их легко расширить при наличии новых лингвистических ресурсов, поскольку в случае расширения RBMT необходимы специальные знания и много времени.

Преимущества и недостатки систем SMT, RBMT и NMT представлены в следующей таблице 5.3.

Таблица 5.3. - Преимущества и недостатки систем SMT и RBMT

Наименование системы	**Преимущества**	**Недостатки**
SMT	1. Наличие большого лингвистического ресурса, служащего для улучшения и ускорения работы программы. 2. Хорошее качество перевода текста по определенной теме. 3. Переведенный текст аналогичен переводу «переводчик-человек».	1. Зависимость от языкового резерва. 2. В случае невозможности перевода текста в попытке найти правильный ответ выдвигается наиболее подходящая гипотеза. 1. Тесная связь с языковым резервом. 2. Отсутствие эквивалента в языковом резерве ограничивает возможность добавления и улучшения переведенного произведения. 3. Невозможность предсказать результат перевода. 4. Недостаток работы при переводе "по правилам". 5. Необходимость мощного устройства
RBMT	1. Возможность редактирования оригинального текста, что повысит качество перевода. 2. Не требует мощного программного обеспечения. 3. При переводе используются грамматические правила. 4. Адекватное качество перевода текстов на общие темы	1. Большая потребность в специальных знаниях обычного пользователя. 2. Необходимость больших затрат рабочей силы и времени со стороны разработчиков. 3. Необходимость поддержки обновления источников лингвистических данных.
NMT	1. Нейронная сеть работает с предложениями, что обеспечивает хорошее качество перевода. 2. Текст не делится на слова и словосочетания, этот метод позволяет учесть смысловую связь внутри предложения. 3. Поскольку NMT не работает с частями предложения, перевод становится более связным и "плавным".	1. Зависимость от языкового резерва. 2. В случае невозможности перевода текста в попытке найти правильный ответ выдвигается наиболее подходящая гипотеза. 3. NMT не всегда правильно переводит необычные имена и редкие слова.

В целом, в связи с ростом спроса общества на информацию, спрос на перевод данных растет с каждым днем, и высокий приоритет отдается развитию машинных переводчиков. Крупные компании, такие как Яндекс, Google, ПРОМПТ и другие, прилагают много усилий для правильного внедрения машинных переводчиков, результатами работы которых пользуется практически каждый пользователь ЭВМ. Следует отметить, что сегодняшнюю жизнь невозможно представить без использования машинных переводчиков, ведь они оказывают большую помощь «человеку-переводчику».

В рамках проекта «Обработка таджикского переводчика на базе технологии Google» предлагается использовать систему SMT для обработки таджикского переводчика на основе технологии Google. В связи с тем, что система RBMT не используется в рамках технологии Google, разработать переводчик таджикского языка на основе этой технологии с использованием системы RBMT невозможно. В настоящее время отсутствие решения вопроса «Data mining» («Интеллектуальный анализ данных») для таджикского языка делает невозможной разработку таджикского переводчика на базе технологии Google с использованием системы NMT.

Нейронные сети и их использование в машинном переводе. Первые исследования в области компьютерной лингвистики начались в 1950-1960-х годах, но разработать совершенный алгоритм машинного перевода печатных и рукописных текстов с одного языка на другой с помощью специальной компьютерной программы не удалось.

В настоящее время существует несколько компьютерных программ, позволяющих осуществлять перевод отдельных слов (электронные словари) и полных предложений с согласованием морфологических, семантических и синтаксических отношений членов предложения. Несмотря на способность выдавать относительно «понятные» фразы, машинный переводчик не был способен переводить грамматические явления.

В наше время существуют машинные переводчики таджикского языка, которые лишь переводят введенное в программу слово с таджикского языка на

другие языки. Например, компьютерная программа «4in1 Dictionary, доступ к которой можно получить на следующем веб-сайте www.tajlingvo.tj. Программа предназначена для перевода слов с русского, английского на таджикский и наоборот, а также помогает облегчить освоение таджикского, английского и русского языков. Программа «4in1-Dictionary» имеет следующие базу данных:

- резерв таджикско-русских языков – 42 000, в том числе словарный запас более 27 000;

- русско-таджикский языковой резерв – 68 000, в том числе более 54 000 слов;

- англо-таджикский языковой резерв – 12 000, в том числе более 5 000 слов;

- таджикско-английский языковой резерв – 24 000, в том числе более 11 000 слов.

В связи с этим рассматриваются способы перевода предложений таджикского языка на другие языки. Основная цель – объяснить алгоритм перевода предложения с помощью нейронных сетей. В эту цель входит решение задачи анализа существующих программ применения машинного перевода. Актуальность работы отражена в проблемах, связанных с машинным переводом. Новизна исследования заключается в описании нового метода машинного перевода. Ее теоретическая значимость и практическая ценность проявляется в том, что использование нейронной сети способствует правильному и точному переводу текстов на другой язык. При работе с переведенными текстами использовался метод сравнительного анализа. Ниже приведены отдельные понятия, необходимые для понимания о машинном переводе:

Apertium – бесплатное программное обеспечение для машинного перевода. Apertium был разработан Университетом Аликанте.

Машинный перевод – обработка перевода текстов (устных и письменных) с помощью компьютерной программы с одного языка на другой.

Как уже говорилось выше, машинный перевод – это перевод текстов с языка оригинала на целевой язык с помощью специальных компьютерных программ. Другими словами, для создания перевода на компьютере в программу вводятся правила и законы слова и его перевода. Система машинного перевода состоит из

двуязычного словаря, содержащего информацию о грамматике, синтаксисе и семантике, а также алгоритмических инструментов, необходимых для перевода. Различаются три типа систем машинного перевода: подход, основанный на правилах, подход, основанный на примерах, и статистический подход.

С помощью данных систему можно реализовать только средствами синтаксиса, при котором слово переводится в слово, основной единицей которого является слово, все синтаксические отношения этого слова связаны с исходным вариантом слова. Таким образом, традиционный способ – упорядочить исходный текст в переведенном тексте, сохраняя при этом структуру переводимых и удаляемых предложений.

Учитывая, что системы используют разные правила и разные переводы переводчиков, результаты перевода показывают, что в настоящее время система машинных переводчиков работает не очень хорошо, поскольку все переведенные тексты регулярно отличаются от исходного содержания.

Вместе с тем, ученые сравнили качество переводчиков с помощью аналитических тестов, среди которых были наиболее популярные три системы (Google, Яндекс и ПРОМТ). При переводе в системе «Яндекс.Перевод» наблюдался правильный перевод художественной литературы, а в Google - переводчике перевод научной литературы.

Все существующие переводчики можно оценить положительно только за дословный перевод, но не за перевод биграммы слова (пары слов), триграммы слов (последовательности из трех слов) и фразы в целом. Неправильный перевод машинных переводчиков (переводчик Google) можно увидеть в следующих оборотах: matter wave - "мавҷи материя " вместо "мавҷи Бройл"; SQUID bias – «ғарази калмар » вместо «ивази калмар» (дастгоҳи интерференсияи квантии суперноқил); ; Greens function - «функсияи сабз» вместо «функсияи Грин». Приведенный пример показывает, что данная проблема существует не только для текстов на таджикском языке, но и для других крупных современных языков.

Современные средства машинного перевода имеют возможность переводить предложение слово в слово, что в целом приводит к потере смысла переведенного

предложения. Самым некачественным результатом данного вида перевода является перевод русских текстов на японский и с японского на русский язык. Основная причина этого в том, что при переводе используется не прямой перевод, а через промежуточный язык (английский). Однако проблема машинных переводчиков может быть успешно решена при предварительном обучении нейронных сетей. Объяснением этому служит то, что статистические инструменты рассматривались только для дословного перевода текстов.

Нейронные сети работают совершенно по-другому. Это предполагает не только возможность нескольких вариантов перевода, но и интеллектуальный анализ предложений и разделение их на «лексические части». Внутри сети находится информация о «лексических частях», соответствующих значению слова.

Искусственная нейронная сеть – это математическая модель, действующая по принципу организации и работы биологических нейронных сетей и основанная на сетях нервных клеток живых организмов. Нейронные сети невозможно запрограммировать, но их можно обучить. Способность к обучению – одно из главных преимуществ нейронных сетей перед традиционными алгоритмами. Методика обучения заключается в нахождении коэффициента корреляции между нейронами-факторами.

Алгоритм работы нейронных сетей при переводе текста с языка оригинала на язык перевода следующий:

1. Ввод текста.

2. Разделение текста в нейросети по значению слов.

3. Обработка нейросетей по уровню и смыслу предложения.

Виды решения задачи с использованием нейронных сетей следующие:

- в комплект входит одна величина, предлагается одна величина, где в программу вводится слово и в результате предлагается перевод этого слова;

- в комплект входит одна величина, предлагается несколько величин, где при вводе слова в программу предлагается несколько вариантов перевода этого слова;

- в комплект входит несколько величин, предлагается одна величина, где при вводе в программу несколько синонимов одного и того же слова, предоставляется одно слово;

- включено несколько параметров, предложено несколько переводов, где при вводе в программу несколько предложений одного языка предлагается перевод всех этих предложений на другой язык.

Самым сложным видом решения этой задачи является то, что наша задача решается таким же образом. Данная технология была использована для одного из сложных китайско-английских языков в системе Google. В результате нейросети снизили количество ошибок перевода на 60%. Результат работы нейронных сетей представлен на рисунке 5.5.

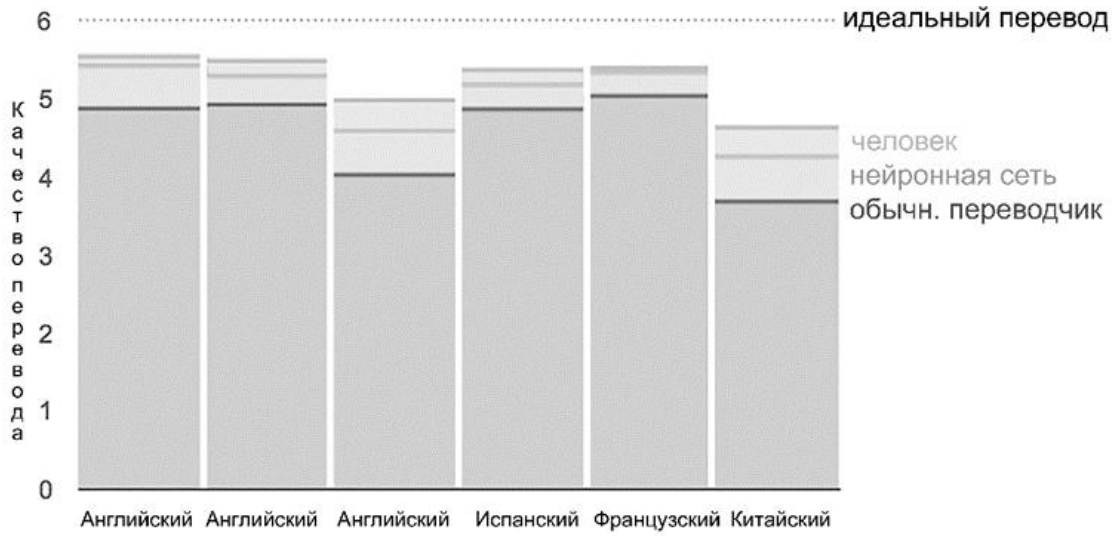

Рисунок 5.5. - Результат работы нейронных сетей

Модель переводчика. Хотя результаты экспериментов дали относительно хорошие результаты, учёные полагают, что компьютер сможет приблизиться к человеческому переводу или сравняться с ним по качеству перевода, только если удастся добавить видео и звук в нейронные сети. Однако, несмотря на активное развитие этой сферы, такого уровня она еще не достигла.

Машинный перевод – это коллективный процесс анализа и интеграции с восстановлением исходного смысла текста переведенного языка. Как показывают

эксперименты, качество нейронного перевода выше качества традиционных методов машинного перевода, однако высокий уровень перевода текстов с полным содержанием до сих пор не замечен разработчиками систем.

§5.4. Логическая структура и анализ артефактов машинного перевода

В XXI веке практическое использование компьютерных средств и информационно-коммуникационных технологий в национальном и международном масштабе считается основным фактором развития человечества в сфере информации. Машинный перевод инструмент, который рассматривается как один из способов создания мультидисциплинарного пространства во всем мире.

В настоящее время машинный перевод – это область, требующая анализа и изучения процессов научных исследований. Благодаря внедрению в жизнь машинного перевода большое количество электронных документов и информации во всемирной паутине стало открытым и доступным для всех пользователей мира, независимо от языка текста. Без использования машинного переводчика тексты на различных иностранных языках становятся недоступными для людей, не владеющих этими иностранными языками. Поэтому в настоящее время машинный перевод считается основным инструментом распространения информации по всему миру. Но в нстоящее время качество и результаты работы машинного переводчика в зависимости от его классификации и практических возможностей требуют устранения орфографических ошибок и содержания получаемого текста.

На практике машинный перевод делится на два класса или вида перевода:

1. По признакам оригинального текста, т. е. в зависимости от жанра и способа его написания. В основном текст готовится из художественных, промышленных, общественно-политических и специальных новостных жанров. Ввиду предварительного анализа структуры художественного текста реализация его машинного перевода считается достаточно сложной. Однако перевод отраслевых текстов относительно прост, поскольку в настоящее время для большинства отраслей доступны серии словарей и отраслевых текстов.

2. От особенностей процесса перевода, т.е. в зависимости от результата перевода в письменной или устной форме. Информационные технологии значительно расширяют возможности автоматического перевода. Однако для реализации устного перевода необходимо решение такой проблемы компьютерной лингвистики, как синтез и распознавание голоса.

С целью анализа практических возможностей и степени автоматизации машинного перевода были классифицированы три группы машинных переводчиков:

- полностью автоматический;
- частично автоматический, с участием редактора;
- перевод переводчиком с использованием электронных словарей.

Для реализации первой и второй группы машинный перевод в рамках его логической структуры использует два основных раздела (рисунок 5.6).

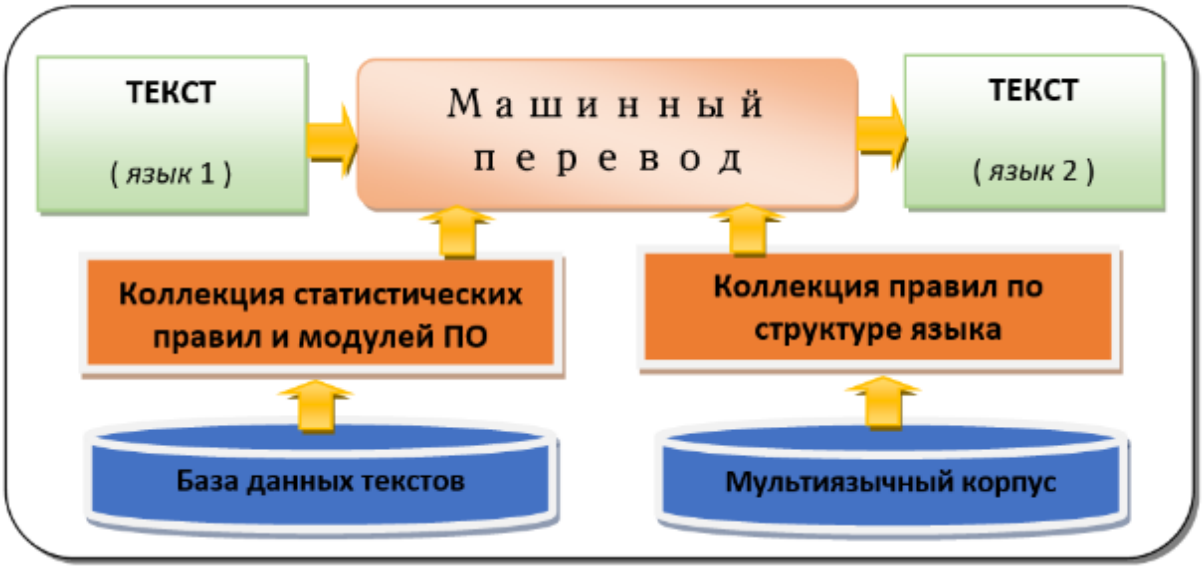

Рисунок 5.6. - Логическая структура машинного переводчика

В первом разделе набор статистических правил и программных модулей обрабатывает тестовые текстовые данные. В этом случае правила формируются на основе математического и статистического инструментария. Алгоритмы, управляемые правилами, преобразуются в программные модули. При этом

качество работы алгоритмов сильно зависит от объема контрольных текстовых данных. Во втором пункте на основе набора языковых правил обрабатывается многоязычный ресурс, то есть максимально возможный параллельный межъязыковой ресурс.

Основные ошибки машинного перевода. В процессе использования машинного перевода в переводимом тексте в основном встречаются две группы ошибок: орфографические ошибки и ошибки в содержании текста. Согласно терминам лингвистической науки, их можно назвать орфографическими и стилистическими ошибками.

Орфографические ошибки – несоблюдение совокупности орфографических правил языка, которые используются для создания текстовых единиц, таких как слова, словосочетания и предложения. По результатам научных исследований выделены следующие виды синтаксических ошибок:

- неверная последовательность слов;
- неправильная структура имени;
- неправильная структура глагола;
- неправильная структура числительного;
- неправильная структура прилагательного;
- неправильное и неполное употребление местоимений;
- неправильное и неполное использование окончания;
- неправильное и неполное использование префиксов.

Стилистические ошибки – это несоблюдение содержания языковых единиц, слова, словосочетания, предложения и их логического содержания. По результатам научных исследований выделено несколько видов стилистических ошибок, возникающих при замене слов:

- исчезновение слов в предложении;
- использование дополнительных или лишних слов;
- использование синонимов;
- использование неправильного перевода слов.

Чтобы планировать и разрабатывать проекты артефактов машинного перевода, прежде всего необходимо избегать перечисленных выше ошибок. Кроме того, следует подготовить набор статистических и языковых правил для обработки текстовых данных, а также составить алгоритмы машинного перевода.

§5.5. Таджикско-русская информационная система автоматического переводчика

Таджикский язык – литературный и разговорный язык таджиков, государственный язык Республики Таджикистан. Таджикский язык также распространен в ряде регионов Узбекистана, Казахстана, Кыргызстана и северного Афганистана с древней культуры и средневековой литературной традиции.

Русский язык – один из восточнославянских языков, национальный язык русского народа, входящий в шестерку самых популярных языков мира по общему числу русскоязычных и восьмой среди стран, использующих его в качестве второго государственного языка.

Большая совокупностьпараллельных текстов называется «параллельным ресурсом» или «параллельным корпусом». Параллельная обработка ресурсов требует выравнивания параллельного текста путем сопоставления предложений в обеих половинах параллельного текста.

На практике параллельный корпус используется для получения перевода текста в определенной форме. С научной точки зрения формирование параллельных корпусов позволяет реализовать важные научно-исследовательские задачи в области компьютерной лингвистики. Развитие параллельного таджикско-российского корпуса укрепит отношения между таджикским и русским народами и поможет народам двух стран улучшить свои знания на обоих языках.

В данном разделе рассматриваются цель и процесс обработки таджикско-российского параллельного корпуса, приводятся способы дальнейшего использования ресурса и машинный перевод.

В настоящее время в Таджикистане машинный перевод с таджикского языка на другие языки и наоборот считается одной из важнейших задач в области компьютерной лингвистики. В частности, необходимо создать параллельный ресурс по переводу текстов с таджикского языка на русский язык, что доказывает важность данной проблемы.

Также в рамках создания параллельного корпуса возникнет необходимость проведения ряда статистических исследований, обработки текстовых данных, исследования компонентов используемых языков.

В параллельный корпус вошли тексты из следующих сфер: политики; классической и современной литературы; истории; права и юриспруденции; журналистики; межгосударственных контрактов и соглашений.

В рамках создания параллельного ресурса были проведены следующие работы:

- выбор правильности текста;
- предварительная обработка текста;
- комментирование источников текста;
- выравнивание текстов;
- разработка алгоритмов обработки текста;
- создание Taj-Rus-Corp с возможностью текстового поиска;
- ввод текстов в параллельный корпус;
- статистический анализ данных;
- создание тестовых модулей машинного переводчика.

Структура базы данных корпуса. База данных параллельного корпуса включает следующую информацию: язык, вид, название документа, автор, источник и год публикации текста. Структура источника данных разработана с использованием системы управления источниками данных MySQL, способной хранить и обрабатывать большие объемы данных. Основные таблицы параллельного хранения состоят из следующих пластов:

- rt – порядковый номер;

- toj – текст на таджикском языке;
- rus – перевод текста на русский язык;
- namud – вид текста;
- matn – название документа;
- muallif – автор;
- manba – источник текста;
- sol – год издания.

Информационная система Taj-Rus-Corp. Для создания параллельного корпуса и непосредственно создания источника текстовых данных использовалась авторская программа *Taj-Rus-Corp*. Эту программу также можно использовать для создания и организации параллельных корпусов других языков. Посредством этой программы создается основа параллельных текстов двух и более языков с функциями управления этими корпусами. Программа поддерживает разные шрифты соответствующих языков в текстах и кодировку Unicode. В программе предусмотрено полуавтоматическое сравнение и выравнивание текстов по предложению или абзацу. Также представлена возможность обработки текста по символам.

Следует отметить, что программа Taj-Rus-Corp позволяет осуществлять настройку, фильтрацию и поиск данных в параллельном хранилище. Перечисленные задачи управления в параллельном корпусе предусматривают возможность использования машинного переводчика.

Программные модули создаются с использованием различных алгоритмов обработки текстовых данных. В качестве элементов поиска в тексте были выбраны униграммы, биграммы, триграммы слов. Также возможно использовать несколько режимов вывода результатов в форматах TXT и HTML (рисунок 5.7.).

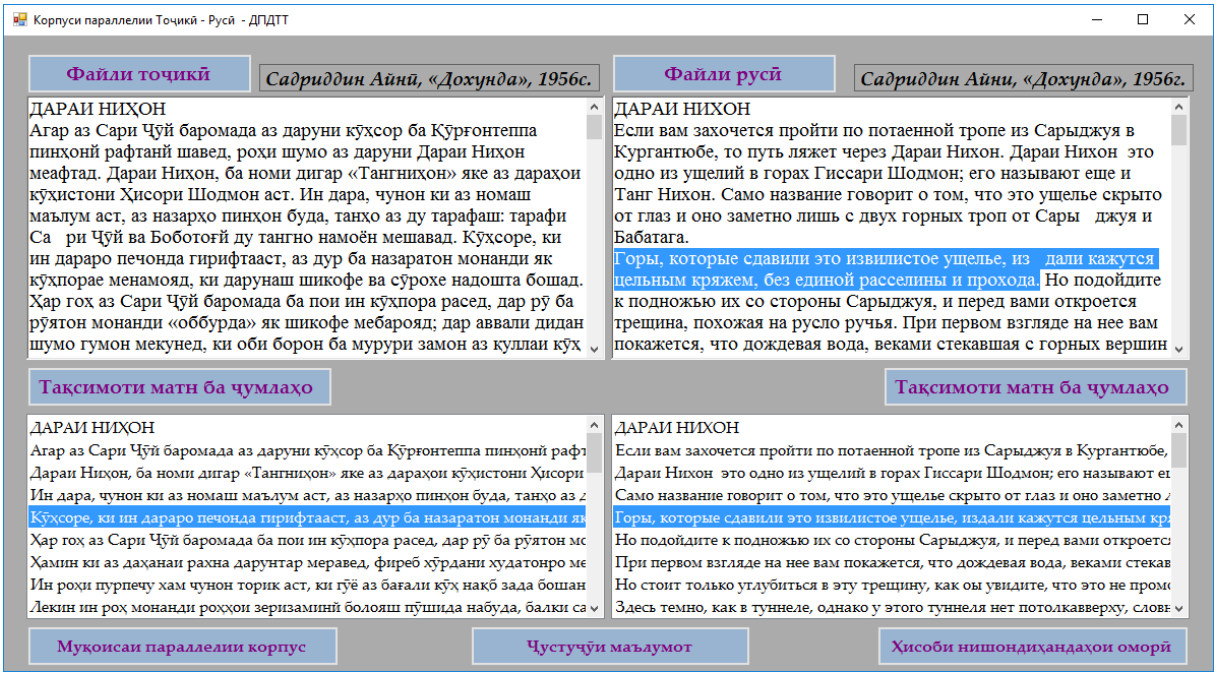

Рисунок 5.7. - Интерфейс выравнивания предложений в программе Taj-Rus-Corp

Статистический анализ данных можно отметить как научно-теоретический аспект параллельного использования корпусов. Учитывается возможность расчета частотности таких элементов в тексте, как слог, слово и словосочетание. Перечисленные функции позволяют учесть необходимые особенности природы языка корпуса.

В программе Taj-Rus-Corp используются следующие алгоритмы поиска: простой текстовый поиск; широкий поиск с использованием нескольких выражений; поиск по текстовым элементам; параллельный поиск на двух языках. Основным инструментом поиска является язык структурированных запросов SQL. Также используется возможности полуавтоматического морфологического анализа с применением часто встречающихся выражений. Функции поиска расширены возможностью поиска словоформ и использования слов, ключевых слов по лемме.

В целом, подводя итог полученным результатам, необходимо отметить, что разработанный параллельный корпус требует дальнейшей обработки и анализа. В будущем планируется улучшить функциональные возможности предварительной обработки текста. Планируется также внедрение автоматического

морфологического анализа. Для получения результатов разрабатываются более мощные инструменты, включая алгоритмы и методы поиска. На базе параллельного корпуса создаются терминологические и специальные словари. Различные статистические методы анализа текста строятся на основе элементов и реальных текстовых данных. Все полученные результаты будут изучены в дальнейшем и послужат основой для разработки новых параллельных корпусов, связанных с таджикским языком как единый модуль, приложение.

Нейронный машинный перевод: использование opennmt для обучения модели переводчика. В рамках исследовательской работы описано использование нейронных сетей в машинном переводе с интерпретацией этапов разработки машинного переводчика с одного языка на любой другой. В данном разделе для решения задачи машинного перевода предлагается использование системы Open-Source Neural Machine Translation (OpenNMT) со свободным исходным кодом PyTorch, который доступен на языке программирования Python. Основная цель дизайна OpenNMT – предоставить благоприятные условия для глубокого изучения и реализации своих идей в области машинного перевода, синтеза, морфологии, преобразования изображений в текст и т.д.

Хотя существуют эффективные системы машинного перевода Microsoft, Яндекс и др., они не свободные источники или ограничены рамками лицензии. Другая система, такая как tensorflow-seq2seq, существует для машинного перевода, но она работает в качестве исследовательского кодирования. OpenNMT не только свободный источник, но также имеет хорошую справочную систему, модульную систему и читаемый код для быстрого и эффективного обучения.

Согласно рекомендациям была определена следующая информация об архитектурной структуре OpenNMT: OpenNMT – это полноценная система для обучения и проектирования моделей нейронного машинного перевода. Эта система представляет собой следующее поколение seq2seq-attn, которое было полностью переработано для упрощения обучения и повышения эффективности машинного перевода (рисунок 5.8).

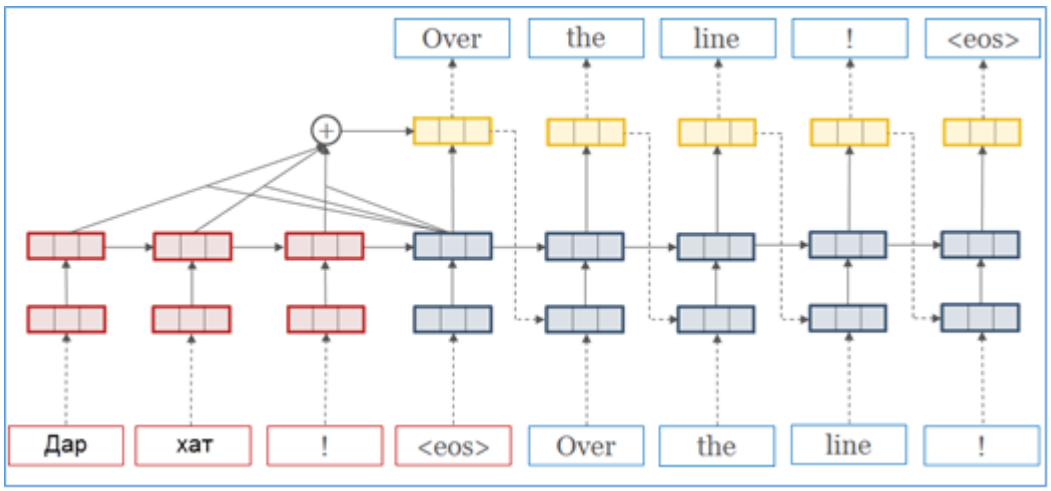

Рисунок 5.8. - Краткие комментарии структуры OpenNMT

Таким образом, предоставляется подробное руководство по настройке каталога PyTorch и использованию инструментов для изучения системы машинного перевода. Полученные результаты включают разъяснения по созданию перевода английского текста на таджикский язык.

Базовая система реализована на основе математической структуры Lua/Torch и может быть легко расширена с использованием стандартных внутренних компонентов нейронной сети Torch.

Настройки основных модулей сбора данных. Основным пакетом, используемым для обучения системы машинного перевода, является PyTorch, в котором реализован модель OpenNMT.

Первым шагом при установке системы является копирование репозитория OpenNMT, который доступен в Интернете. Для обеспечения бесперебойной работы системы рекомендуется установить PyTorch.

Набор данных состоит из файла параллельного корпуса исходного и переводимого языков, имеющему в каждой строке одно предложение, и токен, отделенное пробелами.

В ходе исследования использовалась параллельная база данных предложений английского и таджикского языков, хранящаяся в отдельных файлах. Данные собраны и объединены из различных источников. Затем данные размещаются

таким образом, что создается набор файлов, который выглядит следующим образом:

- src-train.txt: обученный файл с 5000 английскими предложениями;

- tgt-train.txt: обученный файл с 5000 предложениями на таджикском языке;

- src-val.txt: проверочный материал с 1000 предложениями на английском языке;

- tgt-val.txt: проверочный материал с 1000 предложениями на таджикском языке;

- src-test.txt: проверочный материал для оценки, содержащий 1000 предложений на английском языке;

- tgt-test.txt: проверочный материал для оценки, содержащий 1000 предложений на таджикском языке.

Все упомянутые выше файлы расположены в каталоге /data.

В следующем разделе кратко рассматриваются системы машинного перевода и ее проверка. Однако для обеспечения эффективного словарного запаса и приближения к человеческому переводу следует использовать большие ресурсы, содержащие миллионы предложений.

Проверочный материал используется для оценки модели на каждом этапе для определения точки сходства. Обычно он должен содержать не более 5000 предложений. В следующем примере показано, как текстовые данные сохраняются в соответствующих файлах:

Исходный файл:

«Это является развитием, хотя и не фундаментальное, но усиливается».

«В результате делопроизводств во всех предприятиях становятся все более цифровым».

«Использование ИКТ настолько разнообразно, что требует проведения научных исследования».

«Взаимосвязь информационных и коммуникационных технологий в естественных языках».

Файл для перевода:

«Ин як рушд аст, гарчанде ки бунёдӣ нест, аммо шиддат меёбад»

«Дар натиҷа, ҷараёни кории офис дар тамоми корхонаҳо торафт рақамӣ мешавад»

«Истифодаи ТИК ба дараҷае гуногун аст, ки онҳо ба таҳқиқоти илмӣ ниёз доранд»

«Муносибати технологияҳои иттилоотӣ ва коммуникатсионӣ дар забонҳои табиӣ».

Предварительная обработка текстовых данных. Для предварительной обработки текстовых данных для обучения, проверки и разработки словарей создан специальный модуль preprocess.py, который выполняется следующей командой:

python preprocess.py -train_src data/src-train.txt -train_tgt data/tgttrain.txt -valid_src data/src-val.txt valid_tgt data/tgt-val.txt -save_data data/demo

Обучение моделя переводчика. Модуль train.py предназначен для обучения модели переводчика. Основная команда для обучения модели очень проста:

python train.py -data data/demo save_model демо-модель

Приведенная выше команда активирует модель из примера, которая состоит из двухслойного LSTM и имеет 500 скрытых частей для кодирования и декодирования. Чтобы указать использование графического процессора при обучении, в приведенную выше команду необходимо включить аргумент gpuid (например, -gpuid 1 для использования GPU1)

Как правило, модель выполняется по образцу до 100000 шагов, поэтому контрольная точка записывается после каждых 5000 шагов. Затем, если модель адаптируется и точность контроля скорее достигнет точки стабильности, можно остановить будущее обучение и использовать заранее записанную контрольную точку.

Использование модели машинного перевода для перевода текстов. Модуль Translate.py предназначен для перевода. Этот модуль предназначен для перевода незнакомого текста с английского на таджикский. Этот модуль сохраняет результат перевода предложения в отдельный файл.

Обучение модели проводилось на 10000 шагов на NVIDIA GEFORCE 2GB. Обучение на CPU обходится очень дорого, поэтому для обучения модели с большим количеством данных на высокой скорости рекомендуется использовать высокопроизводительный CPU.

Например, для выполнения переводов, приближенных к реальному миру, в модели необходимо обучить большие научные словари и около миллиона предложений, что в то же время требует значительных вычислительных затрат по сравнению с аппаратными требованиями и временем обучения. Для проверки и улучшения работыподготовлено и представлено к использованию веб-приложение для перевода таджикско-русских предложений.

Веб-приложение таджикский переводчик www.tarjumon.tj предназначено для онлайн-перевода текстовой информации с таджикского языка на русский и наоборот. Для повышения качества перевода разработаны также морфологический анализ слова, тезаурус таджикского языка (рис. 5.9).

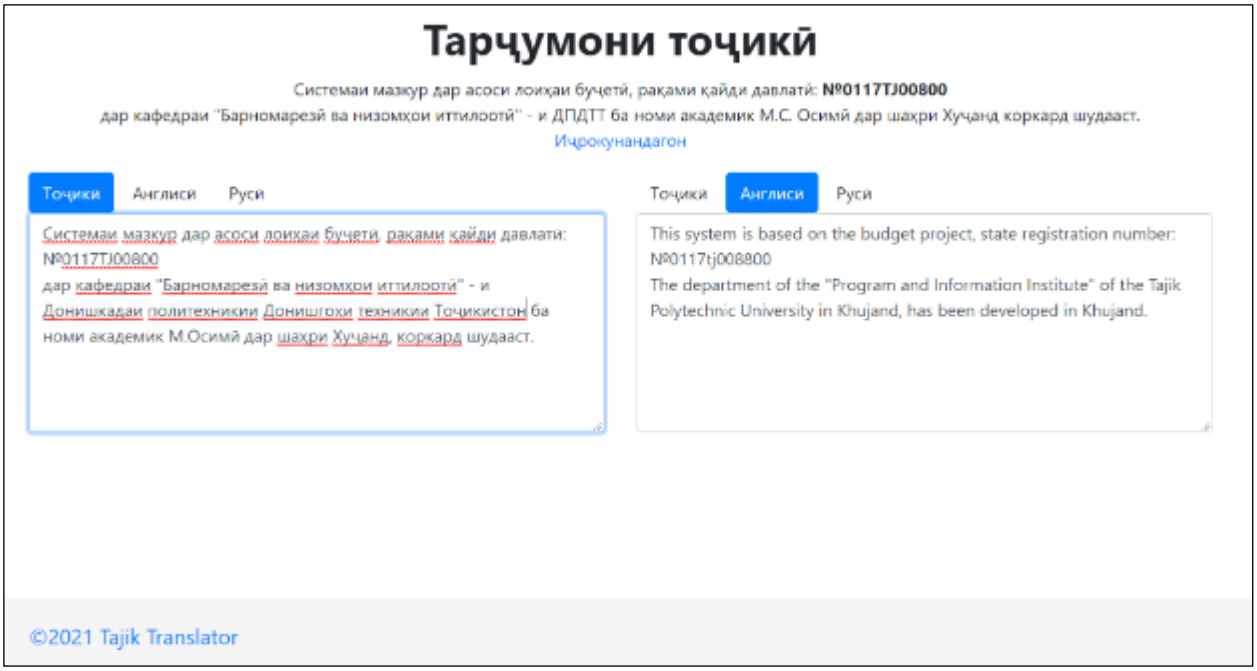

Рисунок 5.9. - Проект таджикского переводчика - www.tarjumon.tj

Веб-приложение также можно использовать как вспомогательную систему для изучения таджикского, русского и английского языков. В настоящее время в

Таджикистане не существует системы автоматического перевода, как веб-приложение www.tarjumon.tj, переводящий на русский, английский и таджикский языки.

Веб-приложение предоставляет возможность автоматического перевода текста на русский и английский с таджикского языка и наоборот. С помощью Web-приложения также можно выполнить морфологический анализ слов и использовать тезаурус таджикского языка.

Выводы по пятой главе

По согласованию со специалистами, имеющими большие навыки в процессе перевода текстов с разных языков на таджикский и наоборот, были выявлены проблемы художественного перевода и его зависимость от машинного перевода.

На основе научных исследований и анализа мировых проектов были проанализированы методы и алгоритмы машинного перевода для перевода текста на таджикский язык.

Выявлены и проанализированы математические модели и методы, используемые в популярной системе машинного перевода онлайн-переводчика Google. На основе этого были изучены возможности и недостатки системы онлайн-переводчика Google.

В процессе обработки машинного перевода в переведенном тексте возникают такие ошибки, как орфографические и стилистические. В связи с этим были разработаны методы и алгоритмы исправления этих ошибок в тексте на таджикском языке.

В переводимом тексте в процесс перевода не передается возможность встречи слов, основанных на именах собственных. По этой причине возникает необходимость замена букв из разных компьютерных алфавитов на компьютерный алфавит таджикского языка. Для решения этой проблемы на основе методов статистического и машинного обучения была разработана система автоматической транслитерации.

С целью информационного обеспечения системы машинного перевода для были разработана параллельные корпуса Taj-Rus-Corp и Taj-Eng-Corp. Корпусы состоят из следующих частей: языка, вида, названия документа, автора, источника и года публикации текста. На первом этапе общее количество элементов запаса определяется следующим образом:

- таджикско-русский языки – 42 000, в том числе словообразовательных более 27 000;

- русско-таджикский языки – 68 000, в том числе словообразовательных более 54 000;

- англо-таджикский языки – 12 000, в том числе словообразовательных более 5 000;

- таджикско-английский языки – 24 000, в том числе словообразовательных более 11 000.

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 была разработана информационная система в виде Web-приложения. Проект доступен в Интернете по адресу www.tarjumon.tj для онлайн-перевода текстовой информации с таджикского языка на русский и английский языки и наоборот.

ГЛАВА 6. ПРОЕКТИРОВАНИЕ, РАЗРАБОТКА И ВНЕДРЕНИЕ КОМПЬЮТЕРНОГО СИНТЕЗА ТАДЖИКСКОЙ РЕЧИ ПО ТЕКСТУ

§ 6.1. Анализ текстовых данных на основе разных слоговых структур

Случайный выбор текстовых данных. В качестве текстовых данных была сделана репрезентативная выборка из 3800 страниц художественных произведений, газетных статей и специальной литературы на таджикском языке приведенные в таблице 6.1.

Таблица 6.1. - Список произведений для отбора текста

№	Автор/жанр	Наименование произведения	Кол-во стр.
1	Абу Алӣ ибни Сино	«Алқонун»	200
2	Абулқосим Фирдавсӣ	«Шоҳнома»	200
3	Садриддин Айнӣ	«Ёддоштҳо»	280
		«Ятим»	220
		«Қаҳрамони халқи тоҷик - Темурмалик»	150
4	Бобоҷон Ғафуров	«Тоҷикон»	200
5	Сотим Улуғзода	«Пири ҳакимони Машриқзамин»	150
6	Назирҷон Турсунов	«Таърихи тоҷикон»	400
7	Ф. Муҳаммадиев	«Куллиёт»	100
8	Ҷалол Икромӣ	«Асарҳои мунтахаб»	100
9	Абдумалик Баҳорӣ	«Бозгашт»	100
		«Соҳили мурод»	100
10	Раҳим Ҷалил	«Одамони ҷовид»	100
11	М.Ғ.Ғаниев	«MS'Word»	50
12	Ҳакими Раҳимзод	«Оила ва оиладорӣ»	150
13	Словарь	«Фарҳанги забони тоҷикӣ»	150
14	Газеты	«Ҷумҳурият»	270
		«Суғд»	280
		«Садои мардум»	200
		«Чархи гардун»	400
Всего			3800

Определение 1. Слог – это наименьшая произносительная единица речи, состоящая из одного или нескольких звуков, составленных на основе фонетического единства [123].

Согласно другим аналогичным определениям, слогом называется звук или сочетание звуков в слове, произносимое одним воздушным ударом [123].

Для изучения закономерностей таджикского языка в связи с понятием слог вводим понятие *слоговая структура слова*. Пусть W - любое слово, имеющее определенную последовательность букв. Сложив гласные с цифрой 1 и согласные с цифрой 0 (букву «й» мы называем согласной), слово W образуем слово – последовательность нуля и единицы $W_{0,1}^{*}$.

Определение 2. Реорганизацию $W \rightarrow W_{0,1}^{*}$ слова мы называем кодировкой слова W, а полученная запись $W_{0,1}^{*}$ представляет собой слоговую композицию или слово-образец W.

Определение 3. Структуру композиции $W_{0,1}^{*}$ назовем количеством образованных букв слова W или количеством символов (двойных знаков) в используемой записи $W_{0,1}^{*}$.

Определение 4. Состав двух слов, если их изображение одинаково в двух записях, назовем *единым,* в противном случае – *разным.*

Обсолютно ясно, что структуры могут быть едиными только при условии охвата одного и того же объема. Также известно, что для любого слова W возможно только одно совпадение в виде $W_{0,1}^{*}$. В свою очередь, по сути, в любом естественном языке $W_{0,1}^{*}$ для какого-либо $W_{0,1}^{*}$ он будет соответствовать нескольким словам W одновременно. Это значит, что разные слова с одинаковым количеством букв могут иметь один и тот же слоговый состав. Например, словам «дилшод», «кардам» и т. д. соответствует элемент «010010».

Дальнейшие результаты в четко выраженном виде получены на основе статистической обработки репрезентативной выборки, представленной в таблице 6.1.

Статистическая закономерность текстовых данных. В этой части работы рассматривается статистическая закономерность текстовых данных, полученных

путем репрезентативной выборки в объеме 1724472 слов, каждое из которых предварительно закодировано в форме $W_{0,1}^{*}$, описано в его слоговом составе (таблица 6.2).

Из этой таблицы 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 6.2. - 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

В большинстве $W_{0,1}^{*}$ было обнаружено только 274 различных частей. Выявлено, что 8 элементов охватывают 50%, а 23 части охватывают 75% таджикских текстов. В таблице 6.3 указано количество элементов в первом столбце в порядке убывания встречаемости, во втором столбце указана запись самого элемента, а в третьем столбце указан процент его встречаемости в текстах.

Таблица 6.3. - 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			

Установлено, что 51 элемент охватывает 90%, а 76 частей охватывает 95% таджикских текстов. Продолжение этой информации показано в таблице 6.4.

Таблица 6.4. - Частота встречаемости слов в слоговом составе (51%-95%)

№	$W^*_{0,1}$	%		№	$W^*_{0,1}$	%		№	$W^*_{0,1}$	%
24	0101011	0,923		42	011	0,421		60	0110101	0,190
25	01010010	0,895		43	010101011	0,404		61	0101001010	0,189
26	1	0,875		44	0101101	0,366		62	010111	0,189
27	010010101	0,869		45	010010011	0,348		63	0100101001	0,189
28	100101	0,810		46	101010	0,321		64	101001	0,188
29	01010100	0,734		47	0101100	0,317		65	0100110	0,187
30	010100101	0,717		48	1010101	0,306		66	1001011	0,185
31	0110	0,716		49	010101001	0,289		67	01001101	0,173
32	01001001	0,660		50	011010	0,281		68	01010010101	0,172
33	010101010	0,601		51	0100100101	0,279		69	1001001	0,170
34	101	0,556		52	0101010101	0,278		70	01001100	0,166
35	01101	0,554		53	101011	0,257		71	0101010100	0,164
36	010110	0,549		54	01010110	0,254		72	010001	0,160
37	0100100	0,533		55	010010010	0,248		73	0101001011	0,158
38	10010101	0,468		56	0100101011	0,244		74	010101101	0,144
39	0101010010	0,443		57	010100100	0,223		75	01000101	0,143
40	1001010	0,438		58	10011	0,195		76	0100101010	0,141
41	01010011	0,432		59	0100011	0,193				

Разновидности слогового состава. Каждый из 274 выявленных частей таджикских слов был «вручную» разделен на слоги (в соответствии с делением слогов тех таджикских слов, находящихся под влиянием тех или иных структур). В результате выялено всего 9 различных слоговых структур:

1, 10, 01, 010, 100, 0100

и

001, 0010, 00100.

В дальнейшем будем называть эти структуры *слоговыми моделями* или *формами слогов*. Первые шесть из них характерны для таджикского языка, а остальные три взяты из других языков.

Частота встречаемости (в процентах) указанных *слоговых моделей* в обработанных текстовых данных приведена в таблице 6.5.

Таблица 6.5. - Частота встречаемости (в %) слоговых моделей

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

Из таблицы видно, что двухбуквенные модели типа *да, ба, ро, на, ни, та, ме, ва, ки* (в кодовой записи - 0 и 1) и др., являются часто всречаемыми. Однако трехбуквенные слова типа *абр, илм, ашк, ишк, умр, орд* (в кодовой записи 100) встречаются особенно редко. Кроме того, слоги 001, 0010 и 00100, заимствованные из других языков, в таджикских текстах встречаются случайно (всего - 0,12%). Следует отметить, что двухбуквенные модели 10 и 01 вместе с трехбуквенным 010 составляют значительную часть слоговых моделей таджикского языка (88,05%). При этом 2, 3 – это среднее количество слогов в таджикских словах.

Алгоритм деления слов на слоги. В этой части работы дано концептуальное описание правил последовательности, реализация которых в виде компьютерных программ позволяет автоматически делить любые таджикские слова на слоги. Процесс сегментации основан на понятии частей слова и в основном использует 6 слоговых моделей.

Пусть любое таджикское слово W представляет собой определенную последовательность букв таджикского алфавита, а $W_{0,1}^{*}$ часть слова W, закодированная запись W в виде набора нулей и единиц. Напоминаем, что $W_{0,1}^{*}$ образуется из W путем соединения согласных букв с цифрой 0 и гласных букв с цифрой 1 в W. Этот алгоритм состоит из двух частей: в первой части осуществляется разделение $W_{0,1}^{*}$ на слоговые модели, во второй части полученный результат непосредственно используется для исходного образа слова W в виде набора слогов.

Часть 1. Таким образом, в таджикском языке существует 6 слоговых моделей: 1; 10; 01; 010; 100; 0100.

В первой части алгоритма, имеющего деление $W_{0,1}^{*}$ на слоговые модели, соблюдаются следующие правила:

1. Начало работы.

2. Ввод слова W.

3. Выполнение реорганизации $W \rightarrow W_{0,1}^{*}$.

4. Подсчет количества единиц k в записи $W_{0,1}^{*}$. Поскольку гласные кодируются цифрой 1, то цифра k на самом деле указывает на количество слогов в слове W.

5. Если $k \neq 1$, то, вероятно, запись $W_{0,1}^{*}$ состоит из одного слога, и этот слог выражается намерением $W_{0,1}^{*}$ с помощью одного из 6-ти ранее упомянутых слогов. Далее обратитесь к пункту 9.

Если $k \neq 1$, то следуйте пункту 6.

6. Если $k = 2$, то запись $W_{0,1}^{*}$ состоит из конкатенации двух шаблонов. Из каких именно моделей состоит $W_{0,1}^{*}$ определяется путем сопоставления $W_{0,1}^{*}$ с одной из различных записей двух моделей, которые получаются путем объединения одного из 6 шаблонов с каждым из 6 слоговых шаблонов. Вероятно, из 6-ти слоговых сочетаний можно получить 36 таких сочетаний пар. Затем перейдите к пункту 9.

Если $k \neq 2$, то переходим к пункту 7.

7. Если $k = 3$, то запись $W_{0,1}^{*}$ состоит из трех слогов. Какие именно слоги составляют $W_{0,1}^{*}$, определяется путем сопоставления $W_{0,1}^{*}$ с одной из разных записей с тремя слоговыми образцами, путем объединения одного из 6 образцов с каждым из 6 слоговых образцов и последующего добавления еще одного из 6 образцов к полученной записи. Вероятно, из 6-сложных сочетаний можно составить 216 таких трехобразных сочетаний. Затем перейдите к пункту 9.

Если $k \neq 3$, то следует следовать пункту 8.

1. Таким же образом узнаваем слоговой состав записи $W_{0,1}^{*}$ в случае $k > 3$, «но» $k \leq 8$, поскольку в настоящее время известно, что в таджикском языке нет слов с числом более 8 слогов.

9. Завершение.

Пример. Пусть $W =$ «хуршед». Тогда согласно пункту 3 реорганизация приведет к $W \to W_{0,1}^{*}$ ба $W_{0,1}^{*}$ = “010010”.

Потом согласно пункту 4 имеем $k = 2$.

Теперь в соответствии с пунктом 6 создается 36 двухмодельных записей:

1) 11 2) 110 3) 101 4) 1010 5) 1100 6) 10100

7) 101 8) 1010 9) 1001 10) 10010 11) 10100 12) 100100

……

18) 010100 19) 0101 ……

22) 010010 ……

25) 1001 ……

30) 1000100 31) 01001 ……

36) 01000100

Приведенное в качестве примера слово «хуршед» кодируется с помощью нулей и единиц отождевляется с 22-й записью. В соответствии с этим закодиров анное слово находит свое слоговое представление.

$W_{0,1}^{*}$ («хуршед») = 010 ⊕ 010, где ⊕ - признак агглютинации, беспробельного соединения (слипания) одного слогового состава с другим.

Часть 2. После разделения $W_{0,1}^{*}$ по слоговым моделям деление начального W слова производится простым способом. Из первой части алгоритма можно записать в память большое количество букв, образованных 1-м слогом, 2-м слогом и т. д. Эти цифры используются для разделения слогов уже в первом слове W. Таким образом, в данном примере при делении $W_{0,1}^{*}$ («хуршед») образовалось 2 слога, причем и первый, и второй слоги состоят из 3 букв. Следовательно, при делении слова = «хуршед» получаем следующий результат: «хур-шед».

Разновидности слогов таджикского языка. На основе алгоритма, представленного в предыдущем разделе, и обработанной на его основе компьютерной программы было проведено статистическое исследование разновидности слогов таджикского языка.

На 3800 случайно выбранных страницах было отмечено 1724472 слова. Из 3259 различных производных слогов было установлено, что более 2044 его различных слогов были идентифицированы ранее [125]. Причина основного различия в окончательных результатах заключается в том, что размер выборки, подлежащей статистической обработке в данной диссертации, более чем в 20 раз превышает размер выборки.

Исследование выборки выявило возможность статистического распределения слогов в текстах на таджикском языке, эмпирическую корреляцию ν = ν (n) между номерами каждого из 3259 различных слогов по порядку словоизменения и соответствующим числом ν в процентах склонения соответствующего слога.

Установлено, что 41 слог, приведенный в таблице 6.6, охватывает 50% таджикских текстов.

Таблица 6.6. - Частота встречаемости таджикских слогов

№	слог	встр	№	слог	встр	№	слог	встр	№	слог	встр
1	и	4,210	12	ва	1,500	23	му	0,968	34	ҳа	0,673
2	да	2,447	13	бо	1,355	24	ли	0,951	35	са	0,647
3	ро	2,347	14	дар	1,325	25	а	0,914	36	за	0,611
4	ба	2,235	15	ди	1,277	26	со	0,833	37	ло	0,602
5	ҳо	2,022	16	ки	1,189	27	си	0,823	38	во	0,562
6	ни	1,827	17	о	1,156	28	но	0,766	39	ла	0,552
7	на	1,796	18	мо	1,149	29	ми	0,760	40	ё	0,548
8	ти	1,665	19	до	1,112	30	би	0,727	41	хо	0,523
9	ри	1,612	20	ра	1,077	31	то	0,722			
10	та	1,552	21	ма	1,071	32	я	0,697			
11	ме	1,508	22	аз	0,986	33	ин	0,693			

Кроме того, выявлено что 148 слогов охватывают 75% таджикских текстов, 418 слогов охватывают 90%, а 683 слога – 95% текстов. Отметим, что все остальные слоги охватывают лишь 5% текстов, начиная с номера 684 по 3259. Поэтому появление каждого отдельного слога из такой группы – событие необыкновенное.

Статистические закономерности слогового состава таджикского языка. На основе формулы К. Шенона определяется информативность слога в словоформах и словоупотреблениях на примере таджикских текстов, насчитывающих более 20 миллионов слов. По ранее полученным данным, в таджикском языке выявлено 3259 различных слогов. При общей обработке текстов было выявлено 64458 различных словоформ из 54 162492 употребленных слов.

Учитывая, что $X = \{x_1,\ldots,x_n\}$ - конечное множество взаимно несовместимых событий, вероятность $p_1,\ldots,p_n$ их возникновения условнао зависит от $p_1+\ldots+p_n=1$. Тогда формула К. Шеннон (формула 6.1) определяет среднее количество информации для одного события.

$$H(p_1,\ldots,p_n) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i \qquad (6.1)$$

В данном случае для расчета информативности слогов внутри слова использовалась та же формула, а в качестве примера был выбран таджикский язык.

Согласно таблице 6.6, в таджикском языке определено 3259 различных слогов. Информация об их частотности, была получена путем обработки набора текстов, состоящего из 54162492 словоупотреблений, среди которых было выявлено 64458 различных словоформ (таблица 6.7.).

Таблица 6.7. - Распределение словоформ и употребление количества слогов

Количество слогов	Количество словоформ	Доля в %	Количество словоупотреблений	Доля в %
1	1389	2,15	15310456	28,27
2	10751	16,68	14673494	27,09
3	22160	34,38	14510438	26,79
4	18880	29,29	7109273	13,13
5	8313	12,90	2000983	3,69
6	2344	3,64	475889	0,88
7	522	0,81	70825	0,13
8	99	0,15	11134	0,02
Всего:	64458	100	54162492	100

В вышеприведенной таблице в 1-м столбце указано количество слогов, составляющих таджикские словоформы. В столбцах 2 и 4 указано количество

словоформ и употребление слов с тем или иным числом слогов. Эти данные выражены в графах 3 и 5 в процентах от общего количества словоформ и словоупотреблений.

Из таблицы следует, что среди словоформ больше трехсложных и четырехсложных слов, при употреблении слова – больше односложных, двухсложных и трехсложных слов.

Пусть $WF^{(l)}$ подмножество словоформ с l слогами, а $l = \overline{1,8}$ и m - порядковое число слогов в слове с l слогами. Информативность m-го слога рассчитывается двумя способами. Сначала из совокупности разных форм слова без учета частоты их встречаемости, затем из совокупности всех случаев употребления слова другими словами, из совокупности одинаковых словоформ, но уже с учетом частота их встречаемости.

Каждый слог последовательно извлекается из списка всех слогов и рассчитывается относительная частота его появления на m-й позиции слога в подмножестве словоформ. По формуле 6.2 информативность m-го слога определяется по подмножеству l словоформ $WF^{(l)}$.

$$H(\lambda_1^m,\ldots,\lambda_{3259}^m) = -\sum_{i=1}^{3259}\lambda_i^m \cdot \log_2 \lambda_i^m \qquad (6.2)$$

При этом индекс i используется для обозначения номера слога в списке из 3259 слогов, отсортированного по убыванию частоты их появления в наборе текстов, и λ_i^m - относительной частотности i-го слога в m-й позиции l -форм слов находится подмножество.

На основе исследований необходимо отметить, что информативность первого слога достигает максимума при односложных словоформах и употреблении односложных слов, а затем, с увеличением числа слогов, резко снижается в словоформах с восемью слогами и использование таких слов сокращается и принимает наименьшее значение.

Результаты расчета информативности по формуле 6.2 представлены в таблицах 6.8 и 6.9.

Таблица 6.8. - Вес m-го слога в форме слова, состоящего из 1 слога

		Порядковый номер *m*-слога							
		1	2	3	4	5	6	7	8
Размер используемого слова	1	10,25							
	2	9,21	9,41						
	3	8,46	7,76	7,83					
	4	7,88	7,86	6,87	6,45				
	5	7,26	7,55	6,97	5,99	5,12			
	6	7,16	7,55	7,26	6,55	5,59	4,27		
	7	6,59	6,72	6,84	6,21	5,61	4,67	3,55	
	8	5,46	5,32	5,86	5,82	5,09	4,56	4,19	3,48

Таблица 6.9. - Вес m-го слога в употребленном слове, состоящем из 1 слога

		Порядковый номер *m*-слога							
		1	2	3	4	5	6	7	8
Размер используемого слова	1	7,26							
	2	6,15	5,20						
	3	5,67	5,43	4,03					
	4	5,32	5,49	4,87	3,03				
	5	5,23	5,39	5,30	4,43	2,41			
	6	5,04	5,28	5,47	5,05	4,19	1,81		
	7	4,93	5,01	5,29	5,14	4,81	3,59	1,46	
	8	4,38	3,91	4,90	4,82	4,42	3,61	3,24	1,65

Для этого путем обработки набора упомянутых текстов был создан список словоформ с указанием их частотности, причем каждая словоформа выражается как связь слогов. Кроме того, выявляется общая тенденция снижения информативности слога с увеличением его порядкового номера.

Графические иллюстрации приведенные на рисунке 6.1. показывают информативность слогов, отдельно для односложных и до восьмисложных слов.

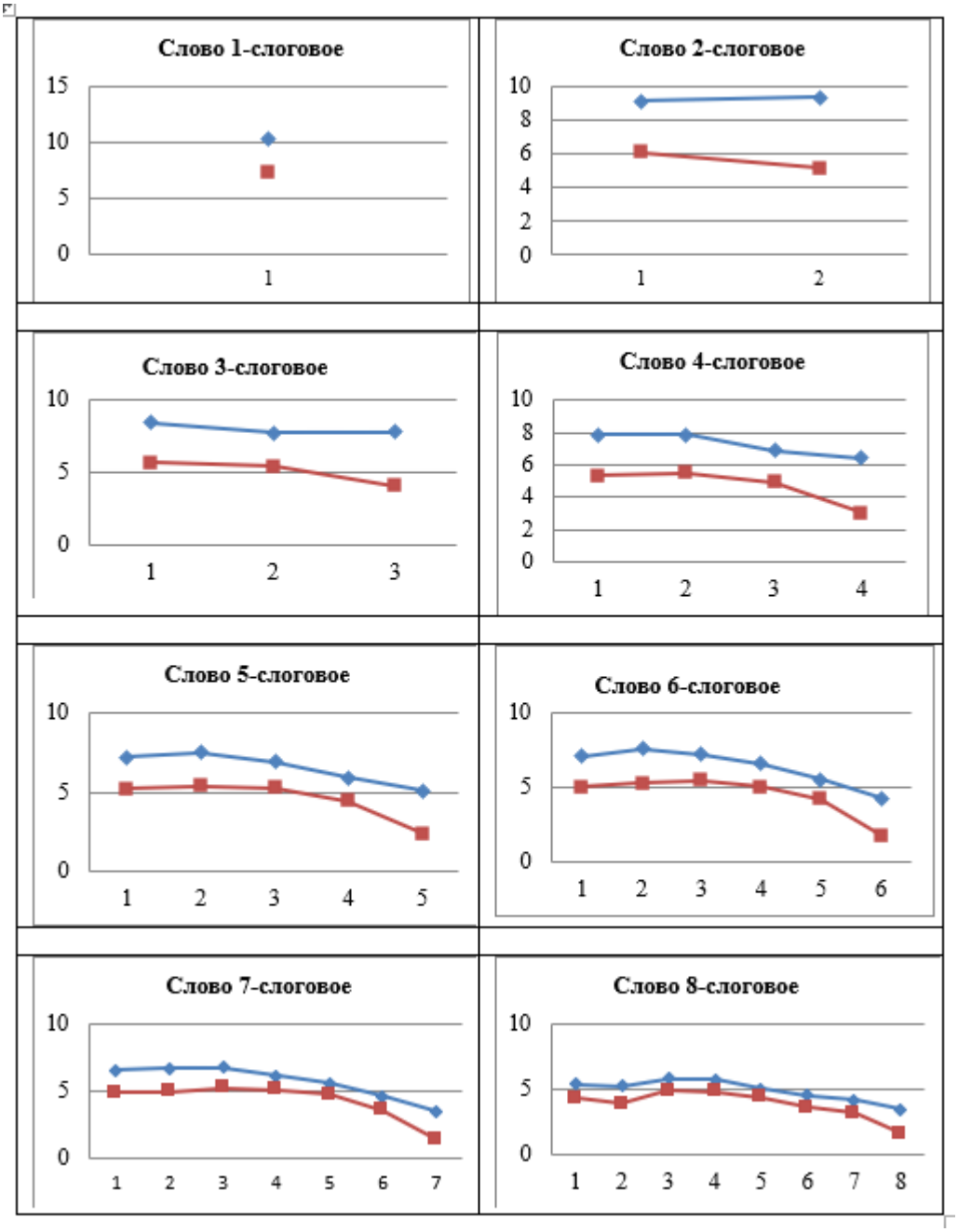

Рисунок 6.1. - Статистические закономерности слогового строя таджикского языка

В этом разделе описана последовательность процедур, используемых при расчете информативности слога в зависимости от его положения в словоформе. Другими словами, речь идет о содержании первого, второго и, наконец, последнего слога в составе образованного слова. Следует отметить, что рассматриваемая задача решается отдельно для подгрупп, состоящих из словоформ с одинаковым числом слогов. В таджикском языке таких компонентов, от односложных до восьмисложных подмножество словоформ.

На риснках кривая, представляющая информационное содержание слогов в словоформах, показана синим цветом и расположена выше, чем аналогичная кривая, представляющая употребление слов, показанная красным цветом. На этих изображениях по оси абсцисс показаны порядковые номера слогов, а по оси ординат – информативность слогов.

§ 6.2. Основы компьютерного синтеза таджикской речи

Фонетика таджикского языка. В таджикском литературном языке шесть гласных фонем: [и], [э], [а], [у], [ӯ], [о]. Выделяют три основных признака, основанных на характерных особенностях гласных: ряд, уровень языка и участие губ. Разница между гласными определяется движением языка в вертикальном направлении, а также вперед и назад. В таджикском языке различают три ряда гласных: передний ряд – [и], [э], [а], задний ряд – [о], [у] и смешанный ряд – [ӯ].

Гласные сегодня отличаются не длительностью и краткостью произношения, а уровнем устойчивости. С этой точки зрения гласные делятся на две группы: устойчивые гласные [э], [ӯ], [о] и нестойкие – [и], [у], [а]. Длительность гласных устойчивой группы изменяется сравнительно меньше в зависимости от изменения фонетических условий, отчасти в зависимости от такта. Долгота гласных неустойчивой группы, наоборот, очень сильно меняется в зависимости от изменения фонетических условий. В ударном и безударном состоянии закрытого слога они равны задержке гласных устойчивой группы, но в безударном открытом слоге они сильно сокращены.

Контраст между двумя группами гласных - устойчивыми и неустойчивыми – отчетливо заметен только в безударных открытых слогах.

Сравнительно четкое сравнение фонем таджикского языка с русскими гласными можно определить следующим образом.

Гласная таджикская фонема [и] более открытая, чем русская [и], при изменении фонетических условий (характера слога, качества соседних звуков, ударения и т.д.) она сильно изменяется как по качеству, так и по длительности.

Гласная [э] - относительно закрытый звук. Благодаря продвинутому уровню языка он считается средним между закрытым русским вариантом [э] и средним. Он входит в группу устойчивых гласных. При изменении фонетических условий он меняется сравнительно меньше, но сохраняет свою устойчивость.

Гласная [а] по звучанию отличается от русской [а] относительно передней артикуляцией. Благодаря влиянию разных фонетических условий оно меняется как по свойству, так и по длительности.

Гласная [у]отличается от русского произношения [у] своей близостью, но не вполне с ним согласуется и входит в число неустойчивых звуков. Оно также сильно меняется при изменении фонетических условий, особенно в плане длительности.

Гласная [о]существенно отличается от русской как по характеру, так и по долготе. Во-первых, она более открыта, чем русская гласная [о], во-вторых, в отличие от русской [о] не изменяется в любом фонетическом положении.

Гласная [ӯ] не имеет аналога в русском языке.

В литературном таджикском языке 23 согласных. Они представлены в таблице 6.10.

Таблица 6.10. - Согласные фонемы таджикского языка

		Губной	Губозубной	Переднеязычный	Межъязыковой	Заднеязычный	Язычковый	Гортанный
Слитный	неносовой	**п б**		**т д**		**к г**	**қ**	
	носовой	**м**		**н**				
	аффриката			**ч ҷ**				
Щелевой	одноартикуляционный		**ф в**	**с з**	**й**		**х ғ**	**ҳ**
	двуартикуляционный			**ш ж**				
	боковой			**л**				
дрожащий				**р**				

Фонема [ж] имеет сравнительно узкий диапазон употребления. Глухие слитные согласные [п], [т], [к] отличаются от соответствующих русских звуков не

слишком большим звуком, который известен в начале слова перед гласными. Взрывные согласные [б], [д], [г] в отличие от русских в конце слова не оглушаются.

Согласная [қ] в русском языке не имеет своего аналога. Звонкая аффриката неязыковой двуартикуляционная переднеязычная [ҷ] как самостоятельная фонема в русском языке не существует.

Звонкий язычковый согласный звук [ғ] и согласный глухой гортанный звук [ҳ] также не имеют своего русского аналога. Преднеязычные двуартикуляционные щелевые согласные [ш] и [ж] акустически отличаются от соответствующих русских звуков повышенной мягкостью. Боковой переднезычный щелевой согласный [л] звучит как средний русский звук между [л] и [ль]. Согласные [в] и [й] отличаются от соответствующих русских звуков своей звучностью, то есть небольшой примесью шума в голосе. Остальные согласные ([м], [н], [ч], [ф], [с], [з], [р]) не имеют заметного отличия от соответствующих русских звуков.

Формирование слого-звукового источника. Слогово-звуковая база состоит из 23259 таджикских слогов, составленных двумя профессиональными чтецами. Формирование источника слог-голос предполагает комплексное решение сложных задач, связанных с реорганизацией слогов в цифровой вид и дальнейшей обработкой звукового файла с помощью программы Cool Edit Pro. Это связано с выбором скорости дискретизации; понижения шума; стабилизации голосового сигнала; разбработки звуковых символов (рис. 6.2).

Рисунок 6.2. - Оцифрованная версия слога

Отметим, что при псоздания реального слогово-голосового источника большая часть голосов была создана голосом профессионального диктора

Каримова Обиджона (телерадиокомпания СМ-1) через микрофон, произносящего различные слоги таджикского языка. Далее возникла необходимость «стандартизации» произношения слогов, которая потребовала редактирование голосов с помощью компьютерной программы Cool Edit Pro.

Редактирование осуществлялось по трем признакам - тону, определенной скорости вибрации голосовых связок, высоте, зависящей от интенсивности голоса и его скорости, и длительности произношения. Отредактированные слоги в аудиоверсии сохраняются в файлах формата WAV.

При произнесении слогов учитывались следующие признаки (описания) голоса: высота, скорость, тон.

Повышение громкости – субъективное восприятие силы звука (абсолютная величина слухового ощущения). Повышение в основном зависит от звукового давления и скорости звуковой вибрации. На повышение голоса также влияют тембр голоса, длительность воздействия голосовых дрожаний и другие факторы.
Абсолютной единицей шкалы громкости является сон. Подъем в 1 сон – это подъем непрерывного тона скоростного синусоидального звука частотой 1 кГц, создающего звуковое давление 2 МПа.

Уровень повышения громкости - относительная величина. Он измеряется в фонах и количественно равен уровню звукового давления (в децибелах – дБ), которое создается синусоидальным тоном со скоростью 1 кГц такого увеличения, как измеряемый голос (равный этому голосу) .

Звук произношения – это особенное ощущение произношения отдельных голосовых источников или группы голосов, скоростной спектр и характер произношения основного тона и обертонов, адекватное воспроизведение прямого звукового поля и скоростного спектра пространственных звуков.

Тон используется как важное средство передачи голоса (выразительность голоса). Главными особенностями произношения являются баланс тонов, описание амплитудной и фазовой скорости, различные виды искажений, нелинейные искажения, процессы перехода, эффект присутствия. Примерами источников

искажений являются громкоговорители, голосовые органы, устройства управления и обычные телефоны.

Тон голоса – одна из единиц измерения громкости голоса. Громкость – это субъективное качество слухового ощущения наряду с высотой и тоном звука, которое позволяет расположить все голоса в таблице от низкого до высокого. Чистота тона высокого голоса зависит, прежде всего, от скорости (с увеличением скорости высота звука увеличивается), а также от его интенсивности. Громкость со сложным спектральным составом зависит от распределения энергии на поверхности шкалы скоростей. Высота голоса измеряется в *мелах (*от слова «мелодия». Звук записывается со скоростью 1 кГц и звуковым давлением 2*10-3 Па на высоте 1000 мел. В диапазоне 20 Гц – 9000 Гц помещается почти 3000 мел. Измерение высоты свободного звука основано на способности человека выравнивать высоту двух голосов или на их соотношении (насколько один голос выше или ниже другого).

Описание слого-звукового источника. Слого-звуковой ресурс занимает на жестком диске 263 Мб и в среднем 40 Кб памяти на один слог. Временной интервал произнесения одного слога меняется в диапазоне 250-400 мс. Межсложная и межсловная длительность может составлять 20–200 мс и 200–2000 мс.

Например, слоги «а», «о», произнесенные мужским голосом, составили всего 13 Кб, а слоги «шахс», «рахш» – 60 Кб. Произношение слогов «а», «и» женским голосом отдельно требовало 16 Кб памяти, тогда как слоги «*заъф*» и «*нашр*» требовали 65 Кб памяти на каждый слог. Краткая информация представлена в таблице 6.11.

Таблица 6.11. - Общее описание слого-звукового источника

Звук произношения	Общее количество слогов	Общее количество памяти (Мб)	Средний объем 1 слога (Кб)	Минимальный объем 1 слога (Кб)	Максимальный объем 1 слога (Кб)
Мужской	3259	130	40	13	60
Женский	3259	133	41	16	65
Всего:	6518	263			

Слого-звуковой источник состоит из звуковых файлов в формате WAV, произносящих 3259 слогов женским и мужским голосом. Описание источника приведено в таблице 6.6. Среднее время произнесения одного слога от слога-звука источника составило 250-400 мс.

Структура WAV-файла. Рссмотрим простейший WAV-файл (Windows PCM). Он состоит из двух полностью разделенных кругов. Один из них - это заголовок файла, другой - диапазон данных. Заголовок файла содержит следующую информацию:

- размер файла;
- количество каналов;
- скорость дискретизации;
- количество битов на сэмпл (небольшой оцифрованный звуковой фрагмент).

Но для полного понимания содержания описания заголовка файла необходимо сказать об информации и оцифровки голоса.

Звук представляет собой движение, которое при оцифровке приобретает вид лестницы. Этот вид образуется потому, что компьютер может выполнить определенную амплитуду (увеличение) за любой короткий период времени, и этот короткий период не всегда является коротким. Длительность этого интервала также определяет уровень дискретизации. Например, наш файл имеет частоту дискретизации 44,1 кГц. Это означает, что короткий интервал времени равен 1/44100 секунды (от величины Gs = 1/c). Современные звуковые карты поддерживают частоту дискретизации до 192 кГц.

Если говорить о том, что связано с амплитудой (подъемом голоса за короткий промежуток времени), то от этого зависит точность голоса. Амплитуда представлена числами, записанными в памяти (файле) 8, 16, 24, 32 бита (теоретически может быть и больше). Поскольку 8 бит = 1 байт, то амплитуда за короткий промежуток времени в памяти (файле) может занимать 1, 2, 3, 4 байта соответственно.

Таким образом, чем больше места в памяти (файле) занимает число, тем больше диапазон сущности этого числа, то есть амплитуды:

- 1 байт - 0..255;
- 2 байта - 0..65 535;
- 3 байта - 0..16 777 216;
- 4 байта - 0..4 294 967 296.

В общем варианте суть амплитуд размещена последовательно. В стерео, например, суть амплитуд сначала уходит в левый канал, потом в правый, потом в левый и так далее. Набор амплитуд и коротких временных интервалов называется *сэмплом* (элемент или часть аудиоданных в цифровой форме).

Алгоритм произношения слова. Сначала исходное слово обрабатывается с помощью алгоритма и разбивается на слоги. Затем для каждого слога слогово-голосового источника выделяется подходящий голосовой файл, а потом с их помощью синтезируется голосовое произношение слова путем размещения межсложного интервала.

Алгоритм произношения чисел. Отметим сразу, что следующий алгоритм направлен на произнесение небольших целых положительных чисел, или равных 1012 – 1 = 999 999 999 999. Вполне понятно, что произношение чисел, больше заданных, требует привлечения дополнительных степеней чисел (*миллиард, триллион и т. д.).*

Наш исследуемый алгоритм состоит из двух частей:

- реорганизации цифрового знака в текстовую форму;
- применения к результату реорганизации, в любом случае к типу текста, полученному алгоритмом произношения слов.

Первая часть – реорганизация числа в его текстовой записи работает с тем же источником информации «число – текстовые изображения». Данный ресурс состоит из трех блоков:

1. Блок цифр от 0 до 19 текст написан на таджикском языке.
2. Десятичный блок состоит из двух вспомогательных регистров с индексами 0 и 1, предназначенных для использования пустых элементов;

3. Блок 9 сотен состоит из вспомогательного регистра с индексом 0, предназначенного для использования пустого элемента. В этих элементах массив цифровое написание сотен с его текстовым написанием выполнен на таджикском языке.

Ниже приведено описание алгоритма произношения цифр.

Часть 1. Последовательность действий по обновлению цифрового знака при написании текста состоит из следующего правила:

1. Старт

2. Введение числа N

3. Написание цифрового текста txt$ = " ".

4. Проверка условия:

если N равно нулю, то txt$ = numb1(0), переход к пункту 8.

5. Подготовка к описанию цифрового текста N.

H1 = H

Правило деления первых трёх степеней первичных (сотни, десятки и единицы) числа N.

triad(1) = N1 - Int(N1 / 1000) * 1000

Правило разделения трех вторых степеней (стотысячных, десятитысячных и единиц тысячных) числа N.

N1 = Int(N1 / 1000)

triad(2) = N1 - Int(N1 / 1000) * 1000

Правило разделения третьей степени (стомиллионной, десятимиллионной и миллионной) числа H.

N1 = Int(N1 / 1000)

triad(3) = N1 - Int(N1 / 1000) * 1000

Правило деления трех четвертых степеней (стомиллиардной, десятимиллиардной и одной миллиардной) числа N.

N1 = Int(N1 / 1000)

triad(4) = N1 - Int(N1 / 1000) * 1000

6. Проверка степени числа N.

N1 = Int(N1 / 1000)

7. Для улучшения работы алгоритма выразим степень числа N от низшей до высшей степени числами 1, 2, 3, 4 и с помощью обычной записи i%={1, 2,3,4}, отметим соответствующие циклы .

7.1. когда i%=1 (однако мы имеем дело с первыми тремя уровнями числа N), сначала в triad (i%) выделяется число сотни k3% - номер элемента в массиве сотня; если k3% = 1, то значение «один сто» принимается как элемент массива numb3; если k3% ≠ 1, то запись этого элемента определяется соответствующим номером элемента массива numb3, если он не делится на одну и ту же тройку чисел(i%) 100, то в конце записи элемент k3% массива numb3 включает суффикс «у»; в тройке чисел triad(i%) отделяется десятичное число k2%; если k2% равно 1, то в числовой triad(i%) необходимо дополнительно выделить количество единиц и затем взять запись соответствующего элемента между числами 10-19 из массива numb1; если k2% не равно 1, то запись записи этого элемента берется из массива numb2; если triad(i%)не делится на 10, то в конце элемента k2% массива numb3 добавляется суффикс «u» (при k2%=3 – суффикс «*ю*»); в числе triad(i%) отделяется номер единицы; из массива numb1 выбирается запись записи элемента k1%-a;

7.2. если i%=2 (однако мы имеем дело с единицами, десятками и сотнями тысяч N чисел), то выполняется операция аналогично пункту 7.1.;

В конце записи добавляется слово «тысячу» или «тысяча»;

7.3. если i%=3 (однако мы имеем дело со степенями единиц, десятками и сотнями миллионов чисел N), то выполняется та же операция, что и п. 7.1.; в конце записи добавляется слово «миллиону» или «миллион»;

7.4. если i%=4 (однако мы имеем дело со степенями единиц, десятичными дробями и сотнями миллиардов чисел N), то операции аналогичные пункту 7.1. выполняется; в конце записи добавляется слово «миллиарду» или «миллиард»;

8. Все полученные записи упорядочены соответствующим образом и образуют число N, записанное прописью. Полученный результат присваивается переменной txt$.

9. Конец.

Часть 2. Значение в первой части производной переменной txt$ – текстовое написание данного числа N произносится с помощью алгоритма произношения слов.

Произношение символов. Общие символы встречаются в таджикских текстах. Список из 14 их наиболее часто используемых символов представлен в таблице 6.12. Для произнесения таких символов создан специальный источник символов – голос.

Таблица 6.12. - Символы и их текстовая запись

№	Символ	Вид слова
1	€	Евро
2	$	Доллар
3	%	Фоиз
4	(	Қавси кушод
5	)	Қавси пӯшида
6	*	Зарб
7	+	Ҷамъ
8	-	Тарҳ
9	/	Тақсим
10	<	Хурд
11	=	Баробар
12	>	Калон
13	§	Параграф
14	№	Рақам

Вполгне очевидно, что этот список можно расширить, добавив другие, менее распространенные символы.

§6.3. Проектирование и разработка алгоритмов синтеза речи

Алгоритм безударного произношения текста. По своей сути этот алгоритм представляет собой реализацию последовательности алгоритма произношения слов с добавлением межсловного пространства. Далее этот алгоритм описывается на уровне блок-схемы, работающей в виде компьютерной программы, обеспечивающей возможность контроля потока безударного произношения текстовых сообщений (рисунок 6.3.).

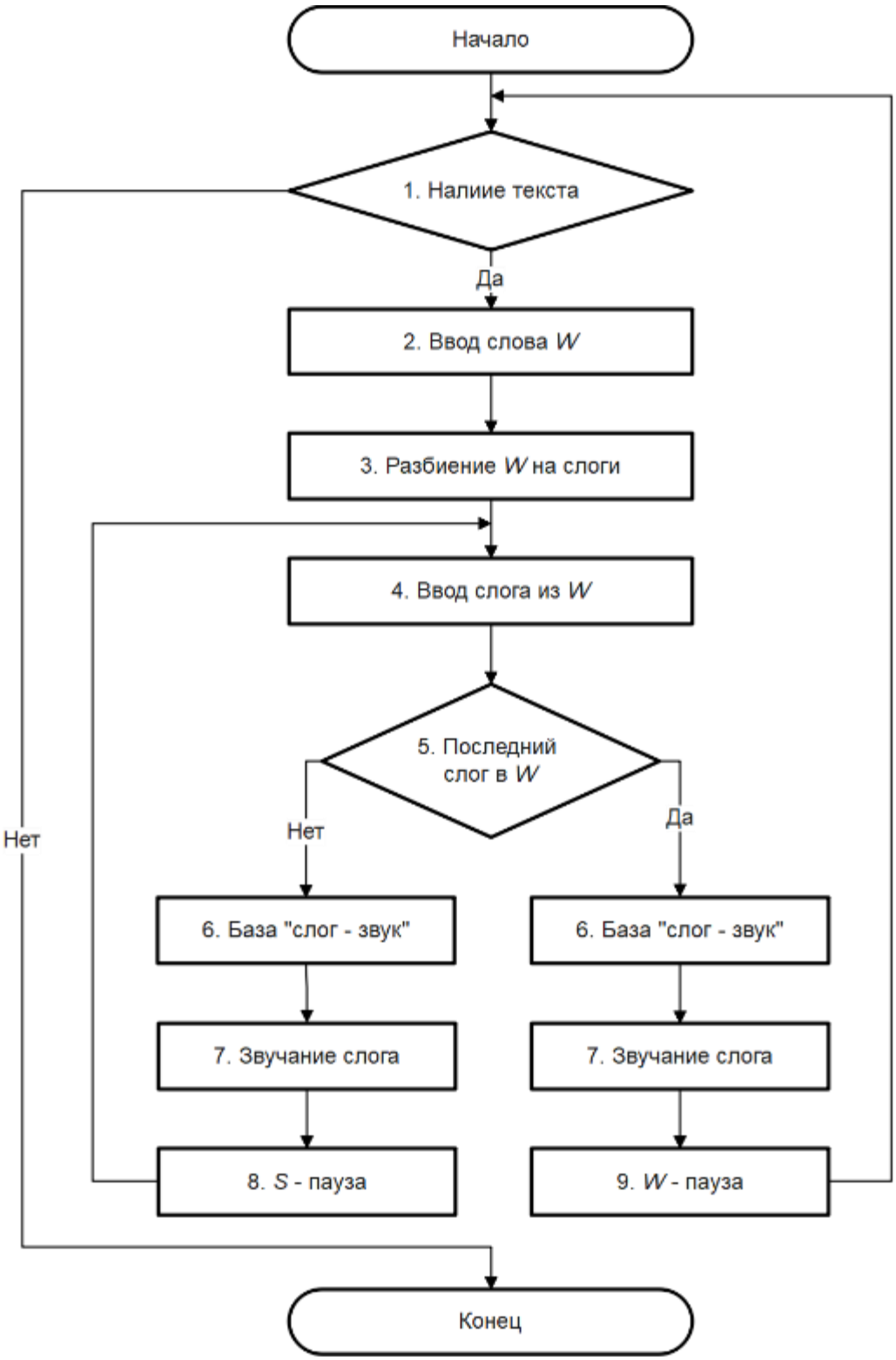

Рисунок 6.3. - Алгоритм безударного произношения слова

Начало – запуск программы произношения.

В блоке 1 проверяется наличие текста. Если его нет, то – конец.

В противном случае он переходит к блоку 2, в котором из текста находится слово W для следующего анализа.

В блоке 3 слово W с помощью алгоритма, выраженного в п. 1.6. подвергается слогоделению.

В блоке 4 в слове W находятся очередные слоги.

В блоке 5 определяется, являются ли найденные слоги последними в слове или нет. В обоих случаях обращается к слого-голосовому источнику, в котором содержится список из 3259 слогов, полученных в ходе статистической обработки случайной выборки объемом 3800 страниц. Каждому слогу определено его произношение.

Учитывая это, в блоке 6, в любом случае, в источнике слог– голос появляется звук, соответствующий данному слогу, который исполняется в блоке 7.

Далее, в зависимости от того, находится ли произносимый слог в последнем слове W или нет, устанавливается межсложный интервал (S – интервал), см. блок 8, или межсловный (W – интервал), см. блок 9. В первом случае временной интервал между моментом окончания произнесения предыдущего слога и началом произнесения следующего слога оказывается меньше временного интервала между моментом окончания произнесения предыдущего слова и началом произнесения следующего слова.

Оказалось, что экспериентальное оценивание S-pause = 20 мсек ва W-pause = 200 мсек достаточно для распознавания голоса компьютерного произношения текстовых данных.

По окончании межсложного интервала происходит возврат к блоку 4, а с окончанием межсловного промежутка - к блоку 1. Алгоритмические правила продолжаются до окончания обработки всего текста.

Алгоритм произношения ударного текста. По сути, этот алгоритм представляет собой реализацию последовательности алгоритма произношения слов, где показано положение ударных слогов с добавлением межсловных интевалов.

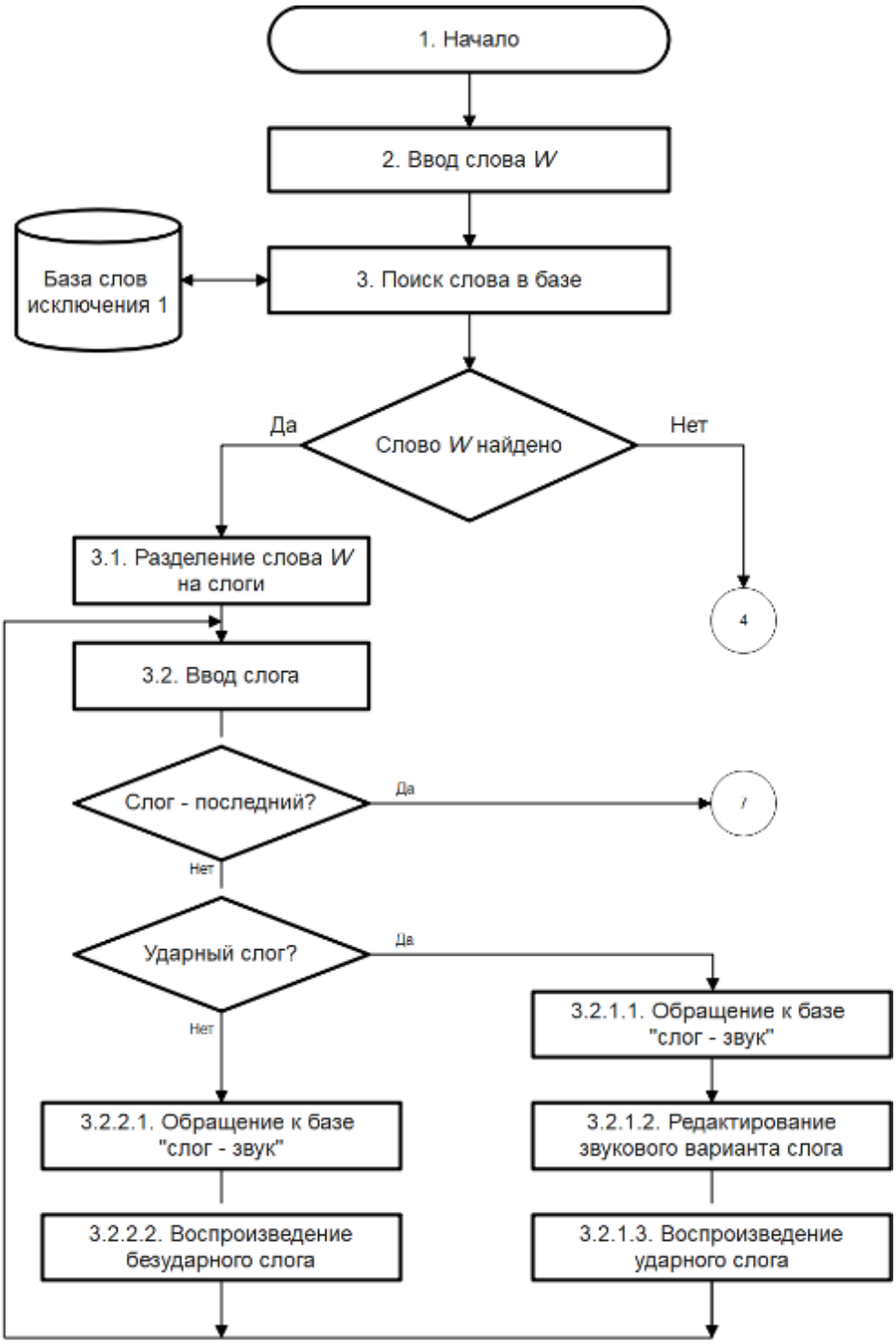

Рисунок 6.4. - Алгоритм произношения ударного текста (часть 1)

Далее этот алгоритм описывается на уровне блока – схема произношения *одного таджикского слова*, (рисунок 6.4.), операция, которая в виде компьютерной программы позволяет возможности направлять процесс произношения ударных текстовых данных с учетом соблюдения межсложных и межсловных интервалов в произносимых текстах.

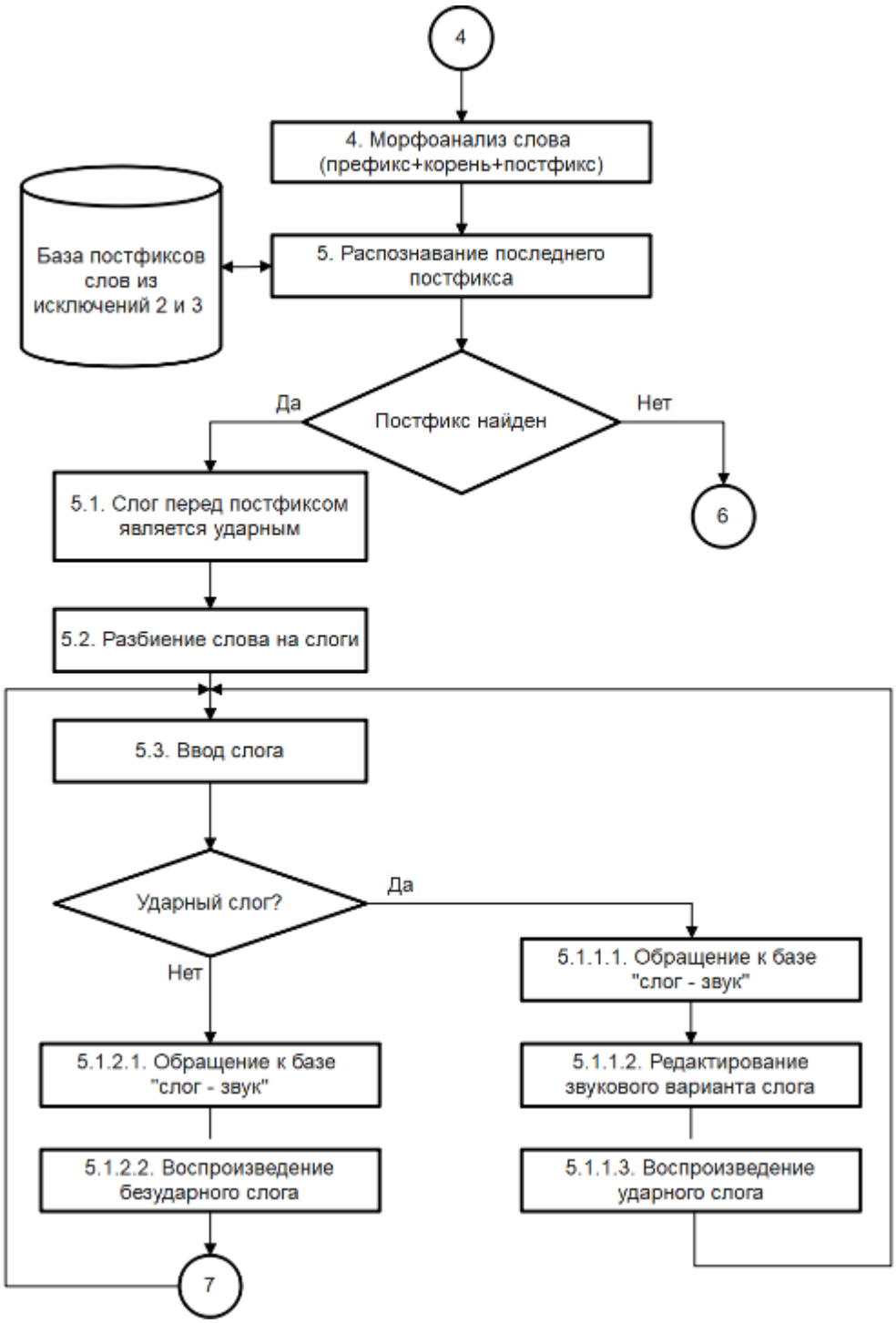

Рисунок 6.5. - Алгоритм произношения ударного текста (часть 2)

Чтобы не перегружать блок-схему, на рисунках 6.5. и 6.6. интервалы не показаны.

1. Старт.

2. Введение слова.

3. Проверка условия: введенное нами слово включено в источник исключения да или нет.

Если да (в любом случае слово включено в исходник и в нем последний слог считается ударный), тогда 3.1. деление слов на слоги;

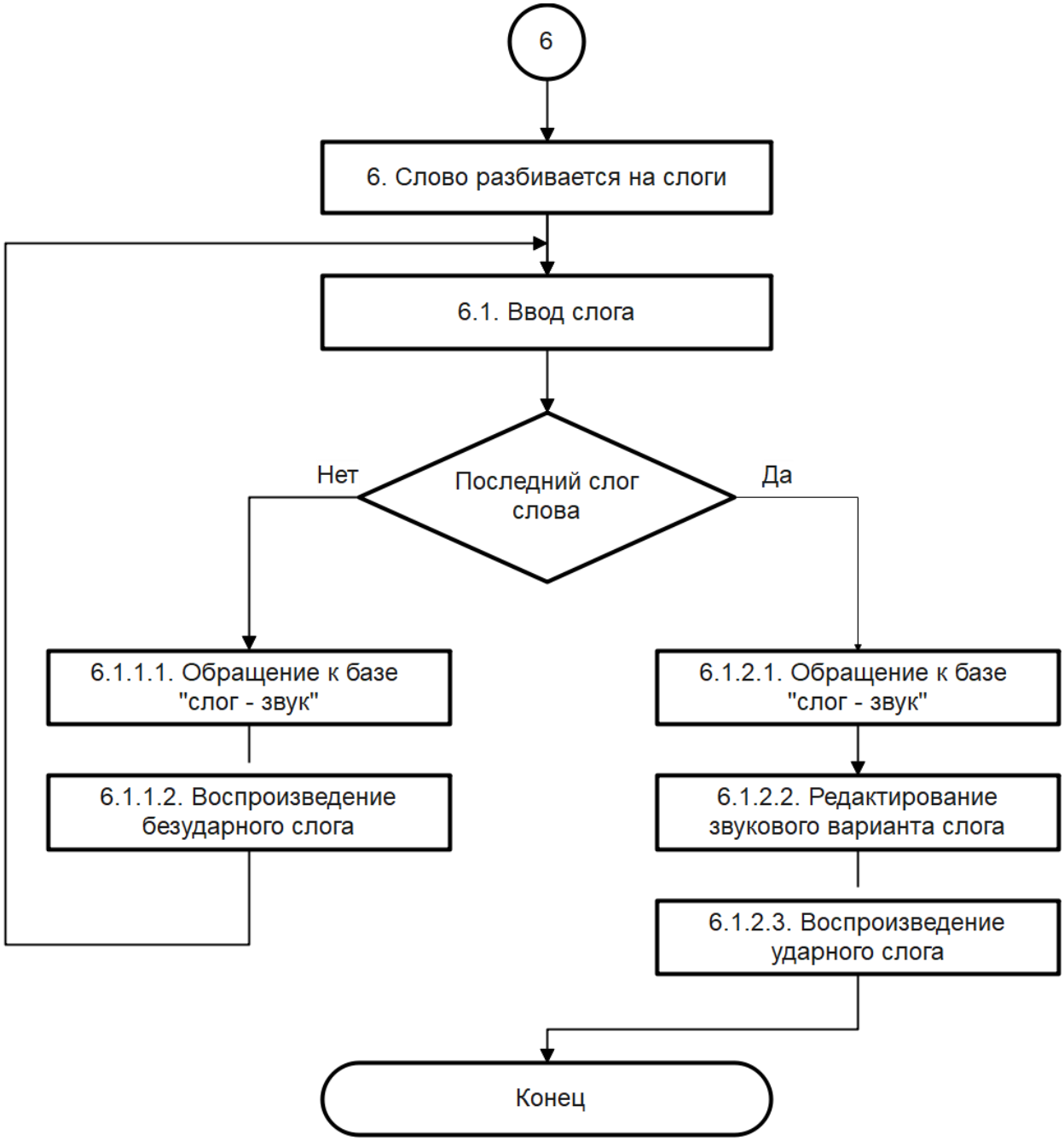

Рисунок 6.6. - Алгоритм произношения ударного текста (часть 3)

3.2. включение следующего слога *и*, если этот слог не последний, то его голосовой образ возникает в источнике слог - голос, однако в блоке 3.2.2.1 и в 3.2.2.2 отмеается его безударное произношение.

Если этот слог последний, то он является ударным, и его обработка производится в блоках 3.2.1.1 и 3.2.1.1 и в 3.2.1.1 с правилами ударного произношения.

Если в пункте 3 определится, что вводимого слова в источнике исключения 1 нет, то дальнейшая его обработка осуществляется через блок 4. Здесь проводится морфемный анализ слова и делается попытка определить, подпадает ли слово под исключения 1 и 2 п. 1.8.

В блоке 5 это реализуется путем определения последнего суффикса в слове (в случае сложного суффикса).

Если последний суффикс есть в исключениях 2 и 3, то обработка слова переходит в блок 5.1 – 5.3 и далее делится на 2 типа: 5.1.2.1, 5.1.2.2 или 5.1.1.1, 5.1.1.2 и 5.1. .1.3 в зависимости от этого ударный или безударный следующий слог.

В случае пункта 5 суффикс слова не относится к исключению 2 или 3, слово не входит ни в одно исключение и ударение стоит на последнем слоге.

Следующая обработка слова происходит в блоке 6. Слово делится на слоги. Слоги от первого до последнего считаются безударными и произносятся соответствующим образом. Если слог последний, то он ударный, и произносится он соответственно.

7. Конец.

Алгоритм морфемного произношения слова. В этом разделе предполагается, что у нас уже есть морфемный словарь таджикского языка с 3 источниками информации: «*приставка-звук*», «*корень-звук*», «*суффикс-звук*».

Далее на уровне блок-схемы (рисунок 6.7) представлено изображение алгоритма произношения таджикских слов на основе составляющих их морфем.

Здесь предполагается, что произносятся только те слова, для которых морфемный анализ дает положительный результат, несмотря на то что слово делится на морфы.

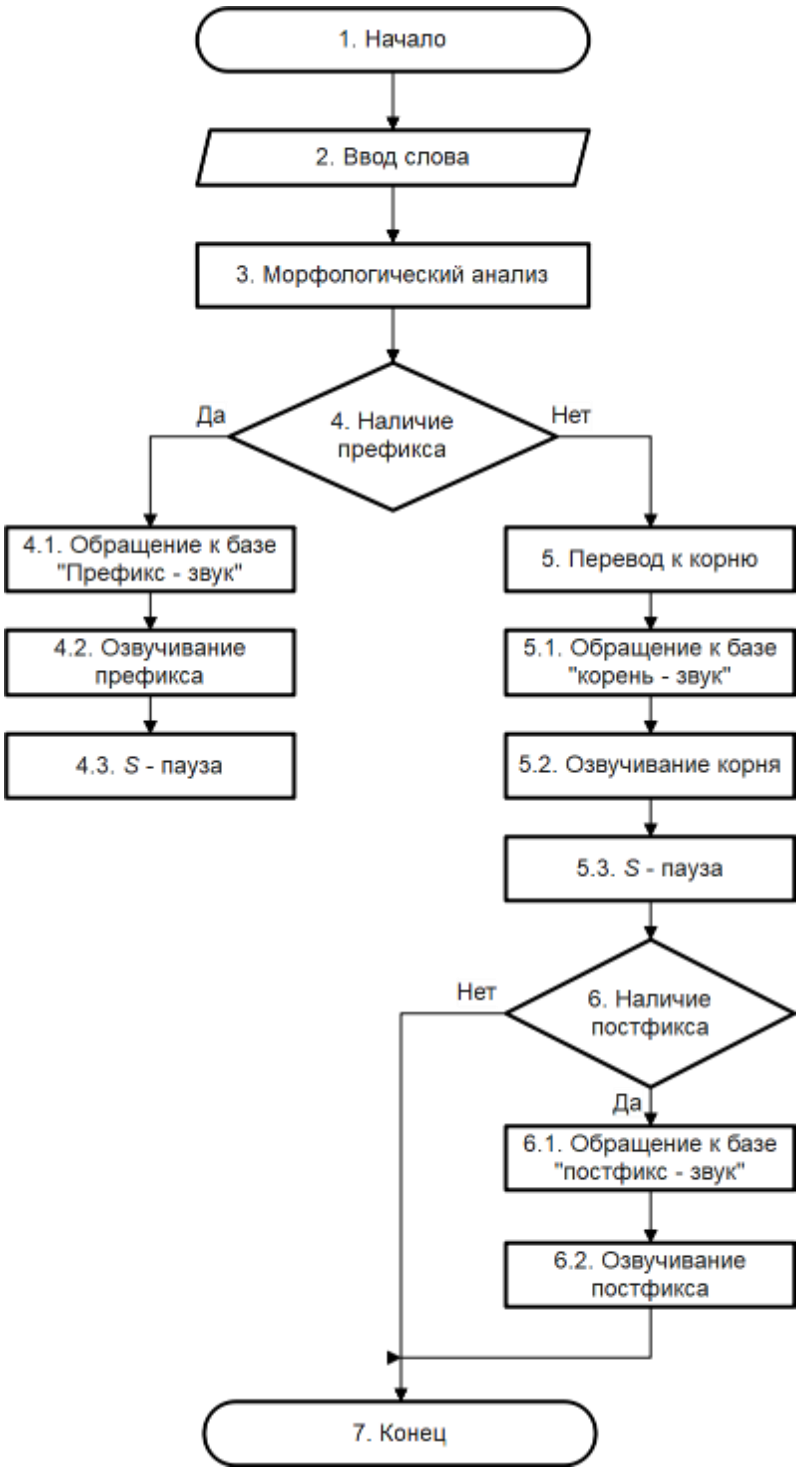

Рисунок 6.7. - Алгоритм анализа морфемного произношения слов

При невозможности анализа морфемного произношения слова оно осуществляется каким-либо другим методом, например, путем конкатенации (сцепления) произношения слогов.

1. Начало – запуск алгоритма.

2. Ввод слова.

3. Слово подвергается морфологическому анализу, в результате которого последовательно представляется слово из 3-х морфем: приставки, ⊕ корня, ⊕ суффикса.

4. Проверка наличия приставки в слове.

Если есть приставка (несмотря на его нееопределенность), то

4.1 обращение к источнику приставки – звуку;

4.2. произношение приставки;

4.3. существование S-интервалов. Переход к пункту 5.

Если проверка в п. 4 показывает, что префикса нет (префикс неопределенный или пустой), то переход к пункту 5.

5. Переход на корень:

5.1. обращение к первоисточнику корня – голосу;

5.2. произношение корня;

5.3. существование S-интервалов.

6. Проверка наличия суффикса в слове.

Если суффикс есть (несмотря на неопределенность суффикса), то

6.1. обращение к источнику суффикса – голосу;

6.2. произнесение суффикса;

6.3. переход к пункту 7.

Если проверка в п. 6 показывает, что префикса нет (префикс – неопределенный), то переход к пункту 7.

7. Конец.

В таджикском языке 66 приставок (простых и сложных) и более 1000 суффиксов. Кроме того, с целью произношения может быть ограничен обширный словарный запас в 50 000 слов. Для того, чтобы реализовать синтез речи, необходимо создать базу данных: префикс – голос, корень – голос и суффикс – голос.

Путем несложных расчетов определено, что для хранения таких источников требуется около 15 Гб постоянной памяти. В случае невозможности анализ слова как последовательность морфов, применяется правило слогового произношения.

О слоговом составе русских слов. В настоящее время таджикский язык содержит большое количество слов, заимствованных из русского языка, преимущественно из существительных. Русские слова встречаются и в других частях речи – глаголах, прилагательных и т.д., но способ их словообразования опирается на правила словообразования таджикского языка. В связи с этим в синтезаторе речи, нацеленном только на произношение таджикских слов, могут возникать ситуации, когда произношение отдельных компонентов русских слов в тексте не осуществляется.

Для выявления его причины в этом разделе предпринята попытка кратко рассмотреть слоговой состав русских слов. С этой целью была взята случайная выборка текстов на русском языке из сети Интернет группы «Известные русские писатели» в объеме почти 100 страниц (108 510 слов). Как и в случае с текстами таджикского языка, сначала кодирование всего выделенного текста производилось с помощью цифр 1 и 0 по отношению к гласным и согласным. Затем обработанный закодированный текст был рассмотрен в материале, представляющем частотность встречаемости русских слов, и описан в виде частей.

Было установлено, что общее количество различных слов, отмеченных в виде элементов, в случайно выбранной выборке составляет ровно 2379. Оказалось, что 50% текста охвачено 26 элементами, которые представлены в таблице 6.13.

Таблица 6.13. - Частота встречаемости русских элементов

№	Состав	Встречаемость, %	№	Состав	Встречаемость, %
1	0	5,52	14	0101010	1,60
2	01	4,85	15	10	1,22
3	1	3,80	16	001010	1,21
4	0101	3,21	17	001001	1,03
5	01010	2,97	18	0101011	0,93
6	001	2,68	19	0010101	0,92
7	010101	2,51	20	0100	0,89
8	01001	2,49	21	01001010	0,86
9	010010	2,26	22	0101001	0,83
10	010	2,11	23	010001	0,80
11	0100101	2,02	24	1001	0,78
12	101	1,93	25	101001	0,69
13	00101	1,63	26	01010101	0,67

Кроме того, 75%, 90% и 95% текста покрыты 103, 323 и 595 частями соответственно. В результате в исследуемой случайной выборке было выявлено 20 слоговых сочетаний (форм), которые представлены в таблице 6.14.

Таблица 6.14. - Слоговые формы русских слов.

№	Слог	Пример	№	Слог	Пример
1	0	в	11	1000	есть
2	1	я	12	00010	стряп
3	01	ты	13	00100	смысл
4	10	он	14	01000	текст
5	010	как	15	001000	власть
6	100	аст	16	000010	взгляд
7	0100	курс	17	000100	вплоть
8	001	кру	18	001000	спасть
9	0010	слад	19	0010000	свойств
10	0001	стро	20	0000100	всплеск

Под номерами 2-7 (в таблице они отмечены серым цветом) показаны слоговые модели таджикского языка.

Таким образом, даже предварительные исследования показывают, что русский язык имеет множество различных слоговых форм (в конечном счете, более 14 форм). Поэтому при создании синтезатора таджикско-русских текстов, основанного на конкатенации слогов, необходимо расширить слоговой источник таджикского языка за счет дополнений, не содержащих русских слогов.

Об алгоритме произношения таджикского текста, содержащего русские слова. В связи с наличием в таджикских текстах большого количества слов, заимствованных из русского языка, возникает важный вопрос, связанный с произношением смешанных текстов. Русские слова, о которых идет речь, в основном существительные. Если, по возможности, охватить наиболее употребительные из них, а затем сделать их слогообразование с последним добавлением таджикского источника *слог-голос*, то мы получим реальную возможность синтеза таджикских текстов с русскими словами.

Алгоритм, который рассмотрен в заголовке этого раздела, по сути ничем не отличается от алгоритмов ударного или безударного произношения таджикских

текстов. Пожалуй, первое, что нужно учитывать, это добавление произносимых русских слогов к таджикскому источнику слог-голос.

§6.4. Система автоматического синтеза речи на таджикском языке

Синтезированная речь в настоящее время создается различными способами, каждый из которых имеет свои достоинства и недостатки. Желаемый синтезатор речи характеризуется двумя основными особенностями – естественностью звука и несложностью получаемой речи. Эти две особенности учитываются при проектировании синтезаторов. Некоторые синтезаторы речи отличаются естественной передачей звука, другие – точностью. В зависимости от поставленных целей для их оформления используются различные методы синтеза речи. Эти методы обычно делят на три группы: *артикуляционный, формантный, гибридный.*

Артикуляционный синтез считается одним из самых сложных методов. Ее представители в Европе и США стремились как можно яснее отразить в числовом формате деятельность гортани человека и процесс звукообразования, чтобы произвести качественную синтетическую речь. До недавнего времени артикуляционный синтез использовался только в научных целях и не привлекал внимания коммерческих организаций. Лишь в последние годы появились некоторые разработаные модели с системой синтезированной речи.

Формантный синтез имитирует человеческую речь с помощью искусственных спектрограмм без использования каких-либо шаблонов. Речевые данные синтезированной речи генерируются аудиомоделью. Такие характеристики, как скорость, громкость и высота тона, со временем накапливаются, создавая форму искусственного речевого сигнала. Большинство систем, основанных на технологии формантного синтеза, создают искусственную речь с «правильным» произношением, поэтому синтезированные речевые данные можно сравнить с естественной человеческой речью.

Системы формантного синтеза имеют некоторые преимущества перед конкатенативными системами, так как в них, во-первых, формантно-синтезированная речь может быть очень четкой благодаря отсутствию характерной для конкатенативных систем звукового шума. В-вторых, формантные синтезаторы обычно представляют собой небольшие программы по сравнению с конкатенативными системами, поэтому они не содержат источника речевых моделей. Их можно использовать в компьютерных системах с минимальной памятью и мощностью процессора.

Наконец, в связи с тем, что формантный синтез предполагает полный контроль над всеми сферами вырабатываемой речевой информации, его объектами могут быть различные рекомендованные просодии (система произношения ударных и безударных слогов, кратких и длинных в речи) или тон произношения не только вопроса и подтверждения, но и вся гамма чувств и эмоций.

Наиболее известные formant-синтезаторы связаны с именем европейского ученого Клатта, определенное представление о первичном formant-синтезаторе можно получить по результатам его исследований.

В основе конкатенативного синтеза или конкатенации лежит заранее записанная связь элементов естественной речи. Такой синтез считается самым простым способом получения ясной и четко звучащей синтезированной речи. В нем одним из важнейших параметров является подбор голосовых отрезков соответствующей длины. Такой выбор делается между короткими и длинными единицами производящими звук. Сравнительно длинные единицы достигают хорошего произношения и высокой степени естественной речи, количество необходимых связей в местах соединения голосовых единиц сокращается. Вместо этого возникает недостаток, заключающийся в неизбежном увеличении с самого начала накопленной памяти компьютера.

Работа с относительно короткими речевыми единицами или фрагментами требует меньше памяти, но процесс их автоматического синтеза становится более трудным.

В современных конкатенативных синтезаторах голосовыми единицами служат фонемы, дифоны, слоги, слова и словосочетания и даже предложения. На первый взгляд кажется, что словам следует отдавать приоритет по сравнению с другими единицами, но в связи с наличием в каждом языке большого и разнообразного количества слов и специальных ниименований, а также из-за разнообразия произношения слов в непрерывной и индивидуальной речи, такой выбор был бы неправильным.

Наиболее распространенными вариантами конкатенативного синтеза являются *параметрический синтез и синтез на основе правил*. Первый вариант относительно гармоничен и гибок с точки зрения измерения силы на основе фонетических единиц: аллофонов, дифонов и слогов. Он расширяет возможности сложных измерений, отвечающих за качество речи содержание форманта, ширина тракта, последовательность тонов, амплитуда сигнала. Это дает возможность подключить звук там, где переход от границы становится незаметным. Преобразование таких параметров как основной поток тона на всей протяженности данных позволяет давать изменения тона и временное описание данных.

Для синтеза используются речевые единицы разной длины: абзацы, предложения, словосочетания, слова, слоги, полуслоги, дифоны. Чем меньше единиц синтеза, тем меньше их количества потребуется для синтеза. Это потребует большего расчета, возникнут проблемы соартикуляции в местах соединения. Преимущество этого метода в том, что он гармоничен, требует мало памяти для защиты первичных материалов и защищает личные характеристики диктора.

Синтез основан на правиле через «неограниченный словарь». Его элементами являются фонемы или слоги, которые соединяются по определенному правилу. Установлено, что для качественного синтеза речи необходимо иметь несколько типов произношения единиц синтеза, например, слогов, что приводит к увеличению словарного запаса первичных единиц независимо от наличия информации о контекстуальная ситуация. Поэтому процесс синтеза приобретает абстрактный характер и превращается из представления о размерах в сложную обработку правил, по которым необходимые измерения рассчитываются на основе

включенного фонетического описания. Эта запись содержит короткое сообщение. Обычно это названия фонетических элементов, например, гласных и согласных со знаками ударения, тональными признаками и временными описаниями. Этот метод позволяет свободно моделировать размеры, хотя само правило моделирования не применяется. Синтезированная речь хуже по качеству, чем естественная, но во всяком случае удовлетворяет текст отчетливостью и четкостью звучания.

Стоит отметить, что среди перечисленных синтезов широкое применение получили *формантный* и *конкатенативный* синтез. Но из них сначала был популярным первый вид. В настоящее время сравнительно популярен второй вид, конкатенативный синтез. По сравнению с ними артикуляциооный синтеза для получения высококачественных изображений кажется более сложным, но вполне вероятно, что в ближайшем будущем он может стать перспективным методом.

Другим распространенным видом синтеза речи является гибридный синтез и синтез на основе Hidden Markov Models (HMM). Гибридный синтез объединил в себе черты формантного и конкатенативного синтеза с целью максимального сокращения голосовых связок в процессе озвучивания речевых элементов.

Концептуальная модель синтеза таджикской речи из текста. В этом разделе работы излагается основная идея синтеза речи из текста, которая, найдя свою реализацию в диссертационной работе, способствует усвоению суть следующих глав. Приведем ряд понятий, которые будут использованы далее.

Текст – это непрерывное предложение, содержащее информацию, организованную по правилам данного языка и знаковой системы. В свою очередь, предложение определяется как совокупность из 7 типов упорядоченных элементов: слово, число, символ, пробел, внутренние знаки препинания (запятая, двоеточие, точка с запятой, тире), внешние знаки препинания (точка, многоточие, вопросительный знак, восклицательный знак) и, наконец, вспомогательные символы конца абзаца (в письменном тексте его нет, но в компьютерном тексте он присутствует).

Понятие, которое мы назвали элементом, должен быть узнаваемым по его обознчаемому значению. Следует отметить, что некоторые элементы могут

отсутствовать в отдельных предложениях, например, цифры, символы, внутренние знаки препинания и т.п., но в случае обязательного присутствия других элементов, например, внешних знаков препинания.

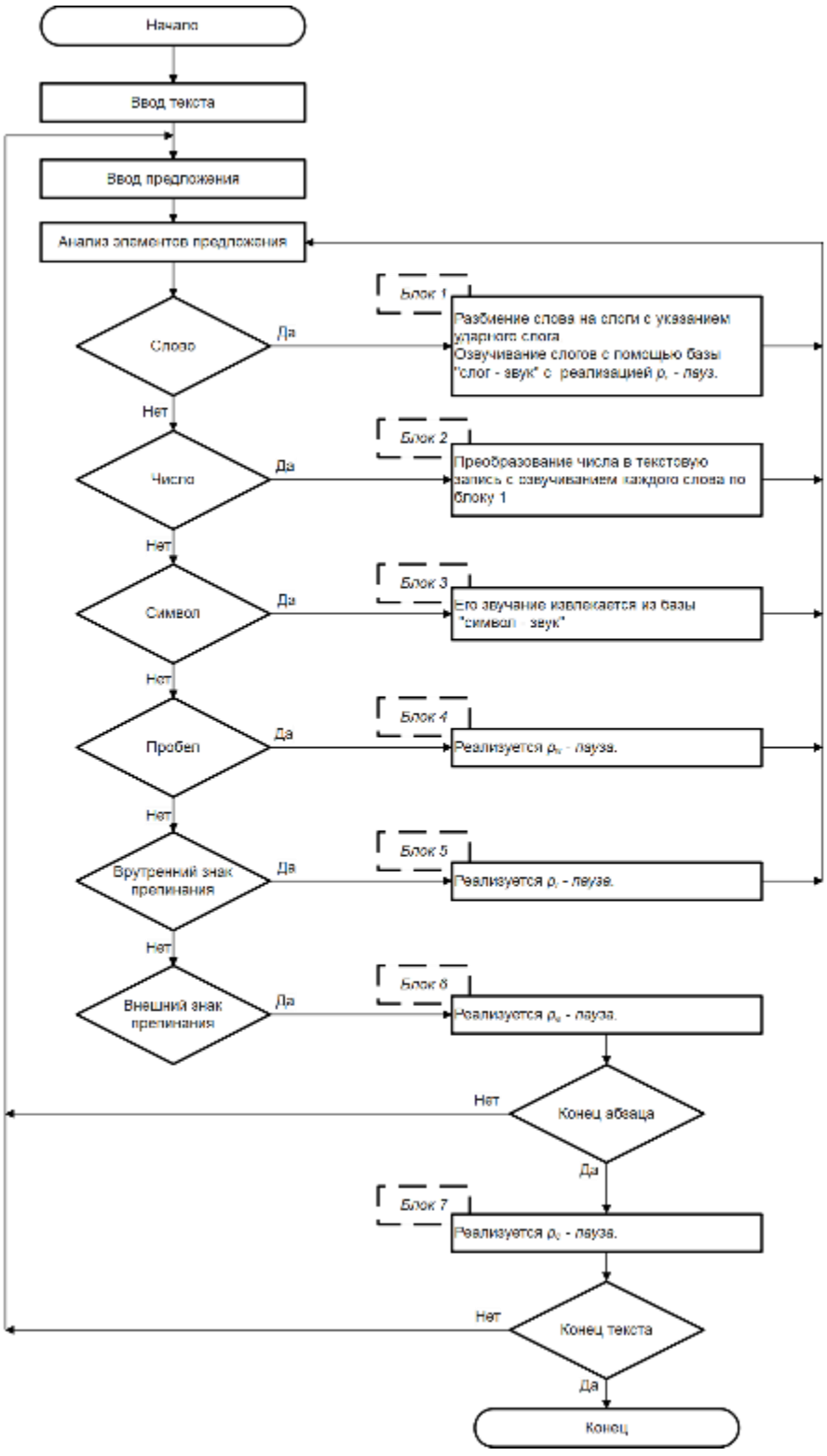

Рисунок 6.8. - Базовая структура синтеза речи из текста

Всего в речи используется пять типов паузы: интервал между слогами при произнесении слова; интервал между словами при чтении предложения соответствует пробелу между словами; интервал, обозначающий внутренние знаки препинания; интервал, обозначающий внешние знаки препинания; интервал, обозначающий конец абзаца.

Теперь мы нашли возможность описать идею синтеза речи из текста в виде принципиальной блочной конструкции.

Синтезатор работает следующим образом. После ввода очередного предложения оно будет проанализировано на основе состава его элементов. Если следующим элементом является слово, то в части 1 оно разбивается на слоги с указанием ударяемого слога, а затем его произношение производится с использованием источника «слог-звук».

Если следующим элементом является число, то в блоке 2 оно преобразуется в текст, а затем осуществляется его произношение через блок.

1. Если следующим элементом является символ, его произношение осуществляется в блоке 3 путем произнесения соответствующего звука из источника «символ – голос».

Если следующим элементом является пробел, внутренний или внешний знак препинания или конец абзаца, то для них из соответствующего блока извлекается соответствующий пробел.

2. Синтезатор речи, изображенный в виде структурной схемы, показывает, что в его основе лежит принцип конкатенации произношения слогов.

Поскольку слог является важнейшей речевой единицей речи, для реализации синтезатора необходимо описать многообразие типов всех слогов согласно естественному языку (задача 1).

Поскольку каждый слог представлен в виде ряда букв и необходим его голосовой отпечаток, необходимо создать источник «слог-голос» (задача 2).

Поскольку синтезатор предполагает произношение цифр и символов, то в первую очередь необходимо преобразовать число в текст (задача 3), затем создать «кодово-голосовой» ресурс (задача 4).

Поскольку в каждом слове есть слог, необходимо разработать автоматическую систему морфемного анализа слова (задание 5).

И, наконец, требуется построить длительность интервалов p_s, p_w, p_i, p_e и p_a таким образом, чтобы получить как можно более четкую естественную речь.

Решение перечисленных задач осуществляется следующими этапами, на которых из текста дается описание примера компьютерного синтезатора таджикской речи. Он организован по правилу конкатенативного синтезатора, а в качестве единицы речи выбирается слог, что в свою очередь указывает на необходимость полного описания многообразия слоговых типов таджикского языка. Решение данной проблемы основано на статистическом исследовании случайной выборки таджикских текстов объемом 3800 страниц, охватывающей 1724472 слова. Кроме того, в нем проведен анализ слогового состава русских слов, поскольку замечено, что в составе таджикского языка имеется большое количество русских слов, а следовательно, и необходимость их произнесения в рамках возник синтез таджикской речи.

Наиболее важные и основные этапы синтеза таджикской речи для любых текстов решаются в следующем порядке:

1. Определение многообразия видов слогов таджикского языка.
2. Разработка алгоритма деления слова на слоги.
3. Определение разновидности типов слогов таджикского языка.
4. Выявление проблем в распознавании слогов.
5. Разработка алгоритма морфемного анализа таджикских слов.
6. Разработка слого-звукового источника на основе таджикских слогов.
7. Подготовка слого-звукового источника для синтеза речи.
8. Разработка алгоритма произношения слова.
9. Разработка алгоритма произнесения цифр и символов.
10. Разработка алгоритма произношения безударного текста.
11. Разработка алгоритма произношения ударного текста.
12. Разработка алгоритма морфемного произношения слов.

13. Разработка алгоритма произношения текста с русскими словами.

14. Разработка программного комплекса обработка синтеза речи.

15. Подготовка программы и технических средств произношения.

16. Корректировка качества произношения слогов и слов путем проведения компьютерных тестов и полной оценки количества слогов для формирования искусственной речи.

17. Проведение компьютерных тестов с произношением морфем и таджикского текста, содержащего русские слова.

Комплексный структурный программированный план. Комплексный структурный программированный план изображен на рисунке 6.9.

Рисунок 6.9. - Структурный план Tajik Text-to-Speech

Блок 1. Подсистема «*Пользовательский интерфейс*» состоит из двух компонентов – «Ввод текста» и «Произнесенная речь», которая имеет одностороннюю связь, однако пользователь имеет возможность вводить текстовые данные и в результате получать речевую версию введенного текста.

Для получения результата блок 1 связан с блоком 2 по двум направлениям – предоставление данных для лингвистического анализа и получение результатов произношения. Блок 1 также работает совместно с блоком 3 непосредственно для использования необходимой информации о настройках программы (выбор женского или мужского голоса, высоты и скорости произношения).

Блок 2. Аналитическая подсистема состоит из двух частей - «Лингвистический анализ» и «Модуль произношения». Первая часть состоит из подмодулей «Проверка текста», «Кодирование текста» и «Деление слов на слоги». Подмодуль «Проверка текста» используется для проверки входных данных, содержащих текстовые элементы, такие как слова, целые числа, символы и знаки препинания. Она проверяет текстовые элементы, формирует целые числа и символы в текстовом формате, а затем передает их на кодирование.

В процессе кодирования создается одноименный подмодуль, где каждое слово W текста формируется из упорядоченного набора $W_{0,1}^{*}$ нулей и единиц напоминаем, что гласные буквы обозначаются цифрой 1, а согласные – цифрой 0, однако представлены все слова с составом их слогов. Закодированный текст передается в подмодуль «Распределение слов по слогам», а слова, разделенные на слоги, подвергаются лингвистическому анализу и передаются в «Модуль произношения».

В указанном модуле формирование речевой информации осуществляется с использованием базы слог-звук информационной подсистемы, ударных слогов, межсложных и межсловных интервалов, а также интервалов, выражающих знаки препинания, например запятые и точки. Модуль произношения является завершающим этапом подсистемы анализа, а аудиотип текстовых данных передается в пользовательский интерфейс.

Блок 3 «Информационная подсистема» включает источники данных «Упорядочение системы» и «слог-звук». Прежде всего он используется для защиты временных данных системы, а также для сохранения слого-звукового источника и статистических данных о звуковых файлах из 3259 слогов таджикского языка. Для работы с этим источником данных используется модуль предоставления доступа, проверки и выбора необходимых данных.

На основе полученных результатов и результатов тестов реализован первый опыт компьютеризированного произношения таджикского текста. Несмотря на то, что предлагаемый таджикский программный комплекс Tajik Text-to-Speech получил положительную оценку по двум характеристикам – естественности произношения и четкости произносимой речи, сами авторы считают его лишь образцом компьютерного синтезатора таджикской речи из текста, которая нуждается в серьезном исследовании в будущем.

§6.5. Проблемы распознавания речи на таджикском языке

Автоматическое распознавание речи – одна из важнейших задач информационных технологий. Проблема распознавания речи возникла с момента зарождения информатики и изучалась вместе с проблемой автоматического перевода с одного языка на другой. Результаты полученных к настоящему времени проведенных исследований недостаточны для применения в виде компьютерной программы с возможностями распознавания речи на естественных языках.

В общем, существует две основные задачи перед распознаванием устной речи. Во-первых, достижение полной точности при наличии ограниченного набора команд хотя бы для одного голоса рассказчика. Во-вторых, самостоятельное распознавание непрерывной связной речи приемлемого качества независимо от говорящего. Несмотря на полувековые научные и практические разработки, обе эти проблемы до сих пор не решены.

Ключевой особенностью речи является то, что он различается по многим параметрам: длительности, скорости, высоте голоса, характеристикам, вызванным

большими изменениями речевого аппарата человека, разными эмоциональными состояниями рассказчика, значительными различиями в голосах разных людей. Два изображения одинаковой длительности речи не совпадают даже у одного и того же человека, записанные в разное время.

Необходимо искать такие измерения звука речи, которые, с одной стороны, могли бы полностью описать его т.е. отличать один речевой голос от другого голоса, с другой стороны, уравнивать упомянутые выше варианты речи. Затем эти измерения следует сравнить с образцами, и это должно быть не простое сравнение ради согласия, а поиск наиболее точного совпадения. Это приводит к необходимости поиска необходимой формы расстояния в пространстве найденных измерений.

Таким образом, процедура распознавания устной речи должна быть основана на использовании подходящей системы измерений, признаков и реализована с помощью разумных алгоритмов.

Сравнительный анализ системы распознавания речи. В настоящее время системы распознавания устной речи существуют, большинство из которых активно совершенствуются и распространяются с открытым исходным кодом. Использование автоматической системы распознавания устной речи необходимо для быстрого и удобного ввода данных. Например, звуки речи используются от приложений на различных смартфонах до систем управления зданиями. Однако, к сожалению, не все алгоритмы и системы автоматического распознавания речи соответствуют ожидаемым от них результатам, и зачастую распознать их невозможно. К счастью, есть такие системы, как Yandex SpeechKit, Alexa, Siri и Google Assistant, которые умеют точно распознавать речь. Проблема таких систем – поддержка других языковых моделей, и эти действующие на данный момент системы не могут распознавать таджикскую речь.

Проблема распознавания звуков таджикской речи заключается в полном отсутствии речевых данных. Причина последнего в том, что количество потенциальных пользователей таджикского языка невелико, большая часть из-за отсутствия таджикского аналога используют русский язык. Развитие таджикского

языка привело к широкому использованию его во всех областях жизни. Поэтому соединить компьютерные достижения с нуждами народа является актуальной задачей сегодняшнего дня.

Одной из важных задач в решении компьютерной лингвистики принадлежит вопросу разработки системы распознавания речи. Поэтому потребность в использовании системы распознавания таджикской речи возрастает с каждым годом, но аналога до сих пор нет. Для решения этой проблемы была поставлена задача создания системы распознавания таджикской речи.

В современном мире в настоящее время существует множество систем распознавания английского, немецкого, русского, а также систем для других популярных языков.

Для решения этой проблемы можно использовать уже готовую систему с открытым исходным кодом. И чтобы решить, какую систему выбрать в качестве ядра, было решено проанализировать популярные системы автоматического распознавания речи, такие как CMU Sphinx и Mozilla deepSpeech. Эти системы были выбраны исходя из частоты упоминаний в современных научных журналах, а также мнений и отзывов разработчиков в области машинного обучения и нейронных сетей. CMU Sphinx и Mozilla deepSpeech сравнивались по точности, скорости распознавания речи и простоте использования.

Для определения точности распознавания речи используется следующая формула:

$$accuracy = \frac{Sk}{S}, \tag{6.3}$$

где,

accuracy – точность распознавания звукового файла;

Sk – количество правильных распознаваний слов;

S – общее количество речевых слов при распознавании.

Для определения точного количества правильного распознавания речи воспользуемся следующей формулой:

$$Sk = S - Se - Sm \quad (6.4)$$

где,

Se – количество неправильных распознаваний речевых слов;

Sm – количество пропущенных слов при распознавании.

Для определения точности распознавания текста используется система стандартных показателей скорости распознавания слов Word Recognition Rate (WRR) ва Word Error Rate (WER), которые имеют следующую формулу:

$$WER = \frac{S+I+D}{T} \text{ (6.5) и } WRR = 1\text{-}WER \quad (6.5)$$

где,

S – количество операций замены слов;

I - количество операций ввода слова;

D - количество исключения словообразовательных операций из распознанного оборота для получения исходной фразы;

T – количество слов в исходном обороте и рассчитывается в процентах.

Для расчета скорости используется критерий Real Time Factor («Расчет в реальном времени»), который является показателем для расчета эффективности алгоритмов, используемых в системах реального времени, или Speed Factor (SF), который решается в рамках теории алгоритмов.

$$SF = \frac{Tr}{T} \quad (6.6)$$

где,

Tr - время распознавания звука;

T – его продолжительность, измеряемая в долях реального времени.

После проведения эксперимента и анализа двух систем был получен следующий результат. В таблице 6.15 в первом столбце для автоматического распознавания представлен текст в голосовой форме на английском языке. Во втором и третьем столбцах показаны результаты распознавания голоса с помощью систем Mozilla DeepSpeech и CMU Sphinx.

Таблица 6.15. - Пример теста на распознавание устной речи

Оригигнальный звучащий текст	**Распознавание голоса с помощью Mozilla DeepSpeech**	**Распознавание голоса с помощью CMU Sphinx**
Food People, as well animals, can't live without food. It is known than a man can't live without food more than a couple of month. Food gives us energy and strength. So it's very important to eat good healthy food. Nowadays a lot of people don't have time for cooking at home. They eat fast food like hamburgers, hot dogs, snacks and pizza is the best case. A lot of schoolchildren eat chips and all kids of snacks. I think it's bad for their health. I try to eat homemade food.	food people as well as animals cantle{can't} live without for{food} it is known that a man can't live without food more than a couple of months who{food} give{gives} us energy and strength so it's very important to [eat] good health{healthy} it for{food} nowadays a lot of people don't have time for cooking at home they it {eats} fast full{food} like hamburgers had dogs{hotdogs} snacks and beating { pizza is } the best case a lot of schoolchildren it chips and all kinds of legs{snaks}. i think it's bad for their help{helth} i tried {try} to homemade food	food pupil{pupil}, as well animals, can{can't} live without food it is known than a man can't live without foot {food} more than a couple of month who {food} gives [us] energy and strength so [it's] very important to [eat] good healthy it for {food} nowadays a lose {lot} of people don't have time for looking{cooking} as {at} home. they eat fast food like hamburgers, hat{hot} dog{dogs} snacks end{and} pizza is the best case a lot [of] schoolchildren eat chips and all kids of legs {snacks} i think is {it's} bad fore{for} their help {health} i try to eat homemade foot{food}

Для сравнения систем распознавания устной речи необходимо минимум 200 часов данных, и поскольку такие ресурсы в настоящее время недоступны, было решено использовать из готового обучающего комплекса. Подобные ресурсы можно найти повсюду в интернете, но доступны только английская или немецкая версии.

В качестве обучения была использована модель Common Voice, созданная 20000 участниками-добровольцами. Они записали около 400 000 сообщений общей продолжительностью около 500 часов.

В результате была сформирована большая база голосовых выражений с различными моделями речевого поведения. Речевая модель была протестирована также такими системами, как LibriSpeech, Google Speech и Apple Dictation.

Сравнительный результат между системами распознавания голоса представлен в таблице 6.16.

Таблица 6.16. - Результаты сравнения точности и скорости

Система	Точные	WER, %	WRR, %	SF
Mozilla DeepSpeech	83,2	16,78/18,9	83,2/81,1	0,6/1
CMU Sphinx	79,8	21,4/22,7	78,6/77,3	0,5/1

После проведенного сравнения стало ясно, что с помощью этих систем можно распознавать слова таджикской речи.

Кроме того, эти системы используют следующие этапы обработки для распознавания речи:

1. Определение акустических характеристик по речевому сигналу.
2. Акустическое моделирование.
3. Лингвистическое моделирование.
4. Расшифровка.
5. Реализация алгоритмов синтеза речи.
6. Проверка соответствия алгоритмов синтеза и распознавания.

В таблице 6.17 приведено сравнение основных возможностей систем автоматического распознавания голоса.

Таблица 6.17. - Сравнение возможностей систем распознавания голоса

Система / возможность	**Mozilla DeepSpeech**	**CMU Sphinx**
Идентификация акустических знаков	MFCC	MFCC, PLP
Акустическое моделирование	HMM,GMM, SGMM, DNN	HMM
Языковое моделирование	FST, N-gramm	N-gramm, FST
Алгоритм распознавания речи	Алгоритм Витерби, Алгоритм двуполосного прямого возврата	Алгоритм Витерби, алгоритм bushderby
Реализация	C/Python (модульное строение)	C/Java (модульное строение)

Проанализировав приведенные выше результаты, было установлено, что Mozilla DeepSpeech не уступает CMU Sphinx по точности, но CMU Sphinx быстрее Mozilla DeepSpeech по скорости. Также было установлено, что обе системы являются модульными, и обе системы могут быть использованы для распознавания слов таджикской речи при создании образца таджикской речи.

Далее необходимо составить речевую модель таджикского языка для анализа и работоспособности этих систем и разработки системы распознавания речи. Также в будущем будет продолжена деятельность в области создания и анализа речевых моделей на разных языках для получения наилучшего результата.

Алгоритм обработки динамического измерения времени. Главная особенность звука речи состоит в том, что он варьирует по многим параметрам: длительности, скорости, высоте, искажениям, вызванным большим изменением речевого аппарата человека, разными эмоциональными состояниями рассказчика, большой разницей голосов разных людей. Два временных представления одного и того же фрагмента речи даже одного и того же человека, записанные в разное время, не совпадают.

Необходимо искать такие измерения звука речи, которые, с одной стороны, могли бы полностью описать его т.е. отличать один речевой голос от другого голоса, с другой стороны, уравнивать упомянутые выше варианты речи. Затем эти измерения следует сравнить с образцами, и это должно быть не простое сравнение для достижения согласия, а поиск наиболее точного совпадения. Это приводит к необходимости поиска необходимой формы расстояния в пространстве найденных измерений.

Таким образом, процедура распознавания устной речи должна быть основана на использовании подходящей системы критериев (признаков) и реализована с помощью продуманных алгоритмов.

Особенностью упомянутого ниже подхода к динамическому обмену измерения времени распознаваемого звука является рассмотрение доминирования речевого сигнала во временном представлении, а не в частотном представлении.

Для реализации распознавания звукового потока следует использовать общедоступные алгоритмы, основанные на обмене показателями времени. Аудиопотоки преобразуются в цифровые сигналы путем преобразования звуковых волн в наборы чисел, разделенных на временные шкалы.

Динамическое преобразование временной шкалы – это алгоритм, основанный на способе расчета пропорционального пути деформации между двумя наборами чисел двух звуковых потоков. Результат работы алгоритма дает значения изменения пути и расстояния между двумя представленными наборами чисел. Чем меньше путь изменения между двумя представленными потоками, тем больше вероятность того, что два звуковых потока идентичны.

Если одно и то же слово произносится двумя разными говорящими, то образуются два разных звуковых потока с разными временными индексами. Например, слово «Хусрав» можно произносить как «Хусрав» или «Хисрав». Алгоритм динамического преобразования временной шкалы решает основную проблему распознавания голоса, правильно сопоставляя слова и затем определяя наименьшее расстояние между двумя произнесенными словами.

Чтобы решить задачу распознавания речи, необходимо скорректировать разницу во времени путем измерения расстояния. Сначала проводится моделирование методом незначительного изменения временной шкалы звукового потока до тех пор, пока не станет возможным распознавание. Метод обмена временными графиками обеспечивает эффективное решение проблемы координации времени. В контексте такого использования голосового текста необходимо использовать его реалистично, чтобы определить совместимость с представленным аудиопотоком. Для обработки звуковых данных всегда используются часовые пояса, что определяет изменение звуковых показателей двух последовательностей звуковых данных.

Основная задача алгоритма обмена графика времени – уравнять две векторные последовательности путем многократного поворота оси времени, пока не будет найдено пропорциональное совпадение между этими двумя последовательностями. Алгоритм действует как линейное отражение для

сопоставления двух звуковых файлов. Например, рассмотрим две числовые последовательности, основанные на двух звуковых файлах:

$$x = [x1\ x2\ \dots\ xn] \text{ и } y = [y1\ y2\ \dots\ yn] \tag{6.7}$$

Сопоставление двух последовательностей осуществляется по сторонам двумерной матрицы: первой по строкам, второй по столбцам. Начальная точка совпадения задается в левой нижней части матрицы. Каждому элементу матрицы присвоена мера расстояния, сравнивающая соответствующие элементы в строках и столбцах. Значения расстояний между двумя точками рассчитываются с использованием евклидова расстояния.

$$\text{Dist}(x,y)=|x-y|=[(x\_1-y\_1\ )\ ^2+(x\_2-y\_2\ )\ ^2+\cdots+(x\_n-y\_n\ )\ ^2\ ]\ ^(1/2) \tag{6.8}$$

Затем, сокращая общее расстояние между двумя последовательностями, получаем пропорциональное согласование двух предложенных звуковых потоков. В результате поиска и прохождения всех возможных маршрутов в матрице вычисляется общее расстояние, т. е. суммарное расстояние двух последовательностей.

Значение наименьшего расстояния можно получить, разделив общее количество расстояний между отдельными элементами при передаче матрицы на общее количество весовых операций. Важно отметить, что для длинной последовательности общее количество переходов от элементов матрицы возвращает большое значение. В связи с этим требуется пропорциональность значений, которая определяется функцией D(i,j).

$$D(i,j) = |t(i) + r(j)| + min\begin{cases} D(i+1,j) \\ D(i+1),j+1 \\ D(i,j+1) \end{cases} \tag{6.9}$$

Рассмотрим функцию, определяющую расстояние динамического обмена размерностью времени между элементами t(i;m) и r(j;n), переходящим от позиций (i,j) к (m,n), которое удовлетворяет начальному условию:

$$D(m, n) = | t(m) - r(n)| \quad (6.10)$$

Алгоритмы распознавания слогов таджикской речи в пространстве колебания и времени. Особенностью нижеприведенного подхода к распознаванию слоговой речи является доминирующий учет звука речи в его временном, а не частотном отношении.

Термин «слоговое распознавание» означает, что в качестве распознаваемых единиц берутся не предложения, слова или морфемы, а слоги – то есть звуки речи или элементы фонетической структуры распознаваемого языка (в нашем случае – таджикского языка).

В целях изучения закономерностей таджикского языка, связанных с понятием слога, дополнительно введем понятие слоговой структуры слова. Для этого гласные в словах следует заменить на цифру 1, а согласные на цифру 0. Например, слово «хуршед» – «010010», «ватан» – «01010».

Указанный слоговый строй таджикских слов делится на слоги согласно делению на слоги тех таджикских слов, которые входят в ту или иную определенную структуру. В результате было найдено всего 9 различных структур слогов – 1, 10, 01, 010, 100, 0100 и 001, 0010, 00100.

Различные типы слогов таджикского языка были получены на основе составленных и интегрированных алгоритмов компьютерных программ. В большом объеме случайно выбранного текста было обнаружено 3259 различных слогов, которые при их произношении составляли «слого-звуковую» основу.

На основе полученных данных: слоговой структуры слов, структуры слогов и, наконец, разнообразия слогов таджикского языка необходимо составить ряд алгоритмов применения слогового членения речи.

Сначала представляется речевой сигнал в колебании и времени в пространстве. Поскольку слоги идентифицированы как узнаваемые единицы речи, мы рассматриваем признаки, которые становятся основой для определения классов фонем с достаточной точностью и скоростью.

Внешний вид рисунка этого графика функции позволяет сделать ряд предположений о произносимых звуках. Как видно из рисунка 6.10, области таблицы для звуков, произносимых без участия голоса (закрытые слоги), существенно отличаются от областей со звонкими звуками (открытые слоги).

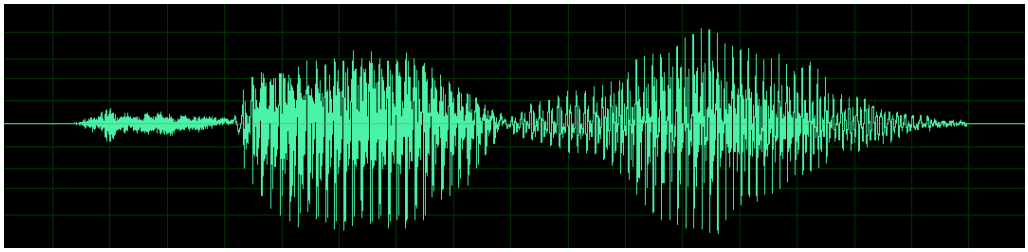

Рисунок 6.10. - Оцифрованное звучание слова «ватан»

Предложенная функция ведет себя по-разному в регионах с разными слогами. Мы можем попытаться найти отличительные черты ее «поведения», которые должны быть измеримы и при измерении это позволит отличить один класс слогов от другого, используя разрешенные значения.

Примером такого символа является величина V с числовым эквивалентом полного изменения функции для дискретного состояния. Здесь n – количество подсчетов в области сигнала, xi – значение i-го отсчета:

$$V = \sum_{k=0}^{n} | x_{k+1} - x_k | \qquad (6.11)$$

Для разделения речевого потока необходимо разработать соответствующие алгоритмы, основанные на структуре слогов, определить процесс образования пауз между слогами.

Алгоритмы определения структуры слогов. На основе упомянутого выше «слог-звук» приводятся слоги структуры «1», состоящие только из одной гласной:

«ё», «у», «е», «ӯ», «а», «о» , «е», «ман», «ва», «ю». Средняя длина слогов в звучащей форме составляет 298 миллисекунд. Определение структуры слогов – «1», «01», «10», «010», «100», «0100». В ходе исследования остальной структуры слогов была получена необходимая информация для разработки дальнейших алгоритмов. Результаты исследования представлены в таблице 6.18 начиная со 2-го числа слоговой структуры.

Таблица 6.18. - Размеры слоговых структур таджикского языка

№	**Структура**	**Среднее время, микросекунда**	**Образцы слогов**
1	1	285	“ӯ”, “ё”, “и”
2	01	330	“ба”, “ро”, “фи”
3	10	315	“ил”, “ор”, “эй”
4	100	455	“аср”, “орд”, “умр”
5	010	375	“дур”, “кор”, “шир”
6	0100	540	“сард”, “бист”, “данд”

Для остановок и процессов им подобным характерно большое число постоянных точек. Обрабатываем звук цифровым проходным фильтром с допустимой полосой 100-200 Гц. В этом случае процесс межсложных речевых сигналов преобразуются в паузы.

Описанные алгоритмы могут отражать определенную точность и стабильность распределения слогов. В результате такой подход к изучению речевого сигнала в его временном выражении позволяет разработать комплексы программ распознавания речи на таджикском языке.

Выводы по шестой главе

По статистической закономерности текстовых данных в таджикском языке при обработке текстовых данных выявлено всего 274 различных структуры слов (элемент, гласная - 1, согласная - 0) в объеме 1724472 слов. Установлено, что 8 единиц («01», «010», «01010», «01001», «10», «0101», «010101», «100101») покрывают 50%, а 23 элементов покрывают 75% таджикских текстов. 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».

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, состоящей из 1724472 слов, 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.

Для обеспечения возможности автоматического синтеза речи были проанализированы правила фонетики таджикского языка, то есть фонемы гласных и согласных. На этапе обработки определялась цифровая структура каждой фонемы. В целях предоставления информации для системы автоматического синтеза речи был создан слогово-голосовой источник данных, состоящий из 2 × 3259 таджикских слогов двумя профессиональными дикторами (мужской и женский голос).

Для программного обеспечения синтеза речи на основе математических моделей и специальных методов были разработаны следующие алгоритмы:

1. Алгоритм озвучивания слов.

2. Алгоритм озвучивания цифр и символов.

3. Алгоритмы безударного и ударного озвучивания текста.

4. Алгоритм озвучивания морфемы слова.

5. Алгоритм озвучивания таджикского текста, содержащего русские слова.

На основе полученных результатов была разработана автоматическая система с возможностью синтеза речи на таджикском языке Tajik text-to-speech.

Концептуальная модель синтеза таджикской речи выражает основную идею синтеза речи из текста. С учетом этого возможность синтезировать речь в SO Windows в качестве диктора доступна в модуле «Tajik Text Narrator». Кроме того,

для пользователей интернета на сайте www.tajlingvo.tj\talaffuz доступна возможность синтеза таджикской речи в формате онлайн.

Научные результаты, полученные в рамках разработки системы автоматического синтеза речи на таджикском языке, в будущем могут быть использованы как основа для решения проблемы распознавания речи на таджикском языке. По этой причине были проанализированы основные проблемы решения задачи автоматического распознавания речи таджикского языка.

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. относительно эффективен. На основе этих показателей предложен алгоритм распознавания речи на таджикском языке на основе анализа слоговой структуры слов, который будет изпользоваться в дальнейших исследованиях.

ЗАКЛЮЧЕНИЕ

В настоящее время информационно-коммуникационные технологии широко используется в различных сферах жизни общества в Республике Таджикистан. Вместе с тем компьютерные программы и информационные системы, созданные на основе современного литературного таджикского языка, встречаются крайне редко. Это касается таких направлений, как обработка электронных словарей, компьютерный тезаурус, автоматической системы проверки орфографии, компьютерный синтез и распознавания речи, автоматической транслитерации, автоматического перевода текста. Результаты исследования, включают решение вышеперечисленных вопросов с использованием 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 [А-4], [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 ЛИТЕРАТУРЫ

1. Барномаи давлатии амалӣ намудани технологияҳои иттилоотӣ-коммуникатсионӣ дар муассисаҳои таҳсилоти умумии Ҷумҳурии Тоҷикистон барои солҳои 2018-2022. Қарори Ҳукумати Ҷумҳурии Тоҷикистон аз 29 сентябри соли 2017 № 443
2. Консепсияи ташаккули Ҳукумати электронӣ дар Ҷумҳурии Тоҷикистон. Қарори Ҳукумати Ҷумҳурии Тоҷикистон аз 30 декабри соли 2011, № 643
3. Конститутсияи Ҷумҳурии Тоҷикистон аз 6 ноябри соли 1994 бо тағйиру иловаҳо аз 26 сентябри соли 1999, 22 июни соли 2003 ва 22 майи соли 2016 (бо забонҳои тоҷикӣ ва русӣ). – Душанбе: Ганҷ, 2016. – 136 с.
4. Қарори Ҳукумати Ҷумҳурии Тоҷикистон дар бораи "Стандарти ҷобаҷогузории давлатии алифбои тоҷикӣ дар клавиатураи компютерӣ" аз 2 августи соли 2004 таҳти № 330
5. Қонуни Ҷумҳурии Тоҷикистон "Дар бораи забони Давлатии Ҷумҳурии Тоҷикистон". Ахбори Маҷлиси Олии Ҷумҳурии Тоҷикистон, соли 2009, №9-10, мод.546
6. Қонуни Ҷумҳурии Тоҷикистон "Дар бораи иттилоот" аз 03.07.2012 №848
7. Қонуни Ҷумҳурии Тоҷикистон дар бораи иттилоотонӣ. http://ncz.tj/system/files/Legislation/40_ru.pdf
8. Стратегияи давлатии "Технологияҳои иттилоотӣ-коммуникатсионӣ барои рушди Ҷумҳурии Тоҷикистон". Қарори Президенти Ҷумҳурии Тоҷикистон аз 5 ноябри соли 2003, № 1174
9. Стратегияи миллии рушди Ҷумҳурии Тоҷикистон барои давраи то соли 2030. Қарори Маҷлиси намояндагони Маҷлиси Олии Ҷумҳурии Тоҷикистон аз 1 декабри соли 2016, № 636
10. Стратегияи рушди инноватсионии Ҷумҳурии Тоҷикистон барои давраи то соли 2020. Қарори Ҳукумати Ҷумҳурии Тоҷикистон аз 30 майи соли 2015, № 354
11. Назарзода С. Имло ва забони адабӣ. – Душанбе, «Андалеб», 2015. – 312 с.

12. Нуров П.Г. Масъалаҳои омӯзиш, таҳқиқ ва рушди забони илмии тоҷикӣ. – Душанбе: Дониш, 2008. - 165 с.
13. Рустамов Ш.Р. Калимасозии исм дар забони адабии тоҷик. - Душанбе, 1972. - 76 с.
14. Фарҳанги имлои забони тоҷикӣ. Муҳаррири масъул Саймиддинов Д. – Душанбе: Шарқи озод, 2013. – 320 с.
15. Фарҳанги тафсирии забони тоҷикӣ. Ҷилди 1. Зери таҳрири С.Назарзода, А.Сангинов, С.Каримов, М.Ҳ.Султон. – Душанбе: Академияи илмҳои Ҷумҳурии Тоҷикистон пажӯҳишгоҳи забон ва адабиёти ба номи Рӯдакӣ. – 2008. 1091с.
16. Фарҳанги тафсирии забони тоҷикӣ. Ҷилди 2. Зери таҳрири С.Назарзода, А.Сангинов, С.Каримов, М.Ҳ.Султон.– Душанбе: Академияи илмҳои Ҷумҳурии Тоҷикистон пажӯҳишгоҳи забон ва адабиёти ба номи Рӯдакӣ. – 2008. 950 с.
17. Анисимов А. В. Компьютерная лингвистика для всех: Мифы. Алгоритмы. Язык. – Наук. думка, 1991. – 325 с.
18. Батура Т.В., Мурзин Ф.А. Машинно-ориентированные логические методы отображения семантики текста на естественном языке: моногр. Институт систем информатики им. А.П. Ершова СО РАН. Новосибирск: Изд. НГТУ, 2008. 248 с.
19. Буч, Г. UML. Руководство пользователя / Г. Буч, Д. Рамбо, А. Джекобсон. - М.: ДМК Пресс; Издание 2-е, стер., 2014. - 432 с.
20. Расторгуева В.С. Краткий очерк грамматики таджикского языка, - с. 529 - 570. В книге «Таджикско-русский словарь» под редакции М.В. Рахими и Л.В. Успенской, Госиздат иностранных и национальных словарей, - М., 1954. - 789 с.
21. Искандарова Д.М. Касымова М.Н. Хрестоматия по теоретической и прикладной лингвистике. - Душнбе: РТСУ, 2005
22. Карчевская М. П., Рамбургер О. Л., Ковтуненко А. С. Обработка данных на VISUAL C#. NET. – 2018.

23. Кривнова О. Ф. Ритмизация и интонационное членение текста в «процессе речи-мысли» (опыт теоретико-экспериментального исследования) //-М.: МГУ. – 2007.
24. Ларман, К. Применение UML и шаблонов проектирования / К.Ларман. - М.: Вильямс, 2015. - 624 с.
25. Ларман, К. Применение UML 2.0 и шаблонов проектирования. Введение в объектно-ориентированный анализ, проектирование и итеративную разработку / К.Ларман. - М.: Вильямс, 2013. - 736 с.
26. Лесников С. В. Направления и разделы лингвистики в систематическом указателе гипертекстового информационно-поискового тезауруса метаязыка лингвистики //Человек в информационном пространстве: межвузовский сборник научных трудов. – 2011. – Т. 2. – №. 10. – С. 214.
27. Марченко А. А. Метод автоматического построения онтологических баз знаний. I. Разработка семантико-синтаксической модели естественного языка //Кибернетика и системный анализ. – 2016.
28. Мюллер, Р.Дж. Базы данных и UML. Проектирование / Р.Дж. Мюллер. - М.: ЛОРИ, 2017. - 420 с.
29. Потапова Р. К. Новые информационные технологии и лингвистика. – URSS, 2005.
30. Потемкин С. Б., Кедрова Г. Е. Семантическое расстояние между предложениями на основе модифицированного расстояния Левенштейна //ББК 88.3 К57. – 2018. – С. 246.
31. Приемы объектно-ориентированного проектирования. Паттерны проектирования / Э. Гамма и др. – М.: СИНТЕГ, 2016. - 366 с.
32. Робин Н. Создаем динамические веб-сайты с помощью PHP, MySQL, JavaScript, CSS и HTML5. 4-е изд. –СПб.: – Издательский дом "Питер", 2016.
33. Сажок Н. Н. Кластеризация слов при построении лингвистической модели для автоматического распознавания речевого сигнала //Кибернетика и вычислительная техника. – СПб.: БХВ-Петебург, – 2012.

34. Солонина А.И. Основы цифровой обработки сигналов. Курс лекций: Учебное пособие, 2-е изд. // А.И. Солонина, Д. Улахович, С. Арбузов, Е. Соловьева. – СПб.: БХВ-Петебург, –2012.
35. Сорокин В.Н. Синтез речи. - М.: Наука, 1992. - 392 с
36. Сулейманов Д.Ш., Гатиатуллин А.Р. Структурно-функциональная компьютерная модель татарских морфем. – 2003.
37. Фаулер, М. UML. Основы. Краткое руководство по стандартному языку объектного моделирования / М. Фаулер. - М.: Символ-плюс, 2016. - 192 с.
38. Фомичев В. А. Формализация проектирования лингвистических процессоров. – МАКС Пресс, 2005.
39. Чистович Л. А., Физиология речи. Восприятие речи человеком / Л. А. Чистович, А. В. Венцов, М. П. Грамстрем и др. // -М.: Наука, –1976. –С 388.
40. Шумаков П. В. ADO. NET и создание приложений баз данных в среде Microsoft Visual Studio. NET. – 2003.
41. Black A. W., Taylor P. A. Automatically clustering similar units for unit selection in speech synthesis. – 1997.
42. Cohen M., Massaro D. Modelling Coarticulation in Synthetic Visual Speech. Proceedings of Computer Animation 93, Suisse. 1993.
43. Grishman, R. Computational Linguistics: An Introduction / R. Grishman. Cambridge etc.: Cambridge University Press, 1986. 193 p.
44. Hausser R. R. A computational model of natural language communication: Interpretation, inference, and production in database semantics. – Springer Science & Business Media, 2006.
45. Hutchins W. J. Machine translation: past, present, future. – Chichester: Ellis Horwood, 1986. – 66 p.
46. Indurkhya N., Damerau F. J. (ed.). Handbook of natural language processing. – CRC Press, 2010. – Т. 2.
47. Johnson M. Attribute-value logic and the theory of grammar. – Center for the Study of Language and Information, 1988.
48. Koehn P. Statistical machine translation. – Cambridge University Press, 2009.

49. Liberman A. M. Speech: A special code. – MIT press, 1996.
50. Mercer R. L. Partial-Wave Analysis of Elastic-Scattering and Inelastic-Scattering of Dirac Particles. – University of Illinois at Urbana-Champaign, 1972.
51. Nirenburg S., Somers H. L., Wilks Y. (ed.). Readings in machine translation. – MIT Press, 2003.
52. Schroeder M. Being for: Evaluating the semantic program of expressivism. – OUP Oxford, 2010.
53. Zen H. et al. Libritts: A corpus derived from librispeech for text-to-speech //arXiv preprint arXiv:1904.02882. – 2019.
54. Шокиров Т. С. Қомусҳои электрони ва вежагиҳои онҳо [Матн] / Шокиров Т.С. //Вестник Таджикского государственного университета права, бизнеса и политики. Серия гуманитарных наук. – 2021. – №. 3 (88). – С. 131-138.
55. Анисимов А. В. Система обработки текстов на естественном языке [Текст] / Анисимов А.В., Марченко А.А. //Искусственный интеллект. – 2002. – №. 4. – С. 157-163.
56. Ашурова Ш. Н. О распознавании автора текста на основе частотности словесных биграмм [Текст] / Ашурова Ш.Н., Тошхуджаев Х.А. //Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2020. – №. 2 (50). – С. 57.
57. Ашурова Ш. Н. Оценка эффективности использования словесных триграмм при идентификации текста [Текст] / Ашурова Ш.Н. // Вестник Технологического университета Таджикистана. – 2017. – №. 4. – С. 51-58.
58. Ашурова Ш. Н. Оценка эффективности использования словесных униграмм при идентификации текста [Текст] / Ашурова Ш. Н., Косимов А. А. //Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2017. – №. 2. – С. 49-54.
59. Ашурова, Ш. Н. О распознавании автора текста на основе частотности словесных униграмм [Текст] / Ш. Н. Ашурова // Вестник ПИТТУ имени академика М.С. Осими. – 2020. – № 3(16). – С. 7-15.

60. Бахтеев К. С. Автоматическая символьная предобработка текстов таджикского языка [Текст] / Бахтеев К. С. // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2012. – №. 4. – С. 37-40.
61. Бахтеев К. С. О применимости укороченных цифровых портретов для идентификации автора текста [Текст] / Бахтеев К. С. // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2020. – №. 2 (50). – С. 25.
62. Бахтеев К. С. О распознавании авторства по усечённым цифровым портретам текста [Текст] / Бахтеев К. С. //Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2018. – №. 4 (173). – С. 82.
63. Белоногов Г. Г. Метод аналогии в компьютерной лингвистике [Текст] / Белоногов Г. Г., Зеленков Ю.Г., Новоселов А.П. //Научно-техническая информация. Сер. 2. – 2000. – №. 1. – С. 21.
64. Быстров И. И. и др. Основы применения онтологии и компьютерной лингвистики при проектировании перспективных автоматизированных информационных систем [Текст] / И.И. Быстров, Б.В. Тарасов, А.А. Хорошилов, С.И. Радоманов //Системы и средства информатики. – 2015. – Т. 25. – №. 4. – С. 128-149.
65. Галунов В. И. Акустическая коммуникация, речь и передача смысловой информации [Текст] / Галунов В. И. //Русский орнитологический журнал. – 2009. – Т. 18. – №. 484. – С. 813-824.
66. Гмурман В. Е. Руководство к решению задач по теории вероятностей и математической статистике [Текст] / Гмурман В. Е., Гмурман В. В., Колосова Т. В. // – Общество с ограниченной ответственностью Издательство ЮРАЙТ, 2015. – С. 404-404.
67. Гращенко Л. А. и др. Концептуальная модель системы русско-таджикского машинного перевода [Текст] / Гращенко Л.А. //Доклады Академии наук Республики Таджикистан. – 2011. – Т. 54. – №. 4. – С. 279-285.

68. Гращенко Л. А. Клиент удаленной автоматизации согласования компьютерных шрифтов таджикского языка [Текст] / Гращенко Л.А. // Доклады Академии наук Республики Таджикистан. – 2011. – Т. 54. – №. 5. – С. 367-370.
69. Гращенко Л. А. Концептуальная модель таджикско-персидской конверсии графических систем письма [Текст] / Гращенко Л.А. //Доклады Академии наук Республики Таджикистан. – 2008. – Т. 52. – №. 2. – С. 111-115.
70. Гращенко Л. А. Модельный стоп-словарь таджикского языка [Текст] / Гращенко Л.А. // Доклады Академии наук Республики Таджикистан. – 2013. – Т. 56. – №. 5. – С. 368-375.
71. Гращенко Л.А. Алгоритм формирования словаря соответствий таджикских и персидских словоформ [Текст] / Гращенко Л.А. //Доклады Академии наук Республики Таджикистан. – 2008. – Т. 51. – №. 5. – С. 339-345.
72. Гращенко Л.А. Опыт автоматизированного анализа повторов в научных текстах [Текст] / Гращенко Л. А., Романишин Г. В. //Новые информационные технологии в автоматизированных системах. – 2015. – №. 18. – С. 582-590.
73. Гуломсафдаров А.Г. О многообразии слогов шугнанского языка [Текст] / Гуломсафдаров А.Г. //Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2010. – №. 1. – С. 49-52.
74. Довудов Г. М. Алгоритм автоматического морфологического анализа таджикских слов [Текст] / Довудов Г. М. //Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2010. – №. 2. – С. 22-26.
75. Довудов, Г. М., Автоматический синтез таджикских словоформ имени существительного [Текст] / Довудов Г. М., Назаров А. А. // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2019. – №. 3. – С. 31-35.

76. Дроздова К. А. Машинный перевод: история, классификация, методы [Текст] / Дроздова К.А. //Вестник Омского государственного педагогического университета. Гуманитарные исследования. – 2015. – №. 3 (7). – С. 156-158.
77. Дулесов А. С. Применение формулы Шеннона и геометрического обобщения для определения энтропии [Текст] / Дулесов А. С., Швец С. В., Хрусталёв В. И. // Перспективы науки. – 2010. – №. 3. – С. 92-95.
78. Евсеева И. В. Когнитивное моделирование комплексных единиц дериватологии [Текст] / Евсеева И. В. //Актуальные проблемы современного словообразования. – 2011. – С. 163-170.
79. Евсеева И. В. Комплексные единицы словообразовательной системы [Текст] / Евсеева И. В. //Вестник Кемеровского государственного университета. – 2011. – №. 3. – С. 188-194.
80. Жиляков Е.Г. Сегментация речевых сигналов на основе анализа распределения энергии по частотным интервалам [Текст] / Е.Г. Жиляков, Е.И. Прохоренко, А.В. Болдышев, А.А. Фирсова, М.В. Фатова // Научные ведомости Белгородского государственного университета. Серия: История. Политология. Экономика. Информатика, Том 18. – 2011. – №7-1 (102). – С. 187-196
81. Заболеева-Зотова А. В. Формализация семантики текста при автоматизации слабоструктурируемых процедур в процессе синтеза технических систем [Текст] / Заболеева-Зотова А. В. //Известия Волгоградского государственного технического университета. – 2006. – №. 4. – С. 36-43.
82. Загоруйко Н. Г. и др. Система ontogrid для автоматизации процессов [Текст] / Н. Г. Загоруйко, В. Д. Гусев, А. В. Завертайлов, С. П. Ковалев, А. М. Налетов, Н. В. Саломатина //Автометрия. – 2005. – Т. 41. – №. 5. – С. 13.
83. Загоруйко Н. Г. Об исследованиях проблемы речевых технологий [Текст] / Загоруйко Н. Г. //Речевые. – 2008.
84. Зализняк А. А. Лингвоспецифичные единицы русского языка в свете контрастивного корпусного анализа [Текст] / Зализняк А. А. //Компьютерная

лингвистика и интеллектуальные технологии. – 2015. – Т. 1. – №. 14 (21). – С. 683.

85. Зарипов С. А. Synthesis model of tajik simple sentence [Текст] / Зарипов С. А. //Вестник Таджикского государственного университета коммерции. – 2020. – №. 1. – С. 295-304.
86. Зарипов С. А. Контекстные модели форм глагола v-ing на таджикском языке [Текст] / Зарипов С. А. //Вестник Технологического университета Таджикистана. – 2015. – №. 1. – С. 47-50.
87. Зарипов С. А. Контекстный анализ и синтез форм глагола [Текст] / Зарипов С. А. //Вестник Технологического университета Таджикистана. – 2013. – №. 2. – С. 54-56.
88. Зарипов С. А. Концепция англо-таджикского двустороннего перевода предикативных основ [Текст] / Зарипов С. А. //Технологический университет Таджикистана-Душанбе-2005. -Библиогр. – 2005. – Т. 1.
89. Зарипов С. А. Об одной модели таджикско-русского перевода простого распространенного перевода [Текст] / Зарипов С. А., Эвазов Х. А. //Труды Технологического университета Таджикистана. – 2005. – №. XII.
90. Зарипов С. А. Таджикские предложно-изофатные модели [Текст] / Зарипов С. А., Ризвонова У. М. //Вестник Таджикского национального университета. Серия естественных наук. – 2018. – №. 2. – С. 58-63.
91. Зарипов С.А. О классификации и моделировании саджа [Текст] / Зарипов С. А., Ниёзбокиев О. С. //Вестник Таджикского национального университета. Серия филологических наук. – 2020. – №. 5. – С. 300-304.
92. Захаров В. Н. Автоматическая оценка подобия тематического содержания текстов на основе сравнения их формализованных смысловых описаний [Текст] / Захаров В. Н., Хорошилов А. А. //Труды. – 2012. – С. 189-195.
93. Зеленков Ю. Г. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов [Текст] / Зеленков Ю. Г., Сегалович И. В., Титов В. А. //Компьютерная лингвистика и

интеллектуальные технологии. Труды международного семинара Диалог. – 2005. – Т. 2005. – С. 188-197.

94. Исмаилов М. А. Алгоритм автоматизированного разбиения слов таджикского языка на слоги [Текст] / Исмаилов М. А. //Доклады АН РТ. – 2000. – Т. 43. – №. 3. – С. 95-99.
95. Исмаилов М. А. Алгоритм определения ударного слога в таджикских словах при отсутствии приставок [Текст] / Исмаилов М. А. //Вестник Технологического университета Таджикистана. – 2013. – №. 2. – С. 49-51.
96. Исмаилов М. А. Алгоритмы анализа префиксов и словоформ, образованных из основ существительных и прилагательных [Текст] / Исмаилов М. А., Гуломсафдаров А. Г.// Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2019. – №. 1. – С. 15-20.
97. Исмаилов М. А. Основы автоматизированного морфологического анализа слов таджикского языка [Текст] / Исмаилов М. А. //Institute of Mathematics of the Academy of Sciences of Tajikistan. – 1994.
98. Исмаилов М. А. Разработка модели словообразования в шугнанском языке [Текст] / Исмаилов М. А., Гуломсафдаров А. Г. //Вестник Таджикского национального университета. Серия естественных наук. – 2016. – №. 1-1. – С. 105-107.
99. Исмоилов М. А. Математическая модель морфологического анализа и синтеза слов таджикского языка [Текст] / Исмоилов М. А. //Доклады АН РТ. – 1998. – Т. 41. – №. 9. – С. 63.
100. Караулов Ю. Н. Лингвистические основы функционального подхода в литературоведении [Текст] / Караулов Ю. Н. //Проблемы структурной лингвистики. – 1982. – Т. 1980. – С. 20-37.
101. Карневская, Е. Б. Взаимодействие аллофонического и свободного варьирования согласных в английской спонтанной речи [Текст] / Карневская Е. Б., Долматова Е. Д. // Сучасні тенденції фонетичних досліджень : збірник матеріалів III Круглого столу з міжнародною участю (19 квітня 2019 р.). – Київ : КПІ ім. Ігоря Сікорського, 2019. – С. 16–20

102. Козеренко, Е.Б. Проектирование многоязычного лингвистического ресурса для систем машинного перевода и обработки знаний [Текст] / Козеренко Е.Б., Лунева Н.В., Морозова Ю.И., Ермаков П.В. // Системы и средства информатики. - М.: «Наука», – 2009. – Вып. 19. – С. 119-141.
103. Козеренко, Е.Б. Лингвистические и металингвистические представления в интеллектуальных многоязычных системах [Текст] / Козеренко Е.Б., Лунева Н.В., Галина И.В., Морозова Ю.И. // Журнал «Искусственный интеллект». - НАН Украины, – 2011. – Том 3. – С. 123-135.
104. Ковалев И.В. Система поиска, анализа и обработки мультилингвистических текстов, интегрированная с информационно-поисковыми системами [Текст] / И. В. Ковалев, , К. В. Полянский, , П. В. Зеленков, , В. В. Брезицкая, Г. А. Сидорова //Сибирский аэрокосмический журнал. – 2013. – №. 1 (47). – С. 48-52.
105. Косимов А. А. О распознавании автора текста на основе частотности буквенных триграмм [Текст] / Косимов А. А. //Вестник ПИТТУ имени академика МС Осими. – 2019. – №. 4. – С. 28-37.
106. Косимов А. А. О распознавании автора текста на основе частотности длин предложений [Текст] / Косимов А. А., Бахтеев К. С. //Доклады Академии наук Республики Таджикистан. – 2020. – Т. 63. – №. 3-4. – С. 180-186.
107. Косимов А. А. О распознавании автора текста на узбекском языке с помощью символьных триграмм [Текст] / Косимов А. А., Зульфикарова П. Э. // Вестник ПИТТУ имени академика МС Осими. – 2020. – №. 2. – С. 24-31.
108. Косимов А. А. Оценка эффективности использования биграмм при идентификации текста [Текст] / Косимов А. А. // Доклады Академии наук Республики Таджикистан. – 2017. – Т. 60. – №. 5-6. – С. 224-229.
109. Косимов А. А. Оценка эффективности использования униграмм при идентификации текста [Текст] / Косимов А. А. //Доклады Академии наук Республики Таджикистан. – 2017. – Т. 60. – №. 3-4. – С. 132-137.
110. Косимов А. А. Применение специфичного цифрового портрета для идентификации авторов произведений [Текст] / Косимов А. А., Бахтеев К. С.

//Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2019. – №. 3. – С. 7-11.

111. Котельников Е. В. Автоматический анализ тональности текстов на основе методов машинного обучения [Текст] / Котельников Е. В., Клековкина М. В. //Компьютерная лингвистика и интеллектуальные технологии. – 2012. – Т. 2. – №. 11. – С. 27.
112. Котельников Е. В. Определение весов оценочных слов на основе генетического алгоритма в задаче анализа тональности текстов [Текст] / Котельников Е. В., Клековкина М. В. //Программные продукты и системы. – 2013. – №. 4. – С. 296-300.
113. Кривнова О. Ф. Генерация тонального контура фразы в системах автоматического синтеза речи [Текст] / Кривнова О. Ф. //Труды Международного семинара по компьютерной лингвистике и ее приложениям «Диалог. – 2000. – С. 211-220.
114. Кривнова О. Ф. Речевые корпусы (опыт разработки и использование) [Текст] / Кривнова О. Ф., Захаров Л. М., Строкин Г. С. //Сборник трудов Международного семинара Диалог. – 2001. – С. 230-236.
115. Лесников С. В. Фреймовое конструирование тезауруса метаязыка лингвистики [Текст] / Лесников С. В. // Вестник Северного (Арктического) федерального университета. Серия: Гуманитарные и социальные науки. – 2011. – №. 4. – С. 84-88.
116. Лобанов Б. М. Алгоритм сегментации текста на синтаксические синтагмы для синтеза речи [Текст] / Лобанов Б. М. // Труды Международной конференции «Компьютерная лингвистика и интеллектуальные технологии» (Диалог’2008).–М.: Наука. – 2008. – С. 323-529.
117. Лобанов Б. М. В. Синтезатор речи по тексту как компьютерное средство «клонирования» персонального голоса [Текст] / Лобанов Б. М., Карневская Е. Б., Левковская Т. //Труды Международной конференции Диалог-2001.–М. – 2001. – С. 265-272.

118. Лобанов Б. М. и др. Фонетико-акустическая база данных для многоязычного синтеза речи по тексту на славянских языках [Текст] / * // Компьютерная лингвистика и интеллектуальные технологии": труды междунар. конф. Диалог. – 2006. – С. 357-363.
119. Людовик Т. В. Автоматическое распознавание спонтанной украинской речи (на материале акустического корпуса украинской эфирной речи) [Текст] / Людовик Т. В., Пилипенко В. В., Робейко В. В. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог»(Бекасово, 25-29 мая 2011 г.). – 2011. – №. 10. – С. 17.
120. Людовик Т. В. Знаки пунктуации в текстах, получаемых при автоматическом распознавании речи [Текст] / Людовик Т. В. //Пятый междисциплинарный семинар" Анализ разговорной русской речи"(АРЗ-2011). – 2011. – С. 19-26.
121. Мамадназаров А. Современная русско-таджикская специальная лексикография [Текст] / Мамадназаров А. //Вестник института языков. – 2014. – №. 1. – С. 6-13.
122. Мамадназаров А. Современные лексические словари английского, русского и таджикского языков [Текст] / Мамадназаров А. //Забон: таҳқиқ ва таълим. – 2019. – С. 14-28.
123. Минаков И. А. Автоматизированное пополнение онтологии на основе знаний, извлеченных в процессе кластеризации [Текст] / Минаков И. А. //Вестник Самарского государственного технического университета. Серия: Технические науки. – 2005. – №. 33. – С. 321-326.
124. Минаков И. А. Сравнительный анализ некоторых методов случайного поиска и оптимизации [Текст] / Минаков И. А. //Известия Самарского научного центра Российской академии наук. – 1999. – Т. 1. – №. 2. – С. 286-293.
125. Михайлов Д. В. Семантическая кластеризация текстов предметных языков (морфология и синтаксис) [Текст] / Михайлов Д. В., Емельянов Г. М. // Компьютерная оптика. – 2009. – Т. 33. – №. 4. – С. 473-480.

126. Михайлов Д. В., Емельянов Г. М. К вопросу автоматизации пополнения базы данных лексических функций в задаче установления смысловой эквивалентности текстов естественного языка [Текст] / Михайлов Д. В., Емельянов Г. М. // Вестник Новгородского государственного университета им. Ярослава Мудрого. – 2007. – №. 44. – С. 45-49.
127. Мурзин Ф. А. и др. Методы определения степени близости предложений на естественном языке на основе грамматики связей [Текст] / Мурзин Ф.А., Батура Т.В., Еримбетова А.С., Бакиева А.М. //Наука и мир. – 2015. – №. 3-2. – С. 61-67.
128. Назаров А. А. Автоматический синтез таджикских словоформ имени прилагательного [Текст] / Назаров А. А. // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2019. – №. 4. – С. 16-18.
129. Нариньяни А. С. ТЕОН-2: от Тезауруса к Онтологии и обратно [Текст] / Нариньяни А. С. //Труды Международного семинара «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука. – 2002. – Т. 1. – С. 199-154.
130. Одинаев А. А. Основные направления прикладных лингвистических исследований в Республике Таджикистан [Текст] / Одинаев А. А. // Вестник Московского государственного лингвистического университета. Гуманитарные науки. – 2015. – №. 25 (736). – С. 41-48.
131. Одинаев А. А. Прикладные лингвистические исследования в республике Таджикистан: состояние и перспективы [Текст] / Одинаев А. А. //Вестник Педагогического университета. – 2015. – №. 4 (65). – С. 73-78.
132. Погорелов Д. А. Сравнительный анализ алгоритмов редакционного расстояния Левенштейна и Дамерау-Левенштейна [Текст] / Погорелов Д. А., Таразанов А. М., Волкова Л. Л. // Синергия наук. – 2019. – №. 31. – С. 803-811.
133. Потапова Р. К. Особенности исследования текста в эпоху цифровой коммуникации [Текст] / Потапова Р. К., Курьянова И. В. //Вестник Волгоградского государственного университета. Серия 2: Языкознание. – 2021. – Т. 20. – №. 2. – С. 5-15.

134. Пруцков А. В. Математико-алгоритмическая формализация моделей морфологического анализа и синтеза словоформ естественных языков [Текст] / Пруцков А. В. // Cloud of science. – 2018. – Т. 5. – №. 4. – С. 729-748.
135. Пруцков А. В. Методы поиска решений в лингвистических автоматизированных обучающих системах [Текст] / Пруцков А. В. //Научно-техническая информация. Серия 2: Информационные процессы и системы. – 2006. – №. 4. – С. 15-18.
136. Рубашкин В. Ш. Словарная поддержка процедур семантической интерпретации предложных связей [Текст] / Рубашкин В. Ш. //Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции" Диалог. – 2005. – С. 430-435.
137. Сажок Н. Н. Речевые информационные технологии и системы [Текст] / Сажок Н. Н. //Управляющие системы и машины. – 2017. – №. 2. – С. 38-45.
138. Сбоев А. Г. Модель системы синтаксического анализа текстов естественного языка на основе статистически отобранных наборов параметров слов [Текст] / Сбоев А.Г., Рыбка Р.Б., Иванов И. И.3, Гудовских Д.В., Молошников И.А., Кукин К.А., Власов Д.С. //Современные информационные технологии и ИТ-образование. – 2013. – №. 9. – С. 422-432.
139. Сбоев А. Г. Продвинутые нейросетевые модели для решения задачи определения тональности [Текст] / А. Г. Сбоев, И. Е. Воронина, Д. В. Гудовских, А. А. Селиванов //Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. – 2016. – №. 4. – С. 178-183.
140. Смирнов С. В. О понятиях «научная школа» и «научное направление» в истории языкознания [Текст] / Смирнов С. В. //Учёные записки Тартуского государственного университета. – 1981. – №. 573. – С. 136-147.
141. Собиров Д. Д. Информационные основы автоматического распознавания глаголов таджикского языка [Текст] / Собиров Д. Д., Гращенко Л. А., Усманов З. Д. //Известия Академии наук Республики Таджикистан. Отделение физико-

математических, химических, геологических и технических наук. – 2011. – №. 3. – С. 41-46.

142. Солиев О. М. О раскладке таджикских букв на компьютерной клавиатуре по схеме русской раскладки [Текст] / Солиев О. М. // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2007. – №. 2. – С. 26-30.
143. Сорокин В. Н. Совместные приближения корня, логарифма и арксинуса [Текст] / Сорокин В. Н. // Вестник Московского университета. Серия 1. Математика. Механика. – 2009. – №. 2. – С. 65-69.
144. Сорокин В. Н. Фундаментальные исследования речи и прикладные задачи речевых технологий [Текст] / В. Н. Сорокин // Речевые технологии. –2008. – №1. С. 18-48.
145. Сулейманов Д. Ш. и др. Многофункциональная модель тюркской морфемы как база данных для лингвопроцессоров [Текст] / Д.Сулейманов, А.Гатиатуллин, А.Альменова, А.Баширов //Филология и культура. – 2016. – №. 2 (44). – С. 143-151.
146. Сулейманов Д. Ш. Корпус татарского языка: концептуальные и лингвистические аспекты [Текст] / Сулейманов Д. Ш., Хакимов Б. Э., Гильмуллин Р. А. //Филология и культура. – 2011. – №. 26. – С. 211-216.
147. Сулейманов Д. Ш. Система семантического анализа ответных текстов обучаемого на естественном языке [Текст] / Сулейманов Д. Ш. //Онтология проектирования. – 2014. – №. 1 (11). – С. 65-77.
148. Усманов З. Д. Алгоритм компьютерного перевода простого нераспространенного английского предложения на таджикский язык [Текст] / Усманов З. Д., Исмаилов М. А., Зарипов С. А. //ДАН РТ. – 2002. – Т. 45. – №. 3-4. – С. 81.
149. Усманов З. Д. Алгоритм настройки кластеризатора дискретных случайных величин [Текст] / Усманов З. Д. //Доклады Академии наук Республики Таджикистан. – 2017. – Т. 60. – №. 9. – С. 392-397.

150. Усманов З. Д. Закономерности статистического распределения частот встречаемости букв в таджикском языке [Текст] / Усманов З. Д., Солиев О. М. //ДАН РТ. – 2003. – Т. 46. – №. 3-4. – С. 59-62.
151. Усманов З. Д. Закономерности статистического распределения частот встречаемости букв в таджикском языке [Текст] / Усманов З. Д., Солиев О. М. //ДАН РТ. – 2003. – Т. 46. – №. 3-4. – С. 59-62.
152. Усманов З. Д. и др. О статистических закономерностях языка эсперанто [Текст] / Усманов З. Д. //Доклады Академии наук Республики Таджикистан. – 2006. – Т. 49. – №. 4. – С. 316-320.
153. Усманов З. Д. Информационные основы автоматизированной таджикско-персидской транслитерации [Текст] / Усманов З. Д., Гращенко Л. А., Фомин А. Ю. //Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2008. – №. 1. – С. 20-26.
154. Усманов З. Д. К вопросу о наилучших раскладках английских и русских символов на компьютерной клавиатуре [Текст] / Усманов З. Д., Солиев О. М. //Программные продукты и системы. – 2004. – №. 4. – С. 41-44.
155. Усманов З. Д. Классификатор дискретных случайных величин [Текст] / Усманов З. Д. //Доклады Академии наук Республики Таджикистан. – 2017. – Т. 60. – №. 7-8. – С. 291-300.
156. Усманов З. Д. Кодирование предложений [Текст] / Усманов З. Д. //Доклады Академии наук Республики Таджикистан. – 2013. – Т. 56. – №. 5. – С. 365-367.
157. Усманов З. Д. Компьютерная коррекция таджикского текста, набранного без использования специфических букв [Текст] / Усманов З. Д., Эвазов Х. А. // Доклады Академии наук Республики Таджикистан. – 2011. – Т. 54. – №. 1. – С. 23-26.
158. Усманов З. Д. Компьютерная коррекция таджикского текста, набранного без использования специфических букв [Текст] / Усманов З. Д., Эвазов Х. А. //Доклады Академии наук Республики Таджикистан. – 2011. – Т. 54. – №. 1. – С. 23-26.

159. Усманов З. Д. Моделирование восприятия мозгом анаграммно искаженного текста [Текст] / Усманов З. Д. //Программные продукты и системы. – 2018. – Т. 31. – №. 3. – С. 448-454.
160. Усманов З. Д. О “наилучшей” раскладке таджикских букв на компьютерной клавиатуре [Текст] / Усманов З. Д., Солиев О. М. //ДАН РТ. – 2004. – Т. 47. – №. 3. – С. 56.
161. Усманов З. Д. О влиянии цифрового портрета текста на распознавание автора произведения [Текст] / Усманов З. Д., Косимов А. А. //Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2020. – №. 3. – С. 36-42.
162. Усманов З. Д. О множестве постфиксов таджикского литературного языка [Текст] / Усманов З. Д., Солиев О. М., Довудов Г. М. //Доклады Академии наук Республики Таджикистан. – 2010. – Т. 53. – №. 2. – С. 99-103.
163. Усманов З. Д. О применимости-классификатора к распознаванию авторства и тематики художественных произведений [Текст] / Усманов З. Д., Косимов А. А. // Новые информационные технологии в автоматизированных системах. – 2019. – №. 22. – С. 174-178.
164. Усманов З. Д. О распознавании авторства таджикского текста [Текст] / Усманов З. Д., Косимов А. А. //Доклады Академии наук Республики Таджикистан. – 2016. – Т. 59. – №. 3-4. – С. 114-119.
165. Усманов З. Д. О слоговой структуре слов шугнанского языка [Текст] / Усманов З. Д., Гуломсафдаров А. Г. //Доклады Академии наук Республики Таджикистан. – 2009. – Т. 52. – №. 9. – С. 681-684.
166. Усманов З. Д. О статистических закономерностях морфемной базы таджикского языка [Текст] / Усманов З. Д., Довудов Г. М. //Доклады Академии наук Республики Таджикистан. – 2010. – Т. 53. – №. 3. – С. 188-191.
167. Усманов З. Д. О статистических закономерностях слогового многообразия таджикского языка [Текст] / Усманов З. Д., Абдухамидов А. А., Исмаилов М. А. //ДАН РТ. – 2002. – Т. 45. – №. 5-6. – С. 9.

168. Усманов З. Д. О формировании базы префиксов таджикского литературного языка [Текст] / Усманов З. Д., Довудов Г. М. //Доклады Академии наук Республики Таджикистан. – 2009. – Т. 52. – №. 6. – С. 431-436.
169. Усманов З. Д. Об анаграммах словоформных N-грамм [Текст] / Усманов З. Д. //Доклады Академии наук Республики Таджикистан. – 2020. – Т. 63. – №. 1-2. – С. 43-48.
170. Усманов З. Д. Об одном обобщении формулы золотого сечения [Текст] / Усманов З. Д. //Доклады Академии наук Республики Таджикистан. – 2014. – Т. 57. – №. 1. – С. 5-8.
171. Усманов З. Д. Об одном цифровом портрете текста и его приложении [Текст] / Усманов З. Д. //Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2019. – №. 3. – С. 35-38.
172. Усманов З. Д. Об оптимальной раскладке символов на клавиатуре [Текст] / Усманов З. Д. //Программные продукты и системы. – 2004. – №. 2. – С. 41-45.
173. Усманов З. Д. Об упорядоченном алфавитном кодировании слов естественных языков [Текст] / Усманов З. Д. //Доклады Академии наук Республики Таджикистан. – 2012. – Т. 55. – №. 7. – С. 545-548.
174. Усманов З. Д. Обзор результатов по применению гамма-классификатора [Текст] / Усманов З. Д. //Известия Национальной академии наук Таджикистана. Отделение физико-математических, химических, геологических и технических наук. – 2021. – №. 3 (184). – С. 62.
175. Усманов З. Д. Основные достижения и перспективы научной школы Таджикистана по вычислительной лингвистике [Текст] / Усманов З. Д. //Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2017. – №. 1. – С. 44-50.
176. Усманов З. Д. Оценка эффективности применения γ-классификатора для атрибуции печатного текста [Текст] / Усманов З. Д. //Доклады академии наук Республики Таджикистан. – 2020. – Т. 63. – №. 3-4. – С. 172-179.

177. Усманов З. Д. Распознавание словоформ таджикского языка [Текст] / Усманов З. Д., Исмаилов М. А., Гафуров Д. А. //ДАН РТ. – 2002. – Т. 45. – №. 5-6. – С. 4.
178. Усманов З. Д. Частотность букв таджикской литературы [Текст] / Усманов З. Д., Косимов А. А.//Доклады Академии наук Республики Таджикистан. – 2015. – Т. 58. – №. 2. – С. 112-115.
179. Усманов З. Д. Частотный метод устранения омонимии таджикских словоформ [Текст] / Усманов З. Д., Довудов Г. М. //Доклады Академии наук Республики Таджикистан. – 2017. – Т. 60. – №. 1-2. – С. 36-41.
180. Усманов З. Д. Частотный морфемный словарь таджикского литературного языка [Текст] / Усманов З. Д., Довудов Г. М. // Доклады Академии наук Республики Таджикистан. – 2010. – Т. 53. – №. 4. – С. 257-262.
181. Усманов З. Д., Довудов Г. М. Концептуальная модель автоматического морфологического анализа таджикских словоформ [Текст] / Усманов З. Д., Довудов Г. М. // Доклады Академии наук Республики Таджикистан. – 2014. – Т. 57. – №. 3. – С. 205-209.
182. Цирульник Л. И. Автоматизированная система клонирования фонетико-акустических характеристик речи [Текст] / Цирульник Л. И. // Информатика. – 2018. – №. 2 (10). – С. 46-55.
183. Цирульник, Л.И. Алгоритмы синтеза просодических характеристик речи по тексту в системе «Мультифон» [Текст] / Л.И. Цирульник, Д.В. Жадинец, Б.М. Лобанов, О.Г. Сизонов // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог’2007, Бекасово, 30 мая – 3 июня 2007 г. – М.: Издательский центр РГГУ, 2007. – С. 550-558.
184. Черник, Н. Н. Сегментация спонтанной речи в языках различных типов [Текст] / Н.Н. Черник // Вестник Белорусского государственного экономического университета. - 2009 - N 4 - С. 101-107.
185. Чучупал В.Я. Диалоговая система цифровой обработки зашумленных речевых сигналов [Текст] / Чучупал В.Я. // В кн.: Тезисы докладов Всесоюзной школы-семинара АРС0-I3, Новосибирск, 1984, с.116.

186. Чучупал В.Я. Реализация метода вычитания спектров для повышения качества и разборчивости речи [Текст] / Чучупал В.Я. // В кн.: Тезисы докладов Всесоюзной школы-семинара АРС0-I2, Киев-Одесса, 1982, с.155-156.
187. Шереметьева, С.О. Методы и модели автоматического извлечения ключевых слов [Текст] / С.О. Шереметьева, П.Г. Осминин //Вестник Южно-Уральского государственного ун-та. –2015. – № 1, т.12. – С. 76-81.
188. Шокиров Т. С. Переводная лексикография и ее особенности [Текст] / Шокиров Т. С., Ахмеджонова М. А. //Вестник Педагогического университета. – 2019. – №. 6 (83). – С. 67-73.
189. Эвазов Х. А. О структуре сложных слов современного таджикского литературного языка [Текст] / Эвазов Х. А. //Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2010. – №. 1. – С. 53-59.
190. Эвазов Х. А. О таджикском компьютерном корректоре [Текст] / Эвазов Х. А. //Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2018. – №. 1. – С. 29-32.
191. Эвазов Х. А. Статистические закономерности таджикского языка, связанные с используемым в нем расширенным кириллическим алфавитом [Текст] / Эвазов Х. А. //Доклады Академии наук Республики Таджикистан. – 2010. – Т. 53. – №. 12. – С. 903-906.
192. Abney S. Partial parsing via finite-state cascades [Text] / Abney S. // J. Natural Language Eng.2. 1996. pp.337–344
193. Ahmed F. Arabic/english word translation disambiguation using parallel corpora and matching schemes [Text] / Ahmed F., Nürnberger A. //Proceedings of the 12th Annual conference of the European Association for Machine Translation. – 2008. pp. 6-11.
194. Ahmed F. Revised n-gram based automatic spelling correction tool to improve retrieval effectiveness [Text] / Ahmed F., Luca E. W. D., Nürnberger A. //Polibits. – 2009. – №. 40. pp. 39-48.

195. Angell R. C. Automatic spelling correction using a trigram similarity measure [Text] / Angell R. C., Freund G. E., Willett P. //Information Processing & Management. – 1983. – Т. 19. – №. 4. pp. 255-261.
196. Berghel H. L. A logical framework for the correction of spelling errors in electronic documents [Text] / Berghel H. L. //Information processing & management. – 1987. – Т. 23. – №. 5. pp. 477-494.
197. Church K. Introduction to the special issue on computational linguistics using large corpora [Text] / Church K., Mercer R. L. //Computational linguistics. – 1993. – Т. 19. – №. 1. pp. 1-24.
198. Church K. W. A stochastic parts program and noun phrase parser for unrestricted text [Text] / Church K. W. //International Conference on Acoustics, Speech, and Signal Processing. – IEEE, 1989. pp. 695-698.
199. Clark A. Combining distributional and morphological information for part of speech induction [Text] / Clark A. //10th Conference of the European Chapter of the Association for Computational Linguistics. – 2003.
200. Clark A. Computational learning theory and language acquisition [Text] / Clark A., Lappin S. //Philosophy of linguistics. – 2010. pp. 445-475.
201. Clark A. The handbook of computational linguistics and natural language processing [Text] / Clark A., Fox C., Lappin S. (ed.). // – John Wiley & Sons, 2012. – Т. 118.
202. Cohen M. M. Modeling coarticulation in synthetic visual speech [Text] / Cohen M. M., Massaro D. W. //Models and techniques in computer animation. – Springer Japan, 1993. pp. 139-156.
203. Collins M. Clause restructuring for statistical machine translation [Text] / Collins M., Koehn P., Kučerová I. //Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). – 2005. pp. 531-540.
204. Collins M. et al. A statistical parser for Czech [Text] / Collins M. et al. //Proceedings of the 37th annual meeting of the Association for Computational Linguistics. – 1999. pp. 505-512.

205. Collins M. Head-driven statistical models for natural language parsing [Text] / Collins M. //Computational linguistics. – 2003. – Т. 29. – №. 4. pp. 589-637.
206. Cooper F. S. Some experiments on the perception of synthetic speech sounds [Text] / Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. //The Journal of the Acoustical Society of America. – 1952. – Т. 24. – №. 6. – С. 597-606.
207. Cooper F. S. The interconversion of audible and visible patterns as a basis for research in the perception of speech [Text] / Cooper F. S., Liberman A. M., Borst J. M. //Proceedings of the National Academy of Sciences. – 1951. – Т. 37. – №. 5. pp. 318-325.
208. Cooper, F. S. The Interconversion of Audible and Visible Patterns as a Basis for Research in the Perception of Speech [Text] / Cooper, F. S., Liberman, A. M., Borst, J. M. // Proceedings of the National Academy of Sciences, - 1951. 37(5), pp. 318-325.
209. Cunha E. L. T. P. An algorithm to identify periods of establishment and obsolescence of linguistic items in a diachronic corpus [Text] / Cunha E. L. T. P., Wichmann S. //Corpora. – 2021. – Т. 16. – №. 2. – С. 205-236.
210. Daille B. Applications of computational morphology [Text] / Daille B., Fabre C., Sébillot P. //Many morphologies. – 2002. – С. 210-234.
211. Damerau F. J. A technique for computer detection and correction of spelling errors [Text] / Damerau F. J. //Communications of the ACM. – 1964. – Т. 7. – №. 3. – С. 171-176.
212. De Luca E. W. Ontology-based semantic online classification of documents: Supporting users in searching the web [Text] / , Nürnberger A., Von-Guericke O. //Proc. of the European Symposium on Intelligent Technologies (EUNITE 2004), Aachen. – 2004.
213. Devadason F. J. A methodology for the identification of information needs of users [Text] / Devadason F. J., Lingam P. P. //IFLA journal. – 1997. – Т. 23. – №. 1. – С. 41-51.

214. Devadason F. J. Online construction of alphabetic classaurus: a vocabulary control and indexing tool [Text] / Devadason F. J. //Information processing & management. – 1985. – Т. 21. – №. 1. – С. 11-26.
215. El-Haj M. Experimenting with automatic text summarisation for arabic [Text] / El-Haj M., Kruschwitz U., Fox C. // Human Language Technology. Challenges for Computer Science and Linguistics: 4th Language and Technology Conference, LTC 2009, Poznan, Poland, November 6-8, 2009, Revised Selected Papers 4. – Springer Berlin Heidelberg, 2011. – С. 490-499.
216. George E. B. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model [Text] / George E. B., Smith M. J. T. //IEEE transactions on speech and audio processing. – 1997. – Т. 5. – №. 5. pp. 389-406.
217. Gezmu A. M. Portable spelling corrector for a less-resourced language: Amharic [Text] / Gezmu A. M., Nürnberger A., Seyoum B. E. // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). – 2018.
218. Grishman R. Information Extraction // The Oxford Handbook of Computational Linguistics [Text] / Mitkov R. (ed.). // Oxford University Press. 2003. pp. 545-559.
219. Harris K. C. et al. Speech recognition in younger and older adults: a dependency on low-level auditory cortex [Text] / Kelly C. Harris, Judy R. Dubno, Noam I. Keren, Jayne B. Ahlstrom, Mark A. Eckert // Journal of Neuroscience. – 2009. – Т. 29. – №. 19. pp. 6078-6087.
220. Hausser R. NEWCAT: parsing natural language using left-associative grammar / Hausser R. // Springer Science & Business Media, 1986. – Т. 231.
221. Hausser R. R. The syntax and semantics of English mood [Text] / Hausser R. R. //Questions and answers. – 1983. pp. 97-158.
222. Hládek D. Survey of automatic spelling correction [Text] / Hládek D., Staš J., Pleva M. //Electronics. – 2020. – Т. 9. – №. 10. p. 1670.
223. Hunt, A. J. Unit selection in a concatenative speech synthesis system using a large speech database [Text] / Hunt, A. J., Black, A.W. // In IEEE ICASSP-06, 1996. Vol. 1, pp. 373-376

224. Ide N. Introduction: Common methodologies in humanities computing and computational linguistics [Text] / Ide N., Walker D. //Computers and the Humanities. – 1992. – T. 26. – №. 5/6. pp. 327-330.
225. Johnson M. How the statistical revolution changes (computational) linguistics [Text] / Johnson M. //Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous. – 2009. – pp. 3-11.
226. Juhár J. Recent progress in development of language model for Slovak large vocabulary continuous speech recognition [Text] / Juhár J., Staš J., Hládek D. //New technologies-trends, innovations and research. – 2012. – pp. 261-276.
227. Klatt D.H. Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Listeners [Text] / Klatt D., Klatt L. // Journal of the Acoustical Society of America, 1990, JASA vol. 87 (2): pp. 820-857.
228. Klatt D. H. Review of text-to-speech conversion for English [Text] / Klatt D. H. // The Journal of the Acoustical Society of America. – 1987. – T. 82. – №. 3. pp. 737-793.
229. Klatt D. H. Speech perception: A model of acoustic–phonetic analysis and lexical access [Text] / Klatt D. H. //Journal of phonetics. – 1979. – T. 7. – №. 3. pp. 279-312.
230. Klatt, D. H. The Klattalk text-to-speech conversion system [Text] / Klatt, D. H. // In IEEE ICASSP-82, - 1982 pp. 1589-1592.
231. Koehn P. Europarl: A parallel corpus for statistical machine translation [Text] / Koehn P. //Proceedings of machine translation summit x: papers. – 2005. pp. 79-86.
232. Koehn P. Moses: Open source toolkit for statistical machine translation [Text] / Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst // Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions. – 2007. pp. 177-180.

233. Kutuzov A. Improving English-Russian sentence alignment through POS tagging and Damerau-Levenshtein distance [Text] / Kutuzov A. //Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing. – 2013. pp. 63-68.
234. Leppin H. Postscript: Border-Crossing Texts [Text] / Leppin H. //Apocryphal and Esoteric Sources in the Development of Christianity and Judaism. – Brill, 2021. pp. 610-617.
235. Levenshtein V. I. On the minimal redundancy of binary error-correcting codes [Text] / Levenshtein V. I. //Information and Control. – 1975. – Т. 28. – №. 4. pp. 268-291.
236. Lhoussain A. S. Adaptating the levenshtein distance to contextual spelling correction / Lhoussain A. S., Hicham G., Abdellah Y. //International Journal of Computer Science and Applications. – 2015. – Т. 12. – №. 1. pp. 127-133.
237. Lobanov B. Language- and speaker specific implementation of intonation contours in multilingual TTS synthesis [Text] / Lobanov B., Tsirulnik L., Zhadinets D., Karnevskaya E. // Speech Prosody: proceedings of the 3-rd International conference. Dresden: 2006. V. 2. pp. 553-556
238. Lopez A. Statistical machine translation [Text] / Lopez A. //ACM Computing Surveys (CSUR). – 2008. – Т. 40. – №. 3. pp. 1-49.
239. Madnani N. Using paraphrases for parameter tuning in statistical machine translation [Text] / Nitin Madnani, Necip Fazil Ayan, Philip Resnik, Bonnie Dorr //Proceedings of the Second Workshop on Statistical Machine Translation. – 2007. pp. 120-127.
240. Massaro D. W. Perceptual units in speech recognition [Text] / Massaro D. W. //Journal of experimental Psychology. – 1974. – Т. 102. – №. 2. p. 199.
241. Massaro D.W. Phonological context in speech perception [Text] / Massaro D. W., Cohen M.M. //Perception & psychophysics. – 1983. – Т. 34. – №. 4. pp. 338-348.
242. Masterman M. Man-aided computer translation from English into French using an on-line System to Manipulate a bi-lingual conceptual dictionary, or thesaurus

[Text] / Masterman M. //COLING 1967 Volume 1: Conference Internationale Sur Le Traitement Automatique Des Langues. – 1967.

243. Masterman M. Semantic message detection for machine translation, using an interlingua [Text] / Masterman M. //Proceedings of the International Conference on Machine Translation and Applied Language Analysis. – 1961.
244. Meersman R. The use of lexicons and other computer-linguistic tools in semantics, design and cooperation of database systems [Text] / Meersman R. //Star. – 2005. – T. 1999. – №. 02.
245. Olive, J. A set of concatenative units for speech synthesis [Text] / Olive, J., Liberman, M. // Journal of the Acoustical Society of America, 65, S130. p.1979.
246. Pujianto E. A grammatical adjustment analysis of statistical machine translation method used by google translate compared to human translation in translating English text to Indonesian. [Text] / Pujianto E. // – 2014.
247. Simard M. Pepr: Post-edit propagation using phrase-based statistical machine translation [Text] / Simard M., Foster G. //Proceedings of Machine Translation Summit XIV: Papers. – 2013.
248. Somers H. Example-based machine translation [Text] / Somers H. //Machine translation. – 1999. – T. 14. pp. 113-157.
249. Staš J. Classification of heterogeneous text data for robust domain-specific language modeling [Text] / Staš J., Juhár J., Hládek D. //EURASIP Journal on Audio, Speech, and Music Processing. – 2014. – T. 2014. – №. 1. pp. 1-12.
250. Stolcke, A. Linguistic Knowledge and Empirical Methods in Speech Recognition [Text] / A. Stolcke // AI magazine. 1997. Vol. 18, № 4. pp. 25-32.
251. Streiter O. W. Example-based NLP for minority languages: Tasks, resources and tools [Text] / Streiter O., De Luca E. //Proceedings of the Workshop “Traitement automatique des langues minoritaires et des petites langues”, 10e conference TALN. Batz-sur-Mer, France. – 2003. pp. 233-242.
252. Taylor P. Heterogeneous relation graphs as a formalism for representing linguistic information [Text] / Taylor P., Black A. W., Caley R. //Speech Communication. – 2001. – T. 33. – №. 1-2. pp. 153-174.

253. Tokuda K. Speech parameter generation algorithms for HMM-based speech synthesis [Text] / K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura // 2000 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 00CH37100). – IEEE, 2000. – T. 3. pp. 1315-1318.
254. Tomarchio J. Computer Linguistics and Philosophical Interpretation [Text] / Tomarchio J. //The Paideia Archive: Twentieth World Congress of Philosophy. – 1998. – T. 17. pp. 79-90.
255. Van Santen J. P. H. Assignment of segmental duration in text-to-speech synthesis [Text] / Van Santen J. P. H. //Computer Speech & Language. – 1994. – T. 8. – №. 2. pp. 95-128.
256. Vitale T. An algorithm for high accuracy name pronunciation by parametric speech synthesizer [Text] / Vitale T. //Computational Linguistics. – 1991. – T. 17. – №. 3. – pp. 257-276.
257. Zamora E. M. The use of trigram analysis for spelling error detection [Text] / Zamora E. M., Pollock J. J., Zamora A. //Information Processing & Management. – 1981. – T. 17. – №. 6. pp. 305-316.
258. Zen H. Statistical parametric speech synthesis [Text] / Zen H., Tokuda K., Black A.W. // Speech communication. – 2009. – T. 51. – №. 11. pp. 1039-1064.
259. Гращенко Л.А. Математические основы автоматизированной таджикско-персидской конверсии графических систем письма: Дисс. канд. физ.-мат. наук. / Гращенко, Леонид Александрович. – Ин-т математики АН Республики Таджикистан, - Душанбе, 2010. - 115 с.
260. Бахтеев К.С. Об идентификации автора текста с помощью γ - классификатора: специальность 05.13.11 – “Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей” : диссертация на соискание ученой степени кандидата технических наук / Бахтеев Камил Саидович. Таджикский технический университет имени академика М.С.Осими, – Душанбе, 2021. – 93 с.
261. Гуломсафдаров А.Г. Разработка подсистем автоматической обработки текстов шугнанского языка: специальность 05.13.11 – “Математическое и

программное обеспечение вычислительных машин, комплексов и компьютерных сетей” : диссертация на соискание ученой степени кандидата технических наук / Гуломсафдаров Абдулназар Гуломназарович. Таджикский технический университет имени академика М.С.Осими, – Душанбе, 2020. – 20 с.

262. Довудов Г.М. Компьютерный морфологический анализ таджикских словоформ: специальность 05.13.11 "Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей" : диссертация на соискание ученой степени кандидата технических наук / Довудов Гулшан Мирбахоевич. Таджикский технический университет имени академика М.С.Осими, – Душанбе, 2018. – 161 с.

263. Зарипов, С. А. Моделирование на таджикском языке английского простого нераспространенного предложения: специальность 05.13.18 "Математическое моделирование, численные методы и комплексы программ": диссертация на соискание ученой степени кандидата физико-математических наук / Зарипов Сайдахмад Асрорович. – Душанбе, 2003. – 89 с. – EDN NODVHV.

264. Косимов, А. А. Разработка основ автоматической системы распознавания автора незнакомого текста (на примере художественных произведений на таджикском языке) : специальность 05.13.11 "Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей" : диссертация на соискание ученой степени кандидата технических наук / Косимов Абдунаби Абдурауфович. Таджикский технический университет имени академика М.С.Осими, – Душанбе, 2018. – 107 с.

265. Солиев, О. М. Математическая модель оптимальной раскладки символов на клавиатуре и её приложения: специальность 05.13.18 "Математическое моделирование, численные методы и комплексы программ": диссертация на соискание ученой степени кандидата физико-математических наук / Солиев Одилходжа Махмудходжаевич. – Душанбе, 2008. – 85 с.

266. Фомин, А. Ю. Структурная типология лексико-морфологической системы таджикского языка: дис. канд. фил. наук. / Фомин, Алексей Юрьевич. – Российско-таджикский славянский университет, - Душанбе, 2010. 134 с.
267. Фомичев, В. А. Метод формального описания содержания сложных естественно-языковых текстов и его применение к проектированию лингвистических процессоров: дис. док. тех.наук. / Фомичев Владимир Александрович. – Московский государственный институт электроники и математики, -М.: 2006. 393с.
268. Проект “BookMania”. [Электронный ресурс] С.Шишминтсев. – URL: http://bookmania.com.ru (дата обращения: 21.10.2022).
269. Проект “Fonix Speech” организация Fonix Co. [Электронный ресурс] – URL: https://www.phoenixspeechtherapy.net (дата обращения: 14.11.2022).
270. Проект “Galaktika-ZOOM”, корпорация Галактика. [Электронный ресурс] – URL: www.galaktika.ru (дата обращения: 25.10.2022).
271. Проект “Govorilka”. [Электронный ресурс] / А.Рязанов. – URL: https://www.vector-ski.ru/vecs/govorilka (дата обращения: 29.10.2022).
272. Проект “Langsoft”. [Электронный ресурс] – URL: www.langsoft.ch (дата обращения: 29.10.2022).
273. Проект “Lexical FreeNet” организация Datamuse Corporation. [Электронный ресурс] – URL: http://home.istar.ca/~obyrne/dict.html (дата обращения: 27.10.2022).
274. Проект “LingSoft” [Электронный ресурс] – URL: https://www.lingsoft.fi (дата обращения: 25.10.2022).
275. Проект “MonoConc/ParaConc”. [Электронный ресурс] – URL: https://monoconc.com/ (дата обращения: 15.10.2022).
276. Проект “Mystem” организация Яндекс. [Электронный ресурс] – URL: http://web-corpora.net/wsgi/mystemplus.wsgi/mystemplus/tagger/mystem/ (дата обращения: 04.11.2022).
277. Проект “netXtract” (Relevant Software Inc.). [Электронный ресурс] – URL: https://www.netxtract.com/ (дата обращения: 08.10.2022).

278. Проект “Ngram Statistics Package – NSP”. [Электронный ресурс] – URL: https://www.d.umn.edu/~tpederse/nsp.html (дата обращения: 20.09.2022).
279. Проект “Open Office Org”. [Электронный ресурс] – URL: http://openoffice.org (дата обращения: 05.11.2022).
280. Проект “Sakrament Text-to-Speech Engine” организация “Сакрамент”. [Электронный ресурс] – URL: https://www.speechtechmag.com (дата обращения: 30.10.2022).
281. Проект “Speech Synthesis and Recognition Laboratory”. [Электронный ресурс] – URL: https://ssrlab.by (дата обращения: 02.11.2022)
282. Проект “Speech technology” Центр голосовых технологий, С-Петербург. [Электронный ресурс] – URL: https://www.speechpro.ru (дата обращения: 13.08.2022).
283. Проект “StarLing”. [Электронный ресурс] / С.А.Старостин. – URL: https://starlingdb.org/memorial/worksru.php (дата обращения: 19.08.2022).
284. Проект “Text-To-Speech Converter for MS Word”. [Электронный ресурс] – URL:https://www.softportal.com/software-4924-text-to-speech-converter-for-ms-word.html (дата обращения: 14.08.2022).
285. Проект “word2vec-toolkit” [Электронный ресурс] – URL: https://groups.google.com/g/word2vec-toolkit/ (дата обращения: 05.08.2022).
286. Проект “WordSmith Tools”. [Электронный ресурс] – URL: https://www.lexically.net/wordsmith/ (дата обращения: 13.08.2022).
287. Проект “ОРФО - Многофункциональная система проверки правописания текстов”. [Электронный ресурс] – URL: https://orfo.ru (дата обращения: 06.08.2022).
288. Проект “Речевые программы”. [Электронный ресурс] / А. Радзишевский. – URL: https://www.websound.ru (дата обращения: 25.07.2022).
289. Проект ABBYY Lingvo-11 (ABBYY Software House). [Электронный ресурс] – URL: https://www.abbyy.com (дата обращения: 26.07.2022).
290. Проект Babylon.com (Babilon.com Ltd.). [Электронный ресурс] – URL: http://online.babylon.com/combo/index.html (дата обращения: 29.09.2021).

291. Проект CMU Artificial Intelligence Repository (Carnegie Mellon University, School of Computer Science). [Электронный ресурс] – URL: https://www.cs.cmu.edu/Groups/AI/0.html (дата обращения: 08.04.2022).
292. Проект CSLU Toolkit (Center for Spoken Language Understanding). [Электронный ресурс] – URL: https://www.cs.cmu.edu/~smrobert /cu_animate.htm (дата обращения: 10.04.2022).
293. Проект "Paai's text utilities" (Dr. J.J. Paijmans, Нидерланд). [Электронный ресурс] – URL: http://paijmans.net/ (дата обращения: 13.12.2022).
294. Проект Portal Language bab.la. [Электронный ресурс] – URL: https://www.babla.ru (дата обращения: 28.04.2022).
295. Проект WordNet (Cognitive Science Laboratory, Princeton University). [Электронный ресурс] – URL: https://wordnet.princeton.edu (дата обращения: 05.10.2022).
296. Комплекс проектов ПИТ-ЗС. [Электронный ресурс] / Научно-исследовательский институт искусственного интеллекта. – URL: https://airi.net (дата обращения: 13.02.2021).
297. Электронный словарь ПРОМТ ("ПРОект МТ"). [Электронный ресурс] – URL: https://www.promt.ru (дата обращения: 06.06.2020).
298. Словари для организации Яндекс. [Электронный ресурс] – URL: https://yandex.ru/support/translate-mobile/dictionary.html (дата обращения: 28.05.2020).
299. Словари Ожегова и Зализняка. [Электронный ресурс] / С.А. Старостин. – URL: https://starlingdb.org/downl.php?lan=ru (дата обращения: 24.05.2020).
300. Онлайн словарь Lexical FreeNet (Datamuse Corporation). [Электронный ресурс] – URL: http://home.istar.ca/~obyrne/dict.html (дата обращения: 26.05.2020).
301. Онлайн словарь “ЭТС” (ETS Publishing House). [Электронный ресурс] – URL: https://www.publishersglobal.com (дата обращения: 24.04.2020).
302. Онлайн переводчик корпорации GOOGLE. [Электронный ресурс] – URL: translate.google.com (дата обращения: 24.08.2021).

303. Онлайн переводчик фирмы PEREVODOV.net (Ectaco). [Электронный ресурс] – URL: http://www.perevodov.net (дата обращения: 26.08.2021).
304. Онлайн переводчик организации ПРОМТ. [Электронный ресурс] – URL: https://www.translate.ru/перевод (дата обращения: 29.07.2021).
305. Онлайн переводчик организации Яндекс. [Электронный ресурс] – URL: translate.yandex.ru (дата обращения: 16.05.2022).
306. Автоматический словарь Мультитран. [Электронный ресурс] – URL: https://www.multitran.com (дата обращения: 14.05.2021).
307. Лемматизатор для дореформенной русской орфографии [Электронный ресурс] – URL: https://textualheritage.org/bulgarian/el-manusctipt-2012/2.html (дата обращения: 24.03.2020).
308. Apache OpenNLP. [Электронный ресурс] – URL: https://opennlp.apache.org/ (дата обращения: 29.10.2019).
309. LEO (Department of Informatics, Technische Universitat, Munchen). [Электронный ресурс] – URL: https://dict.leo.org (дата обращения: 27.10.2019).
310. MS Office. [Электронный ресурс] – URL: https://www.microsoft.com/en-us/microsoft-365/microsoft-office (дата обращения: 10.09.2019).
311. Natural Language Projects at ISI. [Электронный ресурс] – URL: https://www.isi.edu/research-groups-nlg (дата обращения: 10.09.2019).
312. PageMaker. [Электронный ресурс] – URL: https://www.pagemaker.io (дата обращения: 21.08.2020).
313. Quark XPress. [Электронный ресурс] – URL: https://www.quark.com/products/quarkxpress (дата обращения: 01.11.2020).
314. WordPerfect. [Электронный ресурс] – URL: https://www.wordperfect.com (дата обращения: 02.10.2020).
315. WordPro. [Электронный ресурс] – URL: https://www.wordprowiz.com (дата обращения: 29.10.2019).
316. Зеленков Ю.Г. Способ и система для сопоставления исходного лексического элемента первого языка с целевым лексическим элементом второго языка. [Патент] / Зеленков Ю.Г. // – 14.03.2019, – G06F 17/27.

317. Гращенко, Л.А. Таджикско-персидский конвертер графических систем письма. [SOFT] / Гращенко Л.А., Усманов З.Д., Фомин А.Ю. // – 06.03.2009, – № 091TJ.

318. Усманов, З.Д. Таджикский компьютерный морфоанализатор. [SOFT] / Усманов З.Д., Довудов Г.М., Солиев О.М. // – 2011. – ЗИ-03.2.2.220 TJ.

319. Усманов, З.Д. База данных αβ-кодов словоформ для определения автора незнакомого текста [SOFT] / Усманов, З.Д., Косимов А.А., Каюмов М.М. // – 07.06.2021, – №1202100478.

320. Усманов, З.Д. Драйвер раскладки таджикских букв на компьютерной клавиатуре TajGraph 1.0. [SOFT] / Усманов З.Д., Солиев О.М. // - 12.11.2008. - 078TJ.

321. Усманов, З.Д. Таджикский языковый пакет для системы Open Office Org. [SOFT] / Усманов З.Д., Солиев О.М., Давудов Г.М. // – 11.01.2012. – ЗИ-03.2.222 TJ.

322. Усманов, З.Д. Таджикско-русский компьютерный словарь. [SOFT] / Усманов З.Д., Холматова С.Д., Солиев О.М. // – 21.05.2007. – 025TJ.

ПУБЛИКАЦИИ 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] **Х.А. Худойбердиев** 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] **Х.А. Худойбердиев** 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.

СЛОВАРЬ КЛЮЧЕВЫХ ТЕРМИНОВ

№	Русский	Таджикский	Английский
1.	γ-классификатор Усманова	γ –таснифотгари Усмонов	Usmanov γ-classifier
2.	Автоматизированные информационные системы	Низоми (системаи) иттилоотии худкор (автоматикунонидашуда)	Automated information systems
3.	Автоматизированный перевод	Тарҷумаи худкор (мошинӣ)	Machine translation (MT)
4.	Автоматическая обработка текста	Коркарди худкори (автоматии) матн	Automatic text processing
5.	Автоматическая проверка орфографии	Санҷиши худкори имло	Automatic spell check
6.	Автоматическая система	Низоми (системаи) худкор (автоматӣ)	Automatic system
7.	Автоматическая система распознавание автора	Низоми (системаи) худкори (автоматии) шинохти муаллиф	Automatic author recognition system
8.	Автоматический (машинный) перевод	Тарҷумаи худкор	Automatic (machine) translation
9.	Автоматический графематический анализ	Таҳлили худкори (автоматии) графометрикӣ	Automatic graphematic analysis
10.	Автоматический машинный перевод	Тарҷумаи мошинии худкор (автоматишуда)	Automated machine translation
11.	Автоматический семантический анализ	Таҳлили худкори семантикӣ (мазмун)	Automatic semantic analysis
12.	Автоматический синтаксический анализ (парсинг)	Таҳлили худкори синтаксисӣ (сохтор)	Automatic parsing
13.	Автоматический словообразовательный (дериватологический) анализ	Таҳлили худкори (автоматии) калимасозӣ	Automatic word-formation (derivatological) analysis
14.	Адекватности математических моделей	Шабоҳати (баробарии) амсилаҳои математикӣ	Adequacy of mathematical models
15.	Акустическая модель	Тарҳи (амсилаи) садоӣ	Acoustic model
16.	Алгоритм поиска	Алгоритми ҷустуҷӯ	Search algorithm
17.	Алгоритм сортировки	Алгоритми мураттабгардонӣ	Sorting algorithm
18.	Алфавит	Алифбо	Alphabet
19.	Аналитические методы	Усулҳои таҳлилӣ	Analytical methods
20.	Атрибуция текста	Таносубҳои хосиятҳои матн	Text attribution (authorization)
21.	Аффикс	Морфемаи басташуда (аффикс)	Bound morpheme
22.	База данных	Манбаи (пояи) маълумот	Database

№	Русский	Таджикский	Английский
23.	Биграммный шифр	Рамзбандии (рамзгузории) биграммавӣ (дуҳарфа)	Bigram cipher
24.	Булевская (булевая, двоичная) модель	Амсилаи (модели) булӣ (дуӣ, бинарӣ)	Boolean (boolean, binary) model
25.	Векторная модель	Тарҳи (амсилаи) векторӣ	Vector pattern
26.	Вероятностная модель	Тарҳи (амсилаи) эҳтимолӣ	Probabilistic model
27.	Взрывные согласные	Ҳамсадоҳои таркишӣ	Occlusion (occlusive)
28.	Визуальное программирование	Барномасозии визуалӣ	Visual programming
29.	Восприятие (перцепционный)	Идрок (фаҳмиш)	Perception (perceptual)
30.	вычислительный эксперимент	Таҷрибаи ҳисоббарорӣ	Computational experiment
31.	Генерация естественного языка	Тавлиди забони табиӣ	Natural language generation
32.	Гипертекст	Фароматн	Hypertext
33.	Главные члены предложения	Аъзоҳои асосии ҷумла	Principal parts
34.	Грамматика	Имло (грамматика)	Grammar
35.	Грамматическая омонимия	Омонимияи грамматикӣ	Grammatical homonymy
36.	Графематический анализ	Таҳлили графемавӣ	Graphematic analysis
37.	Данные натурного эксперимента	Маълумот оиди таҷрибаи табиӣ	Field experiment data
38.	Декодер	Табдилдиҳанда (рамзкушо)	Decoder
39.	Декодирование	Табдилдиҳӣ (рамзкушоӣ)	Decoding
40.	Демо	Барномаи қаблӣ	Daemon
41.	Деривация	Ҳосилкунӣ (калимасозӣ)	Derivation
42.	Диахронический (подход)	Диахронӣ	Diachronic
43.	Дословный машинный перевод	Тарҷумаи мошинии калима ба калима (таҳтулафз)	Word-by-word machine translation
44.	Единица хранения (в корпусе)	Воҳиди нигаҳдорӣ (дар захира)	Storage unit (in case)
45.	Зависимый (элемент)	Унсури вобаста	Dependent
46.	Иерархия	Тобеият (шохаҳо)	Hierarchy
47.	Извлечение смысла из данных	Дарёфти маъно аз маълумот	Data Mining
48.	Имитационное моделирование	Амсиласозии таҷрибавӣ	Simulation systems системы
49.	Инвертированный файл индекса	Файли индекси баръакс	Inverted index file
50.	Индекс цитирования	Индекси (шохиси) иқтибос	Citation Index
51.	Индексирование	Индексиронӣ (шохисгузорӣ)	Indexing
52.	Интерлингва	Забони миёнарав	Interlingua

№	Русский	Таджикский	Английский
53.	Информационно-поисковая система	Низоми ҷустуҷӯи иттилоот	Information search system (retrieval system)
54.	Информационные системы	Низомҳои иттилоотӣ	Information Systems
55.	Информационный поиск	Ҷустуҷӯи иттилоот	Information search (retrieval)
56.	Искусственный интеллект	Зеҳни сунъӣ	Artificial intelligence
57.	Кириллица	Алифбои (навишти) кириллӣ	Cyrillic
58.	Классификация	Тасниф	Classification
59.	Кластеризация	Кластеризатсия	Clustering
60.	Ключевое слово	Калимаи калидӣ	Key word
61.	Комплексы программ	Маҷмӯи барномаҳои компютерӣ (нармафзорҳо)	Software packages
62.	Компьютерная лингвистика	Забоншиносии (лингвистикаи) ҳисобӣ (компютерӣ)	Computational linguistics
63.	Компьютерный тезаурус	Тезауруси компютерӣ	Computer thesaurus
64.	Конкорданс	Мутобиқат	Concordance
65.	Корпус лингвистический	Захираи забоншиносӣ	Corpus linguistic
66.	Корпус текстов	Маҷмӯи (захираи) матнҳо	Text corpus
67.	Корпусная лингвистика	Забоншиносии (лингвистикаи) захиравӣ	Corpus linguistics
68.	Корпусный менеджер	Менеҷери захира	Corpus manager
69.	Латинское письмо	Хати лотинӣ	Latin script
70.	Лексема	Лексема	Lexeme
71.	Лексикография	Луғатсозӣ (лексикография)	Lexicography
72.	Лексикон	Лексика	Lexicon
73.	Лексическая база данных	Манбаи (пойгоҳи) лексикӣ	Lexical data base (LDB)
74.	Лексическая функциональная грамматика	Грамматикаи амалии лексикӣ	Lexical functional grammar (LFG)
75.	Лемма	Лемма	Lemma
76.	Лемматизация	Шакли луғавии калима (лемматизатсия)	Lemmatisation
77.	Лингвистика	Забоншиносӣ (лингвистика)	Linguistics
78.	Лингвистическая статистика	Омори забоншиносӣ (лингвистикӣ)	Linguistic statistics
79.	Лингвистические ресурсы	Захираҳои (манбаҳои) забоншиносӣ (лингвистикӣ)	Language/Linguistic Resources
80.	Лингвистический спектр	Спектри забонӣ	Linguistic Spectrum
81.	Линейная интерполяция	Интерполятсияи хаттӣ	Linear interpolation
82.	Макроструктура словаря	Макросохтори луғат	Dictionary Macrostructure

№	Русский	Таджикский	Английский
83.	Математическая лингвистика	Забоншиносии (лингвистикаи) математикӣ	Mathematical linguistics
84.	Математические методы	Усулҳои математикӣ	Mathematical methods
85.	Математическое моделирование	Тарҳрезии (амсиласозии) математикӣ	Math modeling
86.	Машинное обучение	Омӯзиши худкор (мошинӣ)	Machine Learning
87.	Метаданные	Метамаълумот (маълумоти асосӣ)	Metadata
88.	Микроструктура словаря	Микросохтори луғат	Dictionary microstructure
89.	Многосложный	Бисёрҳиҷоӣ	Polysyllable (polysyllabic)
90.	Моделирования объектов	Тарҳрезии (амсиласозии) объект	Object modeling
91.	Морфема	Морфема (шакл)	Morpheme
92.	Морфология	Морфология	Morphology
93.	Нейронный машинный перевод	Тарҷумаи худкори (мошинии) нейронӣ	Neural machine translation
94.	Неоднозначность	Номуайянӣ (сермаъноӣ)	Ambiguity
95.	Обобщение	Умумиятбахшӣ	Generalisation
96.	Обработка естественного языка (ОЕЯ)	Коркарди забони табиӣ (КАТ)	Natural Language Processing (NLP)
97.	Общая семантика	Семантикаи (мазмуни) умумӣ	General Semantics
98.	Объект	Объект	Object
99.	Объектно-ориентированное программирование	Барномасозии ба объект нигаронидашуда	Object Oriented Programming
100.	Онлайн-словарь	Луғати электронӣ	Online dictionary
101.	Онтология	Онтология	Ontology
102.	Описательная грамматика	Имлои (грамматикаи) тавсифӣ	Descriptive grammar
103.	Описательный (подход)	Равиши тавсифӣ	Descriptive
104.	Оптическое распознавание символов	Ҳаммонандкунии оптикии аломатҳо (рамзҳо)	Optical Character Recognition (OCR)
105.	Параллельный корпус	Захираи (манбаи) мувозӣ	Parallel corpus
106.	Параллельный текст (битекст)	Матни мувозӣ	Bilingual corpus
107.	Парсер	Таҳлилкунанда (парсер)	Parser (parsing engine)
108.	Пауза	Таваққуф	Pause
109.	Переводной (двуязычный) словарь	Луғати дузабона	Bilingual dictionary
110.	Полнота поиска	Мукаммалии (пуррагии) ҷустуҷӯ	Search completeness
111.	Порог отображения данных	Ҳудуди инъикоси (интиқоли) маълумот	Data Display Threshold
112.	Постпозиция	Мавқеи паси калима	Postposition (postpose)

№	Русский	Таджикский	Английский
113.	Потенциальная пауза	Таваққуфи потенсиалӣ	Potential pause
114.	Предложение	Ҷумла	Sentence
115.	Представление знаний	Муаррифии дониш	Knowledge representation language (KRL)
116.	Прикладная лингвистика	Забоншиносии (лингвистикаи) амалӣ	Applied linguistics
117.	Прикладные проблемы	Масъалаҳои амалӣ	Applied problems
118.	Прикладные программные обеспечения	Таъминоти барномавии (нармафзор) амалӣ	Application software
119.	Программное обеспечение	Таъминоти нармафзор	software
120.	Проектирование программного обеспечения	Лоиҳакашии нармафзор	Software design
121.	Просодия	Хосиятҳои овозӣ (просодия)	Prosody (prosodic feature)
122.	Прямой поиск	Ҷустуҷӯи мустақим	Direct Search
123.	Разметка	Аломатгузорӣ (нишонагузорӣ)	Tagging (annotation)
124.	Разметка (маркировка)	Нишонагузорӣ	Markup
125.	Распознавание речи	Шинохти нутқ (овоз)	Speech recognition
126.	Релевантность	Мувофиқат	Relevance
127.	Репрезентативность	Пешниҳоди дастрас (мувофиқ)	Representativeness
128.	Речевой акт	Амали нутқ	Speech act
129.	Речевой сигнал	Сигнали овозӣ	Speech signal
130.	Сборник текстов	Маҷмӯи матн	Corpora
131.	Свободная морфема	Морфемаи озод	Free Morpheme
132.	Сегмент	Қисм (сегмент)	Segment
133.	Семантика	Маъно (семантика)	Semantics
134.	Семантическая сеть	Шабакаи маъноӣ (семантикӣ)	Semantic web
135.	Семантический синтез	Таҳлили маъноӣ (семантикӣ)	Semantic synthesis
136.	Синтаксическая разметка	Нишонаҳои синтаксисӣ	Syntactic tagging
137.	Синтез речи	Синтези нутқ	Speech synthesis
138.	Слияние	Якҷоясозӣ (муттаҳидшавӣ)	Merger
139.	Словарный запас автора текста	Захираи (фонди) луғавии муаллифи матн	Vocabulary of the author of the text
140.	Словарь	Луғат	Dictionary
141.	Словник	Луғат	Glossary
142.	Словослияние (или словостяжение)	Пайвастшавии калима (омезиши калима)	Portmanteau
143.	Статистика таджикского языка	Омори забони тоҷикӣ	Tajik language statistics
144.	Статистическая языковая модел	Тарҳи (амсилаи) омории забон	Statistical language model

№	Русский	Таджикский	Английский
145.	Статистический машинный перевод	Тарҷумаи мошинии (худкор) оморӣ	Statistical machine translation
146.	Стоп-слова	Калимаҳои хизматрасон	Stop words
147.	Схема деривации	Тартиби калимасозӣ	Derivation tree
148.	Тезаурус	Тезаурус	Thesaurus
149.	Термин	Истилоҳ (термин)	Term
150.	Точность поиска	Саҳеҳии (дурустии) ҷустуҷӯ	Search accuracy
151.	Транскрипция	Транскрипсия (аломатҳои овоз)	Transcription
152.	Транслитерация	Транслитератсия (табдилдиҳӣ)	Transliteration
153.	Фонема	Фонема (садо)	Phoneme
154.	Фонетика	Фонетика	Phonetics
155.	Фундаментальные проблемы	Масъалаҳои бунёдӣ	Fundamental issues
156.	Цепь Маркова	Занҷири Марков	Markov chain
157.	Цифровое изображение	Тасвири рақамӣ	Digital image
158.	Частота слова в документах	Басомади калимаҳо дар ҳуҷҷат	Document frequency (DF)
159.	Частота термина	Басомади истилоҳ (термин)	Term frequency (TF)
160.	Частотность слов	Басомади калима	Frequency of words
161.	Часть речи	Ҳиссаи нутқ	Part of speech (POS)
162.	Численные (вычислительные) методы	Усулҳои рақамӣ (ҳисоббарорӣ)	Numerical (accounting) methods
163.	Шаблон	Намуна (шакл)	Pattern
164.	Электронный словарь	Луғати электронӣ	Machine-readable dictionary
165.	Языковая модель	Тарҳи (амсилаи) забон	Language model
166.	Языковая технология	Технологияи забон	Language Engineering (LE) / Language Technology

ПРИЛОЖЕНИЕ 1. КОПИИ СВИДЕТЕЛЬСТВ О ГОСУДАРСТВЕННОЙ РЕГИСТРАЦИИ ИНФОРМАЦИОННЫХ РЕСУРСОВ И ИНТЕЛЛЕКТУАЛЬНЫХ ПРОДУКТОВ

1. Web-приложение “Автоматические системы обработки информации на таджикском языке – www.tajlingvo.tj"

ВАЗОРАТИ РУШДИ ИҚТИСОД ВА САВДОИ ЧУМҲУРИИ ТОҶИКИСТОН
МУАССИСАИ ДАВЛАТИИ «МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ»
МИНИСТЕРСТВО ЭКОНОМИЧЕСКОГО РАЗВИТИЯ И ТОРГОВЛИ РЕСПУБЛИКИ ТАДЖИКИСТАН
ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «НАЦИОНАЛЬНЫЙ ПАТЕНТНО- ИНФОРМАЦИОННЫЙ ЦЕНТР»

ШАҲОДАТНОМА
дар бораи бақайдгирии давлатии захираи иттилоотӣ

СВИДЕТЕЛЬСТВО
о государственной регистрации информационного ресурса

Номгуй Web-приложение «Автоматические системы обработки информации на таджикском языке»-www.tajlingvo.tj
Наименование (Web-замимаи «Низомхои худкори коркарди маълумот бо забони точикй»- www.tajlingvo.tj
Сарзамин
Страна Республика Таджикистан
Дорандаи
Владелец Худойбердиев Хуршед Атохонович, Усманов Зафар Джураевич
Тахиягар
Разработчик Худойбердиев Хуршед Атохонович
№ қайди давлатӣ
№ государственной регистрации № 4202200496
Ба Феҳристи давлатии захирахои иттилоотии Чумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан 28 апреля 2022 г.

Директор Исмоилзода М.Х

2. Web-приложение таджикский переводчик (tarjumon.tj)

ВАЗОРАТИ РУШДИ ИҚТИСОД ВА САВДОИ ЧУМҲУРИИ ТОҶИКИСТОН
МУАССИСАИ ДАВЛАТИИ «МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ»
МИНИСТЕРСТВО ЭКОНОМИЧЕСКОГО РАЗВИТИЯ И ТОРГОВЛИ РЕСПУБЛИКИ ТАДЖИКИСТАН
ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «НАЦИОНАЛЬНЫЙ ПАТЕНТНО- ИНФОРМАЦИОННЫЙ ЦЕНТР»

ШАҲОДАТНОМА
дар бораи бақайдгирии давлатии захираи иттилоотӣ

СВИДЕТЕЛЬСТВО
о государственной регистрации информационного ресурса

Номгуй Web-приложение таджикский переводчик tarjumon.tj
Наименование (Web-замимаи тарчумони точикй tarjumon.tj)
Сарзамин
Страна Республики Таджикистан
Дорандаи
Владелец Таджикский политехнический институт Таджикского технического университета имени М.С. Осими
Тахиягар
Разработчик Худойбердиев Х.А., Солиев О.М., Солиев П.А., Джолов Г.М., Назаров А.А.
№ қайди давлатӣ
№ государственной регистрации № 4202100482
Ба Феҳристи давлатии захирахои иттилоотии Чумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан 03 декабря 2021 г.

Директор Исмоилзода М.Х

3. Web-сайт “Электронный каталог кодексов Республики Таджикистан”

ВАЗОРАТИ РУШДИ ИҚТИСОД ВА САВДОИ ҶУМҲУРИИ ТОҶИКИСТОН
МУАССИСАИ ДАВЛАТИИ «МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ»
МИНИСТЕРСТВО ЭКОНОМИЧЕСКОГО РАЗВИТИЯ И ТОРГОВЛИ РЕСПУБЛИКИ ТАДЖИКИСТАН
ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «НАЦИОНАЛЬНЫЙ ПАТЕНТНО- ИНФОРМАЦИОННЫЙ ЦЕНТР»

ШАҲОДАТНОМА
дар бораи бақайдгирии давлатии захираи иттилоотӣ

СВИДЕТЕЛЬСТВО
о государственной регистрации информационного ресурса

Номгуй
Наименование Web-сайт «Электронный каталог кодексов Республики Таджикистан»
Сарзамин
Страна Республика Таджикистан
Доранда
Владелец Худойбердиев Хуршед А., Джалолов Икром А.
Тахиягар
Разработчик Худойбердиев Хуршед А., Джалолов Икром А.
№ қайди давлатӣ
№ государственной регистрации № 4202100470
Ба Феҳристи давлатии захираҳои иттилоотии Ҷумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан 25 февраля 2021 г.

Директор М. Исмоилзода

4. Автоматическая система TajSpell-2.0. для проверки орфографии таджикского языка в офисном пакете приложений MS Office 2010-2019

ВАЗОРАТИ РУШДИ ИҚТИСОД ВА САВДОИ ҶУМҲУРИИ ТОҶИКИСТОН
МУАССИСАИ ДАВЛАТИИ «МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ»
МИНИСТЕРСТВО ЭКОНОМИЧЕСКОГО РАЗВИТИЯ И ТОРГОВЛИ РЕСПУБЛИКИ ТАДЖИКИСТАН
ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «НАЦИОНАЛЬНЫЙ ПАТЕНТНО- ИНФОРМАЦИОННЫЙ ЦЕНТР»

ШАҲОДАТНОМА
дар бораи бақайдгирии давлатии захираи иттилоотӣ

СВИДЕТЕЛЬСТВО
о государственной регистрации информационного ресурса

Номгуй
Наименование Автоматическая системаTajSpell-2.0. для проверки орфографии таджикского языка в офисном пакете приложений MS Office 2010-2019
Сарзамин
Страна Республика Таджикистан
Доранда
Владелец Усмонов З.Дж., Солиев О.М., Худойбердиев Х.А., Довудов Г.М.
Тахиягар
Разработчик Усмонов З.Дж., Солиев О.М., Худойбердиев Х.А., Довудов Г.М.
№ қайди давлатӣ
№ государственной регистрации № 4202000456
Ба Феҳристи давлатии захираҳои иттилоотии Ҷумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан 30 июля 2020 г.

Директор М. Исмоилзода

5. Web-приложение Tajik-Russian-Parallel Corpus

ВАЗОРАТИ РУШДИ ИҚТИСОД ВА САВДОИ ҶУМҲУРИИ ТОҶИКИСТОН
МУАССИСАИ ДАВЛАТИИ «МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ»
МИНИСТЕРСТВО ЭКОНОМИЧЕСКОГО РАЗВИТИЯ И ТОРГОВЛИ РЕСПУБЛИКИ ТАДЖИКИСТАН
ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «НАЦИОНАЛЬНЫЙ ПАТЕНТНО- ИНФОРМАЦИОНЕНЫЙ ЦЕНТР»

ШАҲОДАТНОМА

дар бораи бақайдгирии давлатии захираи иттилоотӣ

СВИДЕТЕЛЬСТВО
о государственной регистрации информационного ресурса

Номгуӣ Наименование: Web-приложение Tajik-Russian-Parallel Corpus
Сарзамин Страна: Республика Таджикистан
Доранда Владелец: Худойбердиев Х.А., Солиев О. М., Довудов Г.М., Косимов А.А.
Таҳиягар Разработчик: Худойбердиев Х.А., Солиев О. М., Довудов Г.М., Косимов А.А.
№ қайди давлатӣ № государственной регистрации: № 4201900402
Ба Феҳристи давлатии захирахои иттилоотии Ҷумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан: 30 апреля 2019 г.

Директор: и.о. М. Исмоилзода

6. Web-приложение Tajik-English-Parallel Corpus

ВАЗОРАТИ РУШДИ ИҚТИСОД ВА САВДОИ ҶУМҲУРИИ ТОҶИКИСТОН
МУАССИСАИ ДАВЛАТИИ «МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ»
МИНИСТЕРСТВО ЭКОНОМИЧЕСКОГО РАЗВИТИЯ И ТОРГОВЛИ РЕСПУБЛИКИ ТАДЖИКИСТАН
ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «НАЦИОНАЛЬНЫЙ ПАТЕНТНО- ИНФОРМАЦИОННЫЙ ЦЕНТР»

ШАҲОДАТНОМА

дар бораи бақайдгирии давлатии захираи иттилоотӣ

СВИДЕТЕЛЬСТВО
о государственной регистрации информационного ресурса

Номгуӣ Наименование: Web-приложение Tajik-English-Parallel Corpus
Сарзамин Страна: Республика Таджикистан
Доранда Владелец: Солиев О. М., Худойбердиев Х.А., Назаров А.А., Солиев П.А.
Таҳиягар Разработчик: Солиев О. М., Худойбердиев Х.А., Назаров А.А., Солиев П.А.
№ қайди давлатӣ № государственной регистрации: № 4201900401
Ба Феҳристи давлатии захирахои иттилоотии Ҷумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан: 30 апреля 2019 г.

Директор: и.о. М. Исмоилзода

7. Компьютерный Диктор таджикского текста Computer Tajik Text Narrator

ВАЗОРАТИ РУШДИ ИҚТИСОД ВА САВДОИ ЧУМҲУРИИ ТОҶИКИСТОН
МУАССИСАИ ДАВЛАТИИ «МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ»
МИНИСТЕРСТВО ЭКОНОМИЧЕСКОГО РАЗВИТИЯ И ТОРГОВЛИ РЕСПУБЛИКИ ТАДЖИКИСТАН
ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «НАЦИОНАЛЬНЫЙ ПАТЕНТНО- ИНФОРМАЦИОННЫЙ ЦЕНТР»

ШАҲОДАТНОМА
дар бораи бақайдгирии давлатии захираи иттилоотӣ
СВИДЕТЕЛЬСТВО
о государственной регистрации информационного ресурса

Номгуй / Наименование: «Ровии Матни Тоҷикӣ» (Диктор Таджикского текста)
Сарзамин / Страна: Республика Таджикистан
Доранда / Владелец: Усманов З.Дж., Худойбердиев Х.А., Худойбердиев А.А.
Тахиягар / Разработчик: Усманов З.Дж., Худойбердиев Х.А., Худойбердиев А.А.
№ бақайдгирии давлатӣ / № государственной регистрации: № 4201800386
Ба Феҳристи давлатии захирахои иттилоотии Чумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан: 10 июля 2018 г.

Директор: Ч. Чумъахонзода

8. База данных “Единица измерений текстов произведений таджикских современных поэтов и писателей”

ВАЗОРАТИ РУШДИ ИҚТИСОД ВА САВДОИ ЧУМҲУРИИ ТОҶИКИСТОН
МУАССИСАИ ДАВЛАТИИ «МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ»
МИНИСТЕРСТВО ЭКОНОМИЧЕСКОГО РАЗВИТИЯ И ТОРГОВЛИ РЕСПУБЛИКИ ТАДЖИКИСТАН
ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «НАЦИОНАЛЬНЫЙ ПАТЕНТНО- ИНФОРМАЦИОННЫЙ ЦЕНТР»

ШАҲОДАТНОМА
дар бораи бақайдгирии давлатии захираи иттилоотӣ
СВИДЕТЕЛЬСТВО
о государственной регистрации информационного ресурса

Номгуй / Наименование: База данных «Единица измерений текстов произведений таджикских современных поэтов и писателей
Сарзамин / Страна: Республика Таджикистан
Доранда / Владелец: Усманов З.Дж., Худойбердиев Х.А., Косимов А.А.
Тахиягар / Разработчик: Усманов З.Дж., Худойбердиев Х.А., Косимов А.А.
№ бақайдгирии давлатӣ / № государственной регистрации: № 4201800381
Ба Феҳристи давлатии захирахои иттилоотии Чумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан: 16 мая 2018 г.

Директор: Ч. Чумъахонзода

9. База данных “Единица измерений текстов произведений таджикских классических поэтов и писателей”

ВАЗОРАТИ РУШДИ ИҚТИСОД ВА САВДОИ ҶУМҲУРИИ ТОҶИКИСТОН
МУАССИСАИ ДАВЛАТИИ «МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ»
МИНИСТЕРСТВО ЭКОНОМИЧЕСКОГО РАЗВИТИЯ И ТОРГОВЛИ РЕСПУБЛИКИ ТАДЖИКИСТАН
ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «НАЦИОНАЛЬНЫЙ ПАТЕНТНО- ИНФОРМАЦИОННЫЙ ЦЕНТР»

ШАҲОДАТНОМА
дар бораи бақайдгирии давлатии захираи иттилоотӣ

СВИДЕТЕЛЬСТВО
о государственной регистрации информационного ресурса

Номгуй / Наименование: База данных «Единица измерений текстов произведений таджикских классических поэтов и писателей
Сарзамин / Страна: Республика Таджикистан
Доранда / Владелец: Усманов З.Дж., Худойбердиев Х.А., Косимов А.А.
Тахиягар / Разработчик: Усманов З.Дж., Худойбердиев Х.А., Косимов А.А.
№ кайди давлатй / № государственной регистрации: № 4201800380
Ба Феҳристи давлатии захираҳои иттилоотии Ҷумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан: 16 мая 2018 г.

Директор: Ч. Чумъахонзода

10. Web-приложение проверки уникальности текста на таджикском языке Taj_Text_Plagiat

ВАЗОРАТИ РУШДИ ИҚТИСОД ВА САВДОИ ҶУМҲУРИИ ТОҶИКИСТОН
МУАССИСАИ ДАВЛАТИИ «МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ»
МИНИСТЕРСТВО ЭКОНОМИЧЕСКОГО РАЗВИТИЯ И ТОРГОВЛИ РЕСПУБЛИКИ ТАДЖИКИСТАН
ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «НАЦИОНАЛЬНЫЙ ПАТЕНТНО- ИНФОРМАЦИОННЫЙ ЦЕНТР»

ШАҲОДАТНОМА
дар бораи бақайдгирии давлатии захираи иттилоотӣ

СВИДЕТЕЛЬСТВО
о государственной регистрации информационного ресурса

Номгуй / Наименование: Web-приложение проверки уникальности текста на таджикском языке Taj_Text_Plagiat
Сарзамин / Страна: Республика Таджикистан
Доранда / Владелец: Усманов З.Дж., Солиев О.М., Худойбердиев Х.А., Солиев П.А, Косимов А.А
Тахиягар / Разработчик: Усманов З.Дж., Солиев О.М., Худойбердиев Х.А., Солиев П.А., Косимов А.А
№ кайди давлатй / № государственной регистрации: № 4201800378
Ба Феҳристи давлатии захираҳои иттилоотии Ҷумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан: 16 мая 2018 г.

Директор: Ч. Чумъахонзода

11.Computer tajik Ceossword - компьютерный таджикский кроссворд

Вазорати рушди иқтисод ва савдои Ҷумҳурии Тоҷикистон
Муассисаи давлатии Маркази миллии патенту иттилоот
Министерство экономического развития и торговли
Республики Таджикистан
Государственное учреждение
Национальный патентно - информационный центр

ШАҲОДАТНОМА

СВИДЕТЕЛЬСТВО

дар бораи бақайдгирии давлатии захираи иттилоотӣ
о государственной регистрации информационного ресурса

Номгуй / Наименование: Computer Tajik Crossword – компьютерный таджикский кроссворд
Сарзамин / Страна: Республика Таджикистан
Доранда / Владелец: Усманов Зафар Джураевич, Худойбердиев Хуршед Атохонович, Косимов Абдупаби Абдурауфович
Тахиягар / Разработчик: Усманов Зафар Джураевич, Худойбердиев Хуршед Атохонович, Косимов Абдупаби Абдурауфович
№ қайди давлатӣ / № государственной регистрации: № 4201200224

Ба Феҳристи Давлатии захираҳои иттилоотии Ҷумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан 14 марта 2012 г.

Директор КУРБАНОВ Дж.Дж.

12.Таджикский языковой пакет для проверки орфографии в Microsoft Office

Вазорати рушди иқтисод ва савдои Ҷумҳурии Тоҷикистон
Муассисаи давлатии Маркази миллии патенту иттилоот
Министерство экономического развития и торговли
Республики Таджикистан
Государственное учреждение
Национальный патентно - информационный центр

ШАҲОДАТНОМА

СВИДЕТЕЛЬСТВО

дар бораи бақайдгирии давлатии захираи иттилоотӣ
о государственной регистрации информационного ресурса

Номгуй / Наименование: Таджикский языковой пакет для проверки орфографии в Microsoft Office
Сарзамин / Страна: Республика Таджикистан
Доранда / Владелец: Усманов Зафар Джураевич, Солиев Одилходжа Махмудходжаевич, Худойбердиев Хуршед Атохонович, Довудов Гулшан Мирбахоевич
Тахиягар / Разработчик: Усманов Зафар Джураевич, Солиев Одилходжа Махмудходжаевич, Худойбердиев Хуршед Атохонович, Довудов Гулшан Мирбахоевич
№ қайди давлатӣ / № государственной регистрации: № 4201200235

Ба Феҳристи Давлатии захираҳои иттилоотии Ҷумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан 04 октября 2012 г.

Директор КУРБАНОВ Дж.Дж.

13.Компьютерное озвучивание таджикского текста Tajik Text-to-Speech

Вазорати рушди иқтисод ва савдои
Чумҳурии Тоҷикистон

МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ

ШАҲОДАТНОМАИ

бақайдгирии маҳсулоти зеҳнӣ

Номгӯй — Компьютерное озвучивание таджикского текста Tajik Text-to-Speech

Доранда хукук — Худойбердиев Х.А.

Муаллиф(он) — Усманов З.Дж., Худойбердиев Х.А.

Рақам ва санаи бақайдгирӣ 0411J 04.09.2007

Рақам ва санаи ворид гардидани аризнома

41/07 29.08.2007

Шаҳодатномаи мазкур ба эмонат гузоштани тавсифи маҳсулоти зеҳниро дар ММПИ-и Вазорати рушди иқтисод ва савдои Чумҳурии Тоҷикистон, дар ҳаҷми 1 CD-R тасдиқ мекунад.

ДИРЕКТОР Исҳоқмуллоев Ш.З.

14. Русско-таджикско-русский компьютерный словарь MuliGANJ

ВАЗОРАТИ РУШДИ ИҚТИСОД ВА САВДОИ
ҶУМҲУРИИ ТОҶИКИСТОН
МУАССИСАИ ДАВЛАТИИ
МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ

ШАҲОДАТНОМАИ
бақайдгирии маҳсулоти зеҳнӣ

Номгӯй Русско-таджикско-русский компьютерный словарь MultiGanj. Версия 1.0.0.

Муаллиф (он) Усманов З.Д., Холматова С.Д., Солиев О.М., Худойбердиев Х.А.

Рақам ва санаи бақайдгирӣ 077TJ 12.11.2008 с.
Рақам ва санаи ворид гардидани арзнома
77/08 27.10.2008 с.

Шаҳодатномаи мазкур ба амонат гузоштани тавсифи маҳсулоти зеҳниро дар МД ММПИ-и Вазорати рушди и ва савдои Ҷумҳурии Тоҷикистон дар ҳаҷми 2 саҳ. тасдиқ мекунад.

ДИРЕКТОР Нажмудинов Ш.З.

15.Таджикский текстовый редактор TajikWord (TW)

Вазорати рушди иқтисод ва савдои
Ҷумҳурии Тоҷикистон

МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ

ШАҲОДАТНОМАИ

бақайдгирии маҳсулоти зеҳнӣ

Номгӯй — Таджикский текстовый редактор Tajik Word (TW)

Дорандаи ҳуқуқ — Усманов Зафар Джураевич

Муаллиф(он) — Усманов Зафар Джураевич
Солиев Одилходжа Махмудходжаевич
Худойбердиев Хуршед Атохонович

Ракам ва санаи бақайдгирӣ 030TJ 05.07.2007
Ракам ва санаи ворид гардидани арзнома
30/07 18.06.2007

Шаҳодатномаи мазкур ба амонат гузоштани тавсифи маҳсулоти зеҳниро дар ММПИ-и Вазорати рушди иқтисод ва савдои Ҷумҳурии Тоҷикистон, дар ҳаҷми 1 CD-R тасдиқ мекунад.

ДИРЕКТОР Нажмудинов Ш.З.

ПРИЛОЖЕНИЕ 2. КОПИИ АКТОВ ВНЕДРЕНИЯ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ

«УТВЕРЖДАЮ»
Начальник Управления
по инвестициям и управлению
государственным имуществом
Согдийской области
Ахмедов Ш.Н.
«27» 10 2023 г.

АКТ

о внедрении основных результатов исследований по диссертационной работе Худойбердиева Х.А. на тему «Проектирование и реализация автоматических систем обработки информации на таджикском языке»

В рамках научных исследований в области компьютерной лингвистики таджикского языка достигнуты значительные результаты. Для их достижения выполнен ряд взаимосвязанных задач: проектирование и внедрение систем автоматической обработки информации на таджикском языке с целью формирования и развития электронных словарей, синтеза речи, автоматической проверки орфографии и машинного перевода текста. Полученные результаты соответствуют Национальной стратегии развития Республики Таджикистан на период до 2030 года, соблюдения Закона Республики Таджикистан о государственном языке и положений приведенных в «Концепции формирования электронного правительства в Республике Таджикистан», а именно в применение возможностей искусственного интеллекта и машинного обучения в рамках научно-исследовательских работ в сфере естественных наук.

В результате разработки автоматических систем предложен ряд методологических подходов к анализу, исследованию и автоматической обработке текстовой информации на таджикском языке:

- новые научно-технические положения, математические модели, методы и структуры данных, которые в совокупности формируют теоретические основы системного анализа и исследования текстовой информации;
- новые методы и собственные алгоритмы функционально-структурного и объектно-ориентированного проектирования автоматических систем обработки данных;
- новые модели, методы и программные средства для автоматического озвучивания, проверки орфографии и перевода текстовой информации, которые повышают эффективность применения ИКТ для решения задач в реальных практических проблемах информатизации таджикского языка.

В рамках предложенных методов, моделей и структуры данных спроектированы, теоретически обоснованы и практически апробированы собственные алгоритмы и программные средства автоматической обработки информации на базе трех основных задач: озвучивания, проверки орфографии и перевода. На основе предложенных методологии созданы программные средства

автоматического озвучивания текста Tajik-Text-to-Speech; автоматическая система проверки орфографии TajSpell в пакете программ Microsoft Office; программные модули в Интернет приложении tarjumon.tj. Все полученные результаты реализованы в одном программном комплексе TajLingvo позволяющем: значительно сократить сроки изучения таджикского языка как для пользователей Республики Таджикистан так и за рубежом; увеличить степень обоснованности принимаемых решений в компьютерной лингвистики и задач таджикского языка; обеспечить формирования и использования правильного контента на таджикском языке в сети Интернет.

На основе выше указанного следует, что диссертационное исследование Худойбердиева Х.А. имеет теоретическое и практическое значение по разработке математических и компьютерных моделей информационных систем обработки информации на таджикском языке. Научные результаты облегчают процесс делопроизводства на таджикском языке в организациях, а также ускоряют задачи применения прикладных задач в сферах науки и образования на основе современных компьютерных технологий.

Председатель комиссии: _____ Ш.Ахмедов

Члены комиссии:

_____ М.Раджабзода

_____ Р.Азиззода

_____ Х.Тохирова

_____ Л.Рахимова

НАЦИОНАЛЬНАЯ АКАДЕМИЯ НАУК ТАДЖИКИСТАН
ХУДЖАНДСКИЙ НАУЧНЫЙ ЦЕНТР

735714, г.Худжанд, Северо-восточная пром. зона тел: (83422) 5-78-16, 5-78-13
E-mail: markaziilmiikhujand@mail.ru

Исх. №17/а «14» 03 2023 г.

"Утверждаю"
Директор Худжандского
научного центра НАНТ
_____ Максуди А.Т.
"14" 03 2023 г.

АКТ

о внедрении основных результатов исследований по диссертационной работе Худойбердиева Х.А. на тему «Проектирование и реализация автоматических систем обработки информации на таджикском языке»

Настоящим актом подтверждается, что основные результаты, полученные Х.А. Худойбердиевым в диссертационной работе на тему: «Проектирование и реализация автоматических систем обработки информации на таджикском языке», для получения ученой степени доктора технических наук, научный консультант д.ф.м.н., профессор, Академик НАНТ З.Д.Усманов, включены в учебный план подготовки магистрантов отдела прикладной математики и информатики Худжандского научного центра НАНТ на 2022-2023 учебный год.

Диссертационное исследование имеет теоретическое и практическое значение при разработке математических и компьютерных моделей информационных систем автоматической обработки информации на таджикском языке, а также при решении актуальных задач, связанных с проблемами информатизации таджикского языка в Республике Таджикистан.

Большую практическую ценность исследованиям придаёт возможность его применения в качестве источника в учебном процессе для подготовки магистров направления информационных технологий, на основе которых читается специальные дисциплины "Автоматизированные информационные системы", "Математическая статистика и теория вероятностей ", "Экспертные системы", "Статистический анализ данных в Microsoft Excel". Модели, методы и алгоритмы разработанные Х.А. Худойбердиевым, полезны в применении методов математического и компьютерного моделирования в решении задач разработки информационных систем обработки информации на таджикском языке, в частности автоматизации делопроизводства и управления в учреждениях.

Заместитель директора ХНЦ НАНТ, к.ф.м.н. Эгамов М.Х.

ВАЗОРАТИ
МАОРИФ ВА ИЛМИ
ҶУМҲУРИИ ТОҶИКИСТОН
Муассисаи давлатии таълимии
"Донишгоҳи давлатии Хуҷанд
ба номи академик Бобоҷон Ғафуров"

МИНИСТЕРСТВО
ОБРАЗОВАНИЯ И НАУКИ
РЕСПУБЛИКИ ТАДЖИКИСТАН
Государственное образовательное
учреждение "Худжандский
государственный Университет имени
академика Бободжона Гафурова"

735700, г. Худжанд , проезд Мавлонбекова 1, тел.: (992-3422) 6-52-73, факс: (992-3422) 6-75-18,
e-mail: rector@hgu.tj

№ 01/2356 «24» 05 2023 г.

«УТВЕРЖДАЮ»
Ректор ГОУ ХГУ имени
академика Б. Гафурова, д.и.н.,
профессор Усмонзода А. И.

«24» 05 2023 г.

АКТ

о внедрении основных результатов исследований по диссертационной работе Худойбердиева Х.А. на тему «Проектирование и реализация автоматических систем обработки информации на таджикском языке»

Настоящим актом подтверждается, что основные результаты, полученные Х.А. Худойбердиевым в диссертационной работе «Проектирование и реализация автоматических систем обработки информации на таджикском языке» (научный консультант – д.ф.-м.н., профессор, академик НАНТ З.Д. Усманов), включены в учебный план специальности 1-31 03 03-02 – «Прикладная математика» в кафедре информатики и вычислительной математики и 1-40 01 01 – «Программное обеспечение информационных технологий» в кафедре программирования на 2023-2024 учебный год.

Диссертационное исследование имеет теоретические и практическое значение по разработке математических и компьютерных моделей информационных и методов автоматизации обработки информации на таджикском языке, а также решение актуальных задач в проблемах информатизации таджикского языка в Республике Таджикистан.

Большую практическую ценность исследований следует отнести к его применению в качестве источника познавательного процесса обучения студентов магистрантов направления 1-31 03 03-02 – «Прикладная математика» на основе которых им читается специальные дисциплины «Создание нейронных сетей», «Численные вариационные методы». Модели, методы и алгоритмы разработанные Х.А. Худойбердиевым, полезны в применении методов математического и компьютерного моделирования в решении задач разработки информационных систем обработки информации на таджикском языке, в частности автоматизации делопроизводства и управления в учреждениях.

Декан математического факультета
ГОУ ХГУ им. ак. Б. Гафурова,
к.ф.м.н., доцент Музафаров Д.З.

«ТАСДИҚ МЕКУНАМ»
Ректори ДДҲБСТ
д.и.и., профессор
Шарифзода М.М.
«05» 05 2023 с.

САНАД

оид ба истифодаи амалии натиҷаҳои асосии кори диссертатсионии н.и.ф.-м., дотсент Худойбердиева Х.А. дар мавзуи «Балоиҳагирӣ ва амалигардонии низомҳои ҳудкори коркарди маълумот бо забони тоҷикӣ (Проектирование и реализация автоматических систем обработки информации на таджикском языке)» барои дарёфти дараҷаи илмии доктори илмҳои техникӣ аз рӯи ихтисоси 05.13.11 – "Таъминоти математикӣ ва барномавии мошинҳои ҳисоббарор, муҷтамаъҳо ва шабакаҳои компютерӣ"

Дар Донишкадаи политехникии Донишгоҳи техникии Тоҷикистон ба номи академик М.Осимӣ, кафедраи «Барномарезӣ ва низомҳои иттилоотӣ» раванди корҳои тадқиқотӣ дар соҳаи лингвистикаи компютерӣ ва татбиқи натиҷаҳо барои рушди забони тоҷикӣ ва технологияҳои иттилоотӣ васеъ ба роҳ монда шудааст. Дар доираи тадқиқотҳои илмӣ аз тарафи Худойбердиев Х.А. бо роҳбарии доктори илмҳои физикаю математика, профессор, академики АМИТ Усмонов Зафар Ҷӯраевич якчанд лоиҳаҳои амалӣ коркард карда шудаанд.

Дар кафедраи забони тоҷикии Донишгоҳи давлатии ҳуқуқ бизнес ва сиёсати Тоҷикистон (ДДҲБСТ) лоиҳаҳои номбаршудаи самти лингвистикаи компютерӣ мавриди мухокима қарор дода шуданд.

Зимни шиносоӣ бо муҳтавои лоиҳаҳо ва муҳокимаи онҳо маълум гашт, ки мавзуи тадқиқот мубрам, ҷанбаи назариявию амалиашон боэътимод, навгониҳо вокеиву зарурӣ, доираву дараҷаи татбиқи онҳо фарогир ва боварибахш буда, барои рушди забони тоҷикӣ бо истифодаи имкониятҳои технологияҳои иттилоотӣ равона гардидаанд. Коркард ва ба амал баровардани натиҷаи таҳқиқот бо истифодаи васеи моделҳои математикӣ, дар сатҳи баланди барномарезӣ ба даст оварда шудаанд. Аз ҷумла:

1. Лоиҳаи "Тезауруси забони тоҷикӣ" аз лиҳози фарогирии таркиби луғавии забони тоҷикӣ баробари ҳамсанги "Фарҳанги забони тоҷикӣ" ва "Фарҳанги тафсирии забони тоҷикӣ" буданаш, аз нигоҳи корбасти электронӣ бартариҳои зиёди техникӣ(нишон додан ва хондани калимаҳои бисёргуна, ҳаволаҳо, сермаъноҳо, такрор, анвои маънофарӣ ва ғ.) дорад. Ҷунин вежагиҳо дар ҳар ду фарҳанги номбурда бо усули механикӣ сурат гирифтаанд.

2. Маълум аст, ки баробари ҷорӣ гардидани чопи компютерӣ масъалаи имлои забони тоҷикӣ ва ислоҳи мушкилоти он пеш омада, мутаассифона то ба имрӯз ҳалли худро нёфтааст. Аз ин ҷиҳат иқдоми пажӯҳишгарону барномарезони мавриди муҳокима кобили дастгирист, зеро лоиҳаи пешниҳоднамудаи онҳо "Комплекси барномаҳои TajSpell – тафтиши имлои забони тоҷикӣ" метавонад бархе аз ин мушкилотро бартараф намояд.

3. Масъалаи талаффуз яке аз мубрамтарин қазияҳои овошиноси(фонетика)-и тоҷик буда, то кунун бо истифодаи усули анъанавӣ-таърихӣ баррасӣ мегардад. Ҷорӣ гардидани лоиҳаи "Талаффузи компютерии матни тоҷикӣ" ба ҳалли компютерии талаффузи овозу калимаҳои забони тоҷикӣ ва мувофиқати он бо имлои забони тоҷикӣ кӯмаки назаррасе хоҳад расонд.

4. Лоиҳаи «Web-замимаи корпусҳои параллелии забонҳои тоҷикӣ ва англисӣ" сухани нисбатан навест дар забоншиносии анъавии тоҷик. Зеро он дар замони шуравӣ аз ҷумлаи масоили дувумдараҷаи забоншиносии миллӣ маҳсуб меёфт. Ба шарофати истиқлолияти давлатии Ҷумҳурии Тоҷикистон масъалаи мазкур мубрамияти аввалиндараҷа касб кард. Аз ин лиҳоз Web-замимаи пайкараҳои мувозии забонҳои тоҷикӣ ва англисие, ки омода карда шудаанд, метавонад ба ҳайси дастури амалӣ хизмат намояд.

5. "Web-замимаи корпусҳои параллелии забонҳои тоҷикӣ ва русӣ" лоиҳаест, ки татбиқи он бояд кайҳо сурат мегирифт. Дар бисёр давлатҳои пасошӯравӣ масъалаи мазкур хеле барвақт ҳалли худро ёфтаанд, вале, мутаассифона, ин масъалаи доғи забоншиносии муқоисавӣ - типологӣ дар Тоҷикистон, бо вуҷуди заминаи бою ғанӣ доштанаш(нашри шумораи зиёди луғатҳои тоҷикӣ-русӣ ва русӣ-тоҷикӣ дар замони шуравӣ ва давраи истиқлолият) то ҳол интизори баррасӣ мондааст. Аз ин лиҳоз татбиқи чунин лоиҳаҳоро дер монондан савоб нахоҳад буд.

6. "Тарҷумони автоматии матни забони тоҷикӣ"-ро яке аз лоиҳаҳои зарурӣ ҳисобидан равост. Зеро он метавонад холигоҳи зиёди барномаи мавҷударо пур намояд.

7. "Луғатҳои электронӣ"-и пешниҳоднамудаи мураттибони мазкур мукаммал, аз лиҳози истифода сода ба корбурди интернетӣ хеле мувофиқ буда, барои ҳарчи бехтар дастрас гардонидани фарҳангномаҳои электронӣ, ки алҳол барои бойгардонии он ниёзи зиёд мавҷуд аст, саҳми вокеӣ дошта метавонад.

Бо дарназардошти гуфтаҳои боло кафедраи забони тоҷикии ДДҲБСТ саҳми З.Ҷ.Усмонов ва Х.А. Худойбердиевро дар пешбурди амалияи беҳдошти лингвистикаи компютерӣ ва амалӣ махсус қайд намуда, итминон пайдо карданд, ки татбиқи пешниҳодоти эшон омили рушди соҳаи технологияҳои иттилоотӣ ва забони тоҷикӣ шуда метавонанд. Амалӣ гардидани лоиҳаҳои мавриди тақриз иқдоми мушаххас дар татбиқи бандҳои Қонуни Ҷумҳурии Тоҷикистон "Дар бораи забони давлатии Ҷумҳурии Тоҷикистон"(5-уми октябри соли 2009) ва Қарори Ҳукумати Ҷумҳурии Тоҷикистон аз 28-ноябри соли 2020, № 47 "Дар бораи Барномаи рушди забони давлатӣ барои солҳои 2020-2030" маҳсуб меёбад.

Профессори кафедраи забони тоҷикии
Донишгоҳи давлатии ҳуқуқ, бизнес
ва сиёсати Тоҷикистон, доктори илми
филология, академики Академияи
табиатшиносии Федератсияи Россия Шокириён Т. С.

«УТВЕРЖДАЮ»
Заместитель директора
по учебной части
политехнического института
Таджикского технического
университета имени
академика М.С. Осими
в городе Худжанде
_____ Акрамов А.
«21» 04 2023 г.

АКТ

о внедрении основных результатов исследований по диссертационной работе Худойбердиева Х.А. на тему «Проектирование и реализация автоматических систем обработки информации на таджикском языке»

Комиссия в составе: председателя – к.э.н. Ахмедов У.Х. - заместителя директора по науке и инновации, членов комиссии: к.ф.-м.н., доцент Максудов Х.Т., к.т.н. Довудов Г.М., к.т.н. Левандовский Б.И., к.п.н. Усмановой М.Р. констатирует, что диссертационная работа Худойбердиева Хуршеда Атохоновича на тему «Проектирование и реализация автоматических систем обработки информации на таджикском языке» на соискание ученой степени доктора технических наук, (научный консультант – д.ф.м.н., профессор, Академик НАНТ З.Д. Усманов) рассматривает вопросы, по теоретическому и прикладному направлению углубленного изучения основ компьютерной лингвистики таджикского языка.

Полученные результаты диссертационных исследований:

- математические модели, методы обработки данных на таджикском языке для определения как теоретических, так и практических основ системного анализа текстовой информации;
- компьютерные модели, алгоритмы и программные модули для объектно-ориентированного проектирования автоматических систем обработки данных на таджикском языке;

используются в учебном процессе начиная с 2021-2022 учебного года до сегодняшней времени для проведения спецкурсов для магистрантов специальностей 1-400101 – Программное обеспечение информационных технологий, 1-400301 – Искусственный интеллект и для докторантов PhD по специальностям 6D070400 – Вычислительная техника и программное обеспечение, 6D070300 – Информационные системы в ПИТТУ имени академика М.С. Осими в городе Худжанде.

Председатель комиссии: Ахмедов У.Х.

Члены комиссии:

Максудов Х.Т. Довудов Г.М.

Левандовский Б.И. Усмонова М.Р.

ҶСП «Душанбе Сити Бонк»
Ҷумҳурии Тоҷикистон, ш. Душанбе
кӯчаи Соҳили 5
РМА 510022404
+992 (41) 800 88 55
info1@dc.tj | www.dc.tj

CJSC Dushanbe City Bank
5 Sohili street
Republic of Tajikistan, Dushanbe
TIN 510022404
+992 (41) 800 88 55
info1@dc.tj | www.dc.tj

Председатель ЗАО ДС банк
Исполов Хуршед Назарабдолович

АКТ

по практическому использованию программного комплекса TajSpell 2.0 проверки правописания на таджикском языке в MS Office

TajSpell 2.0 комплекс программных модулей для автоматической проверки правописания текста на таджикском языке, расстановка переносов и компьютерного таджикского тезауруса в пакете программ Microsoft Office. Программный комплекс разработан в рамках научно-исследовательских работ к.ф.м.н., доцента Худойбердиева Х.А. на тему «Проектирование и реализация автоматических систем обработки информации на таджикском языке» для получения ученой степени доктора технических наук.

Программные модули TajSpell 2.0 являются дополнительным расширением для пакета программ Microsoft Office версий 97/2000/XP/2003/2007/2010/2013/2016/2019 и обеспечивают проверку правописания во время набора и редактирования текстовых документов на таджикском языке.

Начиная с 2022 года в компьютерах нашей компании TajSpell 2.0 установлены в качестве подготовки документов на государственном языке, который обеспечивает эффективность документооборота и гарантирует высокое качество подготовки больших объемов текстов на таджикском языке.

Начальник отдела по разработке ПО Солиев О.М.