Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Донецкий национальный технический университет»

На правах рукописи

Большакова Светлана Анатольевна

УДК 004.912

СОВЕРШЕНСТВОВАНИЕ МЕТОДОВ КОМПЬЮТЕРНОЙ ОБРАБОТКИ ТЕКСТОВОЙ ИНФОРМАЦИИ В АСПЕКТЕ ЗАДАЧ, СВЯЗАННЫХ С ОМОНИМИЕЙ И СИНОНИМИЕЙ

Специальность 2.3.1. Системный анализ, управление и обработка информации, статистика (технические науки)

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Донецк – 2025

Работа выполнена в федеральном государственном бюджетном научном учреждении «Институт проблем искусственного интеллекта» Министерства науки и высшего образования Российской Федерации, г. Донецк

Научный руководитель: доктор физико-математических наук, профессор
ШЕЛЕПОВ Владислав Юрьевич
Федеральное государственное бюджетное научное учреждение «Институт проблем искусственного интеллекта» (г. Донецк), главный научный сотрудник

Официальные оппоненты: доктор технических наук, профессор
ВАРЛАМОВ Олег Олегович
Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (г. Москва), профессор кафедры «Системы обработки информации и управления»

доктор технических наук, доцент
ШИШАЕВ Максим Геннадьевич
Институт информатики и математического моделирования Обособленное подразделение Федерального государственного бюджетного учреждения науки Федерального исследовательского центра «Кольский научный центр Российской академии наук» (г. Апатиты), главный научный сотрудник

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт Высшей нервной деятельности и нейрофизиологии РАН (г. Москва)

Защита состоится «15» января 2026 года в 12:00 часов на заседании диссертационного совета 24.2.491.03 при ФГБОУ ВО «Донецкий национальный технический университет» по адресу: 283001, г. Донецк, ул. Артема, 58, I учебный корпус, аудитория 1.203.

Тел. факс: +7 856 301-07-69, e-mail: donntu.info@mail.ru

С диссертацией можно ознакомиться в библиотеке ФГБОУ ВО «ДонНТУ» по адресу: 283001, г. Донецк, ул. Артема, 58, II учебный корпус, а также на сайте: http://donntu.ru

Автореферат разослан «___» _____ 2025 г.

И.о. ученого секретаря
диссертационного совета 24.2.491.03
доктор технических наук, профессор А.О. Новиков

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность исследования. Одной из наиболее сложных задач при автоматической обработке естественно-языковых (ЕЯ) текстов (Natural Language Processsing – NLP) является неоднозначность его единиц, проявляющаяся на всех уровнях, что выражается в явлениях омонимии и синонимии. Разрешение многозначности элементов естественного языка является одной из фундаментальных проблем компьютерной обработки текста**.** Снятие омонимии является необходимым и важным этапом для качественного машинного анализа текстов и, в конечном итоге, понимания и извлечения знаний из них. Для русского языка эта проблема особенно актуальна ввиду наличия очень большого числа омонимичных словоформ. Для слов русского языка примерно в половине случаев имеет место какая-либо форма омонимии, и набор грамматических характеристик оказывается неоднозначным.

Одной из актуальных и социально-значимых NLP-задач является преобразование сложных текстов в тексты, использующие более простой и понятный язык. Такое преобразование текста называется его симплификацией, адаптацией или упрощением. Она может достигаться путем изменения сложных языковых конструкций, а также путем замены слов и словосочетаний более простыми (лексическая адаптация). Эта проблема особенно актуальна для людей, знание языка которых не позволяет в достаточной степени понять сложную текстовую информацию, в частности, для иностранцев, изучающих язык, для людей с первыми симптомами когнитивных нарушений, связанных с возрастом или травмами головного мозга, для детей с задержками речевого развития. Инструменты автоматизированной адаптации могут применяться при разработке приложений для автоматической обработки языка, в том числе для поиска, классификации документов, автореферирования, машинного перевода.

Эффективным способом упрощения ЕЯ-текстов является использование синонимии, поскольку один и тот же смысл может быть выражен различными синтаксическими конструкциями и словами, среди которых можно найти наиболее простую форму выражения.

В русском языке синонимы зачастую обладают различными морфологическими характеристиками, что создает трудности при автоматической замене, связанные с соблюдением правил синтаксиса в адаптированном тексте.

В связи с вышесказанным, совершенствование методов и программных средств снятия омонимии в русскоязычных текстах, и их адаптация с помощью использования более простых и распространенных синонимов, сохраняя правильный синтаксис и смысл текста после упрощения, является актуальной задачей отраслевого значения.

Связь работы с научными программами, планами, темами. Результаты работы внедрены в ФГБНУ «Институт проблем искусственного интеллекта» при выполнении фундаментальных научно-исследовательских

работ: «Исследование и разработка методов снятия омонимии в естественно-языковых текстах внутри парадигмы русского слова» (№Г/Р 0121D000017), «Исследование и разработка методов семантического анализа и интерпретации потоков данных интеллектуальными системами» (№Г/Р 0118D000003), «Исследование и разработка методов обработки данных и естественно-языковых текстов с применением онтологий» (№ гос. учета в ЕГИСУ НИОКТР 123092600030-4).

Степень разработанности темы исследования. В настоящее время создаются NLP-системы анализа текстов романо-германской группы, с помощью которых можно проводить автоматизированную адаптацию для различных целей. Проблеме адаптации медицинских текстов посвящены работы G. Grigonyte, I. Spasic, упрощению текстов из Википедии посвятили свои работы W. Coster и K. Woodsend, вклад в развитие методов упрощения текстов для детей или людей с дислексией внесли J. De Belder, L. Rello, адаптацией текстов для изучающих иностранный язык занимались S. E. Petersen, M. Ostendorf. Для русского языка проблема автоматизированной адаптации является недостаточно исследованной в сравнении с языками романо-германской группы. Разработкой приложений для адаптации русскоязычных текстов занимались В.Г. Сибирцева и Н.В. Карпов. Вместе с тем к настоящему времени проблема лексического упрощения остается открытой для разработки новых методов.

Основой для лексического упрощения текста традиционно выступают исследования в области теории синонимии. Лингвистической основой данного исследования является словарь синонимов З.Е. Александровой, а также частотные словари О.Н. Ляшевской и С.А. Шарова.

Цель диссертационного исследования – повышение эффективности обработки и анализа текстовой информации на основе развития методов компьютерной обработки русскоязычных текстов в контексте задач снятия омонимии и применения способов лексической адаптации путем синонимических замен.

Для достижения поставленной цели сформулированы и решены следующие **задачи**:

–проведен аналитический обзор технологий и методов автоматической обработки текстовой информации;

–реализован алгоритмы определения морфологических параметров словоформ и лемматизации;

–разработаны алгоритмы разрешения частеречной омонимии на основе базы продукционных правил;

–разработан метод упрощения текста путем замены отдельных слов и словосочетаний более простым и более употребительным синонимом с помощью базы правил и меток в базе синонимов;

–сформированы тестовые корпуса: размеченная база синонимов, словарь отглагольных существительных для построения элементов плана текста;

–разработан метод автоматического разбиения текста на абзацы как

семантически однородные фрагменты;

–разработан метод автоматического построения элемента плана текста;

–выполнена программная реализация предложенных методов и алгоритмов в единой системе обработки и анализа текстовой информации и проведена оценка их эффективности.

Объектом исследования являются русскоязычные тексты.

Предмет исследования – методы автоматического снятия омонимии и автоматической адаптации текстов на русском языке.

Методология и методы исследования. Исследование базируется на методах компьютерной лингвистики и методах NLP для проведения морфологического и синтаксического анализа; методах технологий извлечения знаний для построения базы продукционных правил, позволяющий снимать омонимию и сохранять правильный синтаксис; методах объектно-ориентированного программирования для программной реализации системы адаптации русскоязычных текстов.

Научная новизна полученных результатов заключается в следующем.

1) Получили дальнейшее развитие методы автоматического разрешения омонимии на основе гибридного подхода, использующего как декларативные знания в виде словарей, так и базу продукционных правил, что позволило снять частеречную омонимию предикативов и предикативных словосочетаний, деепричастий, групп наречие-существительное с точностью 99,3%.

2) Впервые предложен метод упрощения текста, использующий специально размеченную базу синонимов и набор правил соблюдения синтаксиса, что позволяет осуществлять лексическую замену слов и словосочетаний с соблюдением правильного синтаксиса и сохранением семантики текста с точностью выше 96%.

3) Получили дальнейшее развитие методы автоматического разбиения текста на абзацы как семантически однородные фрагменты за счет введенной величины, учитывающей частоту встречаемости слова и длину отрезка текста, где оно встречается.

Теоретическая значимость научных результатов, полученных в ходе диссертационного исследования, заключается в развитии методов компьютерной обработки русскоязычных текстов за счет создания лингвистических баз знаний, направленных на снятие омонимии и лексическую адаптацию.

Практическое значение работы. Предложенные методы снятия омонимии и лексической адаптации в русскоязычных текстов могут быть применены при разработке широкого круга систем автоматизированного упрощения текстов на русском языке, используемых для подготовки текстов для детей или взрослых, изучающих русский язык как иностранный, для людей, страдающих различными нарушениями восприятия, препятствующими пониманию лексически сложных текстов (афазия, нарушения слуха и т.д).

Разработанные методы и алгоритмы, а также размеченные текстовые корпуса и базы синонимов могут быть использованы как компоненты в NLP-системах различного назначения: машинного перевода, информационного поиска, автоматического реферирования, классификации текстов и пр.

Методы компьютерной обработки текстовой информации нашли применение в работе федерального государственного бюджетного научного учреждения "Республиканский академический научно-исследовательский и проектно-конструкторский институт горной геологии, геомеханики, геофизики и маркшейдерского дела" (ФГБНУ "РАНИМИ") при обработке массивов текстовой информации, что подтверждается справкой о внедрении №04.02-07/34/1 от 05.02.2025 г.).

Результаты и выводы работы нашли применение при выполнении фундаментальных научно-исследовательских работ в ФГБНУ «Институт проблем искусственного интеллекта», что подтверждается справкой о внедрении №173/1/01-01 от 01.07.2025 г.).

Положения, выносимые на защиту.

1) Установлено, что использование декларативных знаний в виде словарей совместно с базой продукционных правил снятия частеречной омонимии предикативов и предикативных словосочетаний, деепричастий, а также групп наречие-существительное, обеспечивает существенное повышение точности разрешения омонимии.

2) Показано, что применение специально размеченного текстового корпуса в виде базы синонимов, а также базы продукционных правил позволяет осуществлять синонимические замены слов и словосочетаний с сохранением семантики текста и правильного русского синтаксиса с точностью выше 96%.

Соответствие паспорту специальности. По направлению исследований, содержанию научных положений и выводов, существу полученных результатов диссертационная работа соответствует паспорту специальности 2.3.1. Системный анализ, управление и обработка информации, статистика (технические науки) по областям исследований: п.3 «Разработка критериев и моделей описания и оценки эффективности решения задач системного анализа, оптимизации, управления, принятия решений, обработки информации и искусственного интеллекта»; п. 4. «Разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений, обработки информации и искусственного интеллекта»; п. 5. «Разработка специального математического и алгоритмического обеспечения систем анализа, оптимизации, управления, принятия решений, обработки информации и искусственного интеллекта».

Обоснованность и достоверность научных положений обеспечивается полнотой теоретических и практических исследований, положительной оценкой на научных конференциях и семинарах, выполненными публикациями.

Апробация результатов работы. Основные научные положения и результаты диссертационной работы доложены и обсуждены на семинарах и конференциях: VII Международная научно-техническая конференция «Современные информационные технологии в образовании и научных исследованиях» (Донецк, 23 ноября 2021 г.), международный научный круглый стол «Искусственный интеллект: теоретические аспекты и практическое применение» (г. Донецк, 2020-2024), а также II Всероссийская школа Национального центра физики и математики для студентов, аспирантов, молодых ученых и специалистов по искусственному интеллекту и большим данным в технических, промышленных, природных и социальных систем (г. Саров, 25-29 ноября 2024 г.).

Личный вклад автора. Основные научные результаты диссертации, которые заключаются в разработке методов автоматического снятия омонимии и автоматической адаптации текстов на русском языке, а также разработке программных средств, входящих в состав системы снятия омонимии и адаптации текста получены соискателем лично. Постановка задач исследования, формулирование основных положений работы, разработка структуры и содержания работы выполнены совместно с научным руководителем.

Публикации по теме диссертации. Содержание диссертационного исследования изложено в 17 публикациях, из которых 2 размещены в изданиях, входящих в Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук, утверждённый ВАК РФ (к-1 и к-2), 7 – в других рецензируемых научных изданиях, а также получено 1 свидетельство о регистрации программы для ЭВМ.

Структура и объем работы. Диссертационная работа содержит 172 страниц машинописного текста и состоит из введения, четырех разделов, заключения, списка литературы из 121 источника на 15 страницах и 7 приложений на 29 страницах. Основной текст, изложенный на 141 страницах, иллюстрируется 9 рисунками и содержит 19 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертационной работы, проведен краткий обзор и анализ научной литературы по теме диссертации, сформулирована цель и задачи исследований, представлена научная новизна, теоретическая и практическая значимость работы.

В первом разделе приведен обзор технологий и методов компьютерной обработки текстовой информации: анализ современных представлений о синонимии и омонимии, описаны методы и способы работы с ними при автоматизированной обработке текстовой информации, в том числе и автоматизированного упрощения текста.

Во втором разделе приведено описание основных структур сформированного словаря словоформ для лемматизации, его представление в виде префиксного дерева и алгоритм индексирования строк данного словаря.

При формировании морфологического словаря за основу взят словарь русских словоформ М. А. Хаген «Полная парадигма. Морфология». В разделе описана блок-схема алгоритма функционирования метода снятия омонимии (рисунок 1). Входными данными являются строка *T*, содержащая текст (предложения, включая правильные знаки препинания). Производится сегментация на предложения и токенизация текста. Исходный текст преобразуется в цепочку токенов (массив *T*), для каждого из которых производится поиск грамматических параметров и леммы.

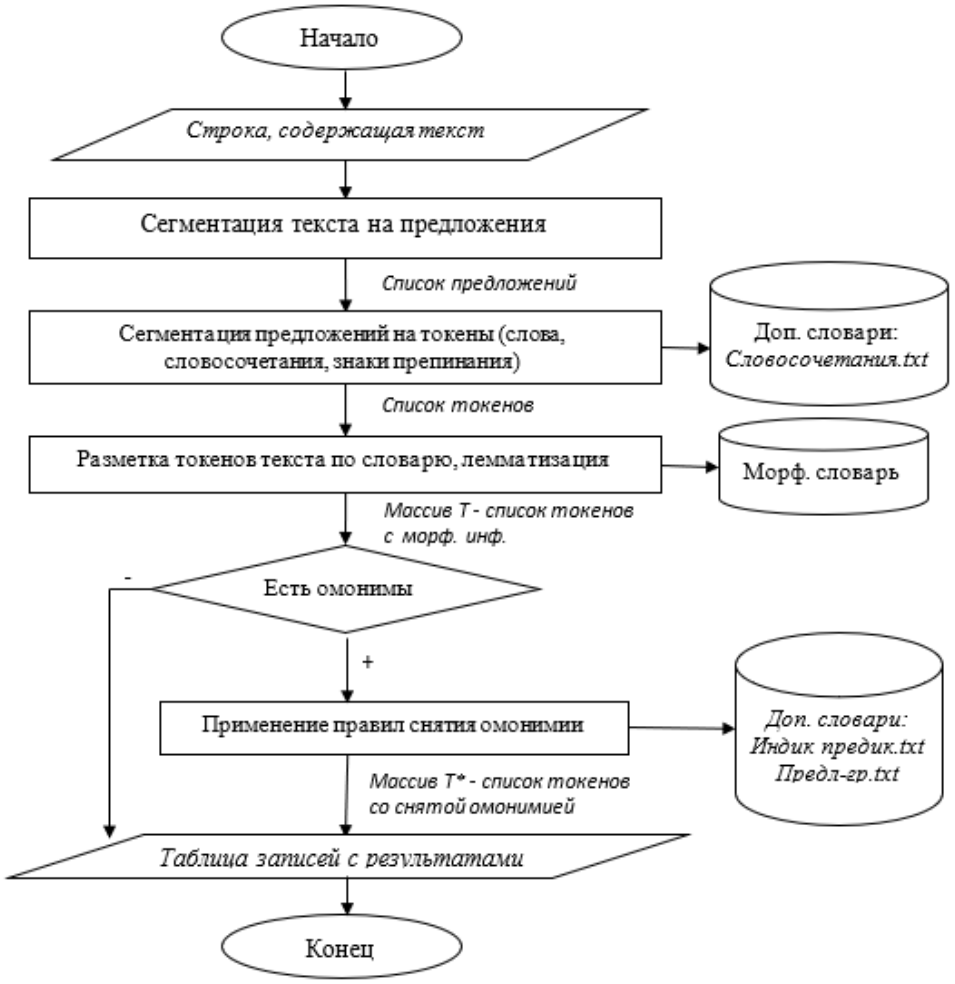

Рисунок 1 – Блок-схема алгоритма функционирования метода снятия омонимии

Третий раздел содержит описание методов и алгоритмов снятия омонимии в русскоязычных текстах, полученных автором и их программной реализации. Описана разработанная база продукционных правил для снятия частеречной омонимии и блок-схема алгоритма применения правил снятия омонимии (рисунок 2) для снятия частеречной омонимии предикативов и предикативных словосочетаний, деепричастий, групп наречие-существительное.

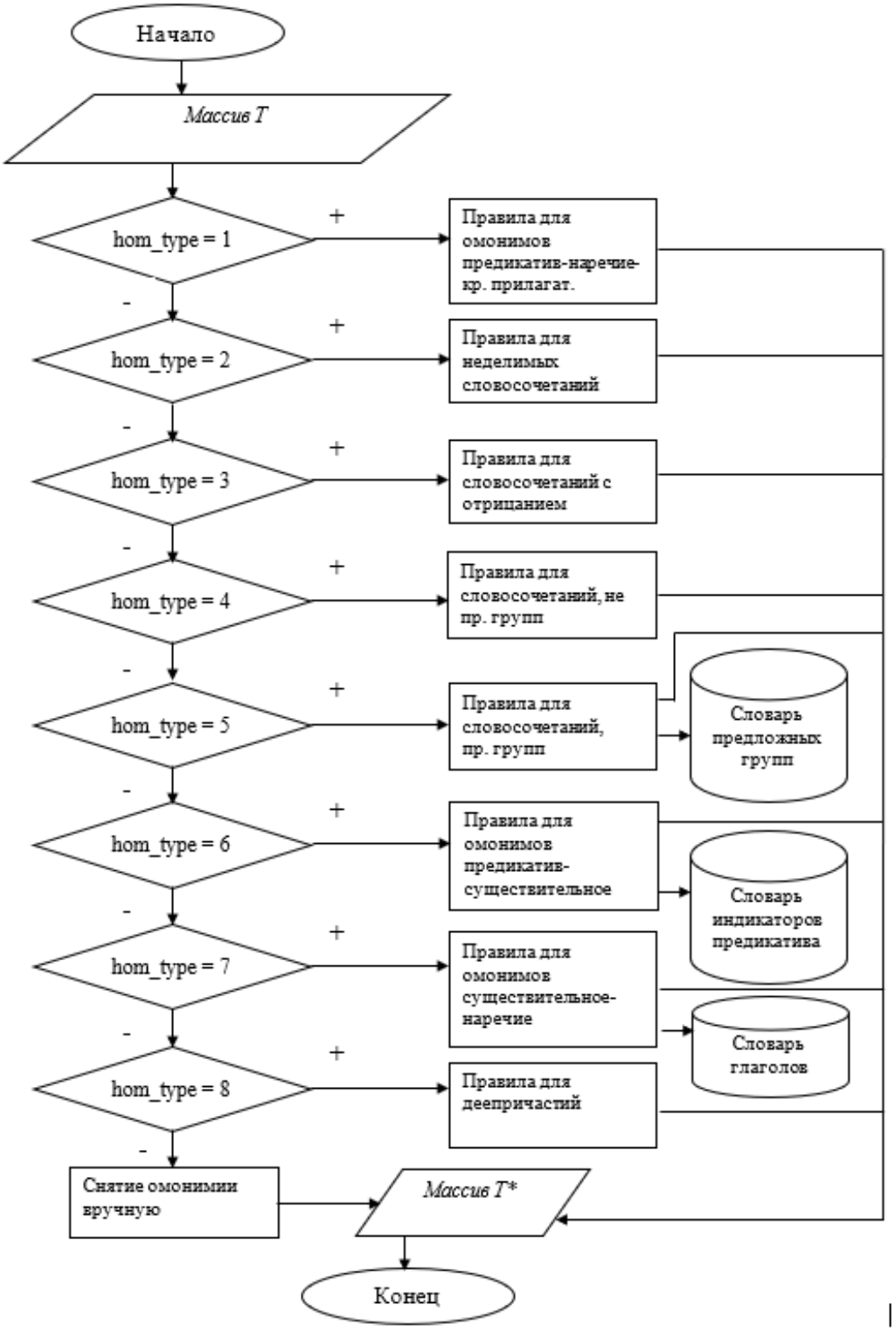

Рисунок 2 – Блок-схема алгоритма применения правил снятия омонимии

В подразделе 3.1 описываются основные этапы работы системы снятия омонимии. Входными данными является массив *T*, содержащий список токенов с морфологической информацией и леммой. С помощью правил снятия неоднозначности (описаны в п. 3.2-3.6) из нескольких вариантов разбора слова выбирается один. Выбранный вариант разбора помечаются специальной пометкой «!» (массив *T**).

Разработанный метод анализирует правильные русские предложения с правильно поставленными знаками препинания. При решении вопроса об определении омонима достаточно ограничиваться содержащим его отрезком предложения между двумя соседними знаками препинания. Опишем это формально. Введем множество знаков препинания *P*:

P {".",",",":",";","!","?","\"","-","(",")","[","]","…" }.

Контекстный фрагмент для снятия омонимии определим как отрезок между двумя элементами множества *P*. Пусть *T[i]* не является знаком препинания, тогда находим ближайшее $T[i_{start}] \in P$ такое что $i_{start} >= 0$ *и* $i_{start} < i$ и ближайшее $T[i_{end}] \in P$, $i_{end} > i$.

Подразделы 3.2 – 3.3 посвящены омонимии предикативов и предикативных словосочетаний. Предикатив – часть речи, связанная с функцией сказуемого в предложении. Наиболее частым в предложении является глагольное сказуемое. Сказуемое может также выражаться другими частями речи: существительным, прилагательным (традиционная реализация). Однако в последний перечень неестественно включать наречие (выражает дополнительную характеристику действия или качества). Поэтому для описания сказуемого в предложении типа «Мне холодно» была введена часть речи «категория состояния» или «предикатив».

В подразделе 3.2 разбирается наиболее часто встречающийся случай омонимии предикатив-наречие-краткое прилагательное, когда омоним выражен одним словом и является единственным кандидатом на предикатив. Сформулированы правила для снятия омонимии. Основная идея данных правил: если сказуемое выражено традиционным способом, то омоним не является предикативом.

В подразделе 3.3 впервые обсуждается омонимия словосочетаний, которые могут быть предикативами, составляющих одну словарную статью. Они могут также выступать в роли наречий, местоимений, частиц и вводных словосочетаний. Во всех этих случаях словосочетание интерпретируется как одна словарная единица. С другой стороны, оно может в ряде случаев требовать разбиения на отдельные слова.

Рассмотрим словосочетание «по дороге». Семантически здесь возможны три ситуации.

1. Речь идет о движении: «двигаться по дороге», «идти по дороге», «ехать по дороге», «скакать по дороге» и так далее. В этом случае словосочетание состоит из двух отдельных слов – предлога и существительного в предложном падеже, которые вместе образуют предложную группу.

2. Речь идет о чем-то, что делается одновременно с основным движением, попутно: «По дороге зайдем в магазин». В этом случае «по дороге» – наречие и трактуется как одна единица словаря.

3. Словосочетание означает совпадение целей, интересов и так далее: «Мне с ним по дороге». «Нам по дороге с этой партией». В этом случае «по дороге» – предикатив и снова трактуется как одна единица словаря.

Существуют словосочетания-омонимы, которые не являются предложными группами и при этом в одних случаях их нужно трактовать как единое целое (одна словарная статья), а в других разбивать на отдельные слова. Пример: «куда там». Наконец, таковыми являются многие словосочетания с отрицанием. Пример: «не грех».

Затем приведены результаты о снятии омонимии большинства словосочетаний, которые могут быть лишь предикативами и предложными группами и не содержат отрицательных частиц *НЕ* и *НИ.*

В подразделах 3.4 – 3.6 изложены результаты об омонимии предикатив-существительное, омонимии деепричастий и омонимии наречие-существительное.

В разделе 3 приведены численные исследования точности разработанных методов с помощью их программной реализации Предложенный метод дает высокую точность морфологической разметки, что объясняется обширной базой правил для конкретных словосочетаний, охватывающих практически все возможные варианты снятия омонимии.

Проведен сравнительный анализ эффективности предложенного метода снятия омонимии с известными аналогами. Выбраны два морфоанализатора русского языка, являющихся бесспорными лидерами: MyStem и Pymorphy2. Данные морфоанализаторы не распознают предикативы и предикативные словосочетания, а омонимию наречие-существительное и деепричастий снимают лишь частично (точность снятия омонимии 22% и 24%). Наилучшие показатели точности имеет авторский метод снятия омонимии за счет анализа контекста омонимов и применения сформулированных правил снятия омонимии (точность снятия омонимии 99,3%).

В четвертом разделе в рамках разработки единой системы обработки и анализа текстовой информации предложены алгоритмы семантической обработки текста: метод лексической адаптации текста с помощью синонимических замен; метод автоматического разбиения сплошного текста на абзацы как семантически однородные фрагменты; автоматического создания элементов плана за счет использования словаря отглагольных существительных. Приведены результаты численных исследований эффективности предложенных методов и алгоритмов.

В подразделе 4.1 представлена общая схема работы системы обработки и анализа текстовой информации (рисунок 3). Разработанная система состоит из двух блоков: блока морфологической обработки и семантической обработки текста. Алгоритмы первого блока описаны в разделе 3, в результате лемматизации и снятия омонимии на вход блока семантической обработки поступает массив пар $\left\{\left\{\langle \text{лемма}_j^i, \text{МИ}_j^i \rangle\right\}_{j=1}^{N\_j}\right\}_{i=1}^{M}$ для каждого *j*-го токена *i*-го предложения текста, *N_j* – количество токенов в предложении *j*, *M* – количество предложений текста.

В подразделе 4.2 описана работа с синонимией для автоматического упрощения (адаптации) русскоязычных текстов. Представлен метод обратного применения синонимических рядов с восстановлением правильного русского

синтаксиса. Для этого проработан объемный лингвистический материал и сформирована размеченная база синонимов, занимающая более двухсот страниц приложений к диссертации.

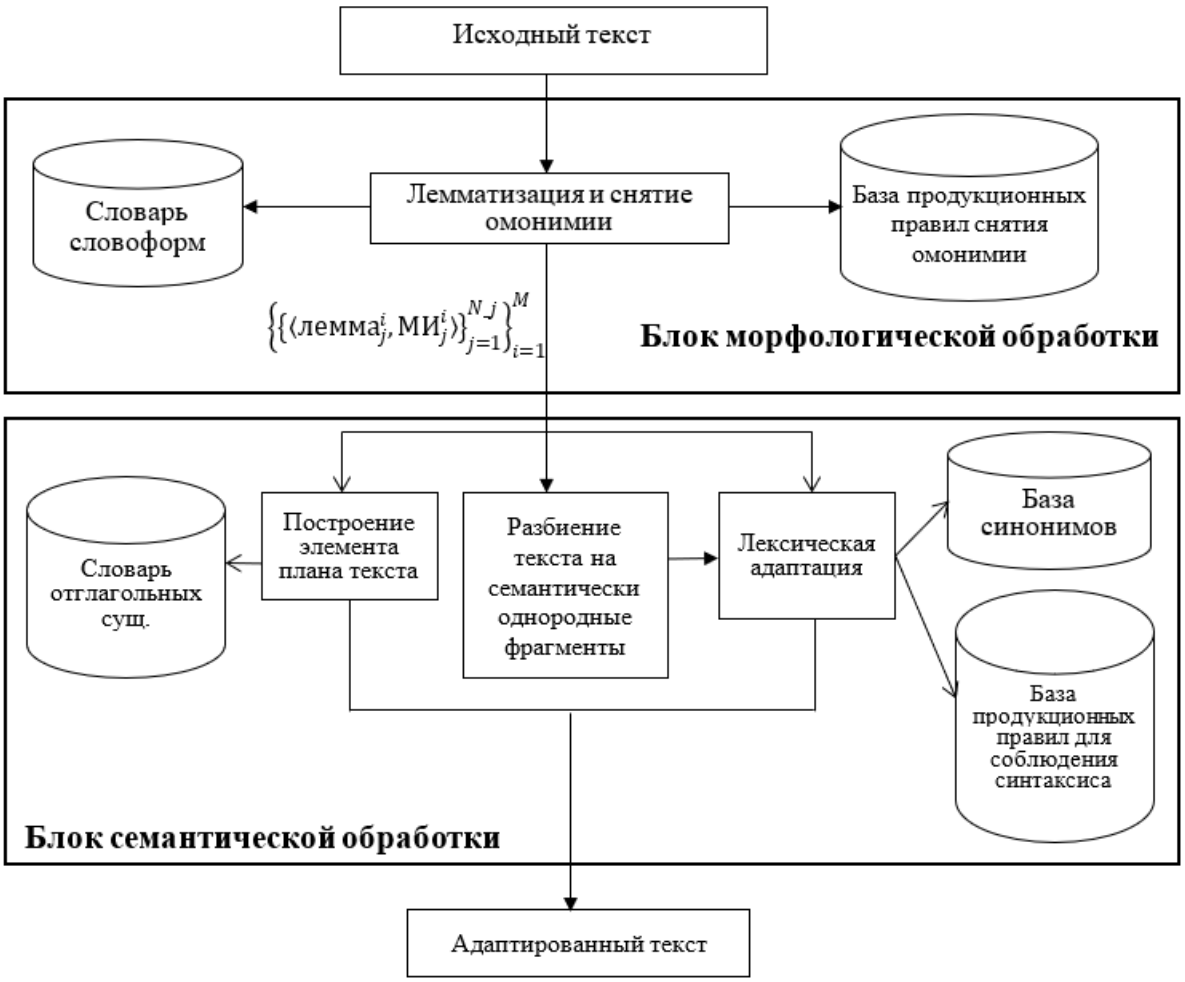

Рисунок 3 - Общая схема работы системы обработки и анализа текстовой информации

В подразделе 4.2.1 описывается формирование размеченной базы синонимов на основе данных из открытых источников. В качестве лингвистической основы в работе используется издание Александровой З.Е. «Словарь синонимов русского языка: Практический справочник» // М.: Рус. яз., 2001г. В основе организации данного словаря лежит понятие синонимического ряда. Синонимический ряд начинается доминантой, за ним следуют синонимы – члены ряда. Идея обсуждаемого метода упрощения русскоязычного текста кратко формулируется так: член синонимического ряда, встретившийся в тексте, должен быть заменен соответствующей доминантой, т.к. доминанта является более общим и более употребительным синонимом.

При создании программы автоматической замены нельзя использовать доминанты и синонимические ряды словаря З.Е. Александровой совершенно

формально, т.к. не все члены синонимического ряда могут быть заменены на доминанту.

Например: *телевизор! телик, ящик, телевизионный ящик.* Нельзя каждый раз «ящик» заменять на «телевизор», т.к. у этого слова более широкое значение.

В связи с этим был проведен анализ и сокращение предлагаемых синонимических рядов. Кроме того, с помощью электронного издания О.Н. Ляшевской и С.А. Шарова «Частотный словарь современного русского языка (на материалах Национального корпуса русского языка)» М.: Азбуковник, 2009 для каждой группы проанализирована частотность доминанты и членов синонимического ряда, и при возможности в качестве доминанты выбран синоним с наибольшей частотностью.

База синонимов разделена на 3 части: база отдельных слов, словосочетаний и неизменяемых словосочетаний. Для системы синонимических замен для замены отдельных слов создается файл *База.txt*, представляющий собой последовательность групп, каждая из которых начинается соответствующей доминантой (отмечается «!»), а затем содержит отрезок синонимического ряда (см. приложение Б). Пример такой группы:

бесконечность !
безграничность
безбрежность
безмерность
бескрайность
беспредельность

В базу включены только начальные формы слов и для работы с ней используется лемматизация исходного текста. В базе синонимов могут встречаться слова, которым соответствуют несколько доминант. Тогда программа адаптации при замене выдаст несколько вариантов в скобках, так что нужное может быть выбрано одним щелчком мыши. Это относится и к заменам словосочетаний (см. разделы ниже).

Также подраздел 4.1 посвящен разметке Базы с целью восстановления после замены правильного синтаксиса. Для этого используются следующие обозначения, проставляемые при необходимости в базе справа от соответствующих членов синонимического ряда.

| - означает, что последующая запись относится к заменяемому слову;

/ - означает, что последующая запись относится к слову, которое непосредственно предшествует заменяемому;

\ - означает, что последующая запись относится к слову, которое непосредственно следует за заменяемым;

Остальное разъясним на конкретном примере. Запись «\ (с) тв-дат» означает, что творительный падеж с предлогом «с» должен быть заменен на дательный падеж без предлога. Пример: «Раскланялся с соседями» программа превратит в «Поклонился соседям».

Подраздел 4.2.2 описывает основные этапы работы системы синонимических замен (рисунок 4).

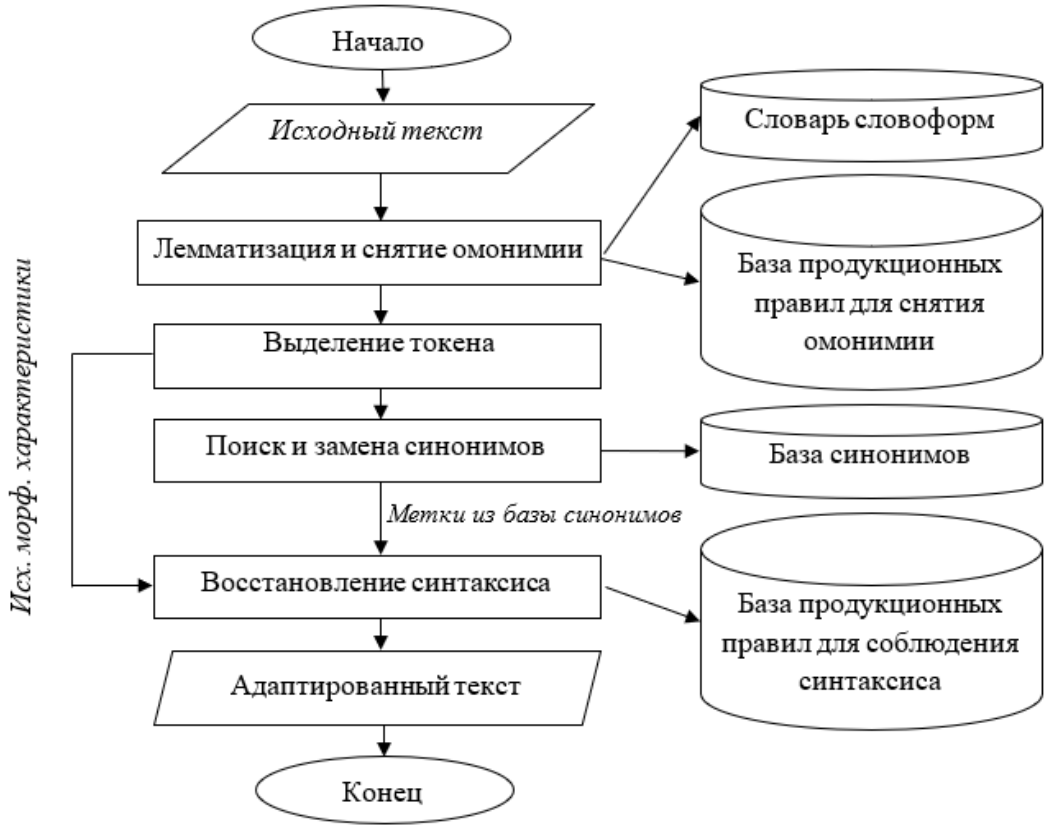

Рисунок 4 - Блок-схема алгоритма реализации метода синонимических замен слов

Подраздел 4.2.3 посвящен синонимическим заменам отдельных слов и неизменяемых словосочетаний, сформулированы первоначальные правила порождения нужной формы для слова-замены при замене одиночных слов.

В подразделе 4.2.4 описаны правила синонимических замен словосочетаний, сформулированы первоначальные правила порождения нужной грамматической формы для слова-замены при замене словосочетаний. Также подраздел описывает процедуру выделения в тексте словосочетания для замены при свободном порядке слов.

Подраздел 4.2.5 содержит примеры тестирование работы системы синонимических замен на материалах Национального корпуса русского языка. Проведенные численные исследования показали, что разработанный метод упрощения текста позволяет успешно осуществлять замену отдельных слов и словосочетаний в тексте с соблюдением правильного синтаксиса и сохранением семантики текста с точностью выше 96%.

В подразделе 4.3 изложен авторский алгоритм разбиения текста на семантически однородные фрагменты. Он основан на использовании наиболее часто встречаемых существительных (ключевых словах) и местах их концентрации в тексте.

Формируется список лемм всех слов текста, которые являются существительными и встречаются в тексте не менее двух раз (ключевые слова). Для каждой леммы подсчитывается a – количество вхождений соответствующих словоформ (частота), b – номер первого предложения, где

встречается слово и *c* – номер последнего предложения, где оно встречается. Далее вычисляется отношение

$$a/(c-b+1) \qquad (1)$$

Список ключевых слов отсортирован по величине *a*. В нем находится первый максимум отношения (1) и дальнейшие операции этапа выполняются с соответствующим словом. Программа выделяет в качестве абзаца отрезок текста от предложения номер *b* до предложения номер *c* включительно. Затем она подсчитывает количество предложений в образовавшихся предыдущем и последующем абзацах, выделяет наибольший из них и выдает вышеописанную информацию о ключевых словах в этом абзаце. Этим заканчивается первый этап алгоритма. Далее этот этап повторяется по отношению к абзацу, выделенному на первом этапе, и так далее.

После рассматривается абзац, в первом предложении которого есть словоформа местоимения «он» или «она» или «оно». Если в нем нет предшествующего существительного в том же роде и числе, то абзац присоединяется к предыдущему. Это делается для всех таких абзацев. Объединяются также соседние абзацы, на стыке которых оказался общий фрагмент прямой речи. Наконец, абзац, состоящий из одного предложения, присоединяется к меньшему из соседних.

В подразделе 4.4 предлагается алгоритм автоматического создания элемента плана текста на основе использования отглагольных существительных. Пусть есть предложение, описывающее некоторое действие или событие с использованием переходного глагола. Полагаем, что одним из наиболее общих выразителей важнейшего смысла, заключенного в таком предложении, может служить отглагольное существительное. Разработанная программа, заменяет глагол соответствующим отглагольным существительным и винительный падеж существительного (прямое дополнение) родительным.

Например, результатом работы программы с предложением: «По телевидению передают важное сообщение» будет словосочетание «Передача сообщения». Оно и является носителем основной информации.

Программа использует файл *Глаг-сущ.txt.* Файл состоит из групп, каждая из которых содержит две строки. Первая включает набор глаголов, а вторая – соответствующее отглагольное существительное. Группы разделены пробельными строками.

Пример:

доверять, передоверять, доверить, передоверить, доверяться, передоверяться, довериться, передовериться

доверие

доминировать

доминирование

и т.д.

Получившийся словарь глаголов – отглагольных существительных представляет собой некоторый самостоятельный лингвистический продукт.

Работа содержит 7 приложений, содержащие лексические базы для синонимических замен отдельных слов, словосочетаний и неизменяемых по форме словосочетаний, численные исследования.

ЗАКЛЮЧЕНИЕ

Диссертация является законченной научно-исследовательской работой, в которой получено решение актуальной научно-технической задачи повышения эффективности обработки и анализа текстовой информации в контексте решения задач снятия омонимии и применения способов лексической адаптации. Основные научные результаты и выводы состоят в следующем.

1. Анализ состояния исследований в области обработки текстовой информации показал, что «узким» местом стандартных подходов разрешения омонимии являются предикативы и предикативные словосочетания, деепричастия, группы наречие-существительное. Представляется наиболее перспективным использовать: синонимические замены для лексического упрощения текста на основе базы синонимов и правил, позволяющих соблюдать правила синтаксиса; словарные методы для лемматизации совместно с методами для разрешения омонимии, основанными на правилах, для чего необходимо формализовать лингвистические знания для снятия омонимии в представительную базу правил; префиксные деревья как структуру данных для представления морфологического словаря.

2. Для формирования словаря русских словоформ для лемматизации использован словарь русских парадигм, находящийся в открытом доступе, а также префиксное дерево внутреннего представления множества всех словоформ, которое позволяет проводить эффективный поиск всех словоформ, соответствующих заданной последовательности символов. Словарь пополнен новыми словоформами, лемма добавлена в каждую его строку. Объем словаря составляет более 4 млн. словоформ для более 130 тыс. лемм, а лемматизация происходит за один проход с той же скоростью, что и поиск вхождений анализируемой словоформы.

3. Предложен декларативно-процедурный метод автоматического разрешения частеречной омонимии для предикативов и предикативных словосочетаний, деепричастий, а также групп наречие-существительное. Помимо морфологического словаря, где предикативные неделимые словосочетания помечены как цельные единицы, метод использует:

– размеченные словари предложных групп, индикаторов предикатива для снятия омонимии предикатив-существительное, глаголов, употребляемых с наречием или существительным для снятия омонимии наречие-существительное

– продукционная база правил на основе словарей и содержащихся в них меток, которая дополнена разработанными для конкретных словосочетаний правилами для случаев нерегулируемых метками.

Метод снимает частеречную омонимию предикативов и предикативных словосочетаний, деепричастий, групп наречие-существительное с точностью 99,3%.

4. Разработан метод автоматического разбиения текста на абзацы как семантически однородные фрагменты за счет предложенного отношения, учитывающего частоту встречаемости слова и длину отрезка текста, где оно встречается. Предложенный подход является статистическим, поэтому не требует специальных лингвистических знаний, кроме морфологического словаря и простых правил, учитывающих анафорические ссылки, характеризуется малой вычислительной сложностью и высокой точностью.

5. Для построения элемента плана текста сформирован текстовый корпус – словарь отглагольных существительных, объемом более 10 000 групп, содержащих глаголы и соответствующие им существительные. Использование этого словаря позволяет формировать элемент плана текста, заменяя глагол в предложении соответствующим отглагольным существительным и винительный падеж существительного, являющегося прямым дополнением, родительным.

6. Для формирования размеченной базы синонимов использованы словари синонимов, находящие в открытом доступе. Для сохранения семантики проведен анализ и сокращение синонимических рядов, проанализирована частотность членов синонимического ряда с целью выбора доминанты, а также проведена разметка записей в словарях и предложен механизм обработки меток для соблюдения правила синтаксиса в упрощенном тексте.

7. Разработана база продукционных правил для сохранения правильного синтаксиса после лексической адаптации текста, позволяющая проводить корректную замену отдельных слов, словосочетаний одним словом и словосочетаний словосочетанием.

8. На основе размеченной базы синонимов и базы правил соблюдения синтаксиса разработан метод упрощения текста путем замены фрагмента текста более простым и употребительным синонимом. На материалах Национального корпуса русского языка проведена оценка его эффективности по критериям: сохранение семантики, соблюдение синтаксиса и упрощение. Разработанный метод позволяет успешно осуществлять замену отдельных слов и словосочетаний в тексте с соблюдением правильного синтаксиса и сохранением семантики текста с точностью выше 96%.

Разработанные методы и алгоритмы, а также размеченные текстовые корпуса и базы синонимов могут быть использованы для задач адаптации, поисковой оптимизации и автоматического реферирования текстов, а также автоматическом переводе. Помимо этого, может быть указан ряд практических приложений адаптации: подготовка учебных материалов, текстов художественной литературы для иностранцев, изучающих русский язык; преобразование сложных текстов в тексты на понятном языке для людей, знание языка которых не позволяет в достаточной степени понять сложную текстовую информацию, в частности, для людей с первыми симптомами когнитивных нарушений, связанных с возрастом или травмами головного мозга, для детей с задержками речевого развития.

Перспективы дальнейшей разработки темы связаны с расширением области применения разработанных методов и алгоритмов для решения других задач компьютерной обработки текстовой информации. Например, можно исследовать возможности использования этих методов для автоматического определения тональности текста, извлечения информации, машинного перевода и других задач. Кроме того, дальнейшее развитие темы может включать разработку специализированных онлайн-приложений и инструментов для облегчения работы с текстовыми данными.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ АВТОРОМ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в рецензируемых научных изданиях, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата и доктора наук:

1) Ниценко, А. В. Автоматическая лексическая адаптация русскоязычных текстов / А. В. Ниценко, В. Ю. Шелепов, **С. А. Большакова** // Искусственный интеллект и принятие решений. – 2025. – № 1. – С. 82-94. – DOI 10.14357/20718594250107. – EDN NJYPOD.

2) **Большакова, С. А.** Система автоматической адаптации русскоязычных текстов и ее практическая значимость / **С. А. Большакова** // Проблемы искусственного интеллекта. – 2024. – № 3(34). – С. 45-54. – DOI 10.24412/2413-7383-2024-3-45-54. – EDN HJUWYF.

3) Ниценко, А. В. О снятии омонимии «предикатив-предложная группа» для некоторых русских словосочетаний / А. В. Ниценко, В. Ю. Шелепов, **С. А. Большакова** // Проблемы искусственного интеллекта. – 2023. – № 2(29). – С. 49-57. – EDN YAUQXU.

4) **Большакова, С. А.** О снятии омонимии «предикатив-предложная группа» для некоторых распространенных словосочетаний в русскоязычных текстах / **С. А. Большакова** // Проблемы искусственного интеллекта. – 2023. – № 1(28). – С. 11-17. – EDN OKVSFX.

5) **Большакова, С. А.** К вопросу об автоматическом снятии омонимии русских деепричастий / **С. А. Большакова**, А. В. Ниценко, В. Ю. Шелепов // Проблемы искусственного интеллекта. – 2021. – № 4(23). – С. 37-45. – EDN CNHQDL.

6) Ниценко, А. В. Об автоматическом снятии омонимии предикативных словосочетаний. Результаты работы с национальным корпусом русского языка / А. В. Ниценко, В. Ю. Шелепов, **С. А. Большакова** // Проблемы искусственного интеллекта. – 2021. – № 3(22). – С. 46-56. – EDN FTFCCH.

7) Ниценко, А. В. О снятии омонимии словосочетаний, которые могут быть предикативами / А. В. Ниценко, В. Ю. Шелепов, **С. А. Большакова** // Проблемы искусственного интеллекта. – 2021. – № 1(20). – С. 53-62. – EDN IHBIAO.

8) Русское синтаксическое управление при словесных заменах. О словах с функциями наречия и существительного / А. В. Ниценко, В. Ю. Шелепов, **С. А. Большакова**, К. С. Ивашко // Проблемы искусственного интеллекта. – 2020. – № 2(17). – С. 46-57. – EDN VCQSHH.

9) О словесных заменах, сохраняющих смысл русского предложения / А. В. Ниценко, В. Ю. Шелепов, **С. А. Большакова**, К. С. Ивашко // Проблемы искусственного интеллекта. – 2020. – № 1(16). – С. 63-74. – EDN ONSRWE.

Публикации в иных изданиях:

10) Ниценко, А. В. Лексико-синтаксический метод снятия омонимии в русскоязычных текстах / А. В. Ниценко, В. Ю. Шелепов, **С. А. Большакова** // Речевые технологии. – 2023. – № 2. – С. 40-48. – EDN UCLBQC.

Публикации по материалам конференций:

11) **Большакова, С. А.** Автоматизированная система упрощения русскоязычных текстов / **С. А. Большакова** // II Всероссийская школа Национального центра физики и математики для студентов, аспирантов, молодых ученых и специалистов по искусственному интеллекту и большим данным в технических, промышленных, природных и социальных системах. Тезисы. – г. Саров: ФГУП «РФЯЦВНИИЭФ» - 2024. - С. 29-31.

12) **Большакова, С. А.** Практическое применение системы автоматической адаптации русскоязычных текстов / **С. А. Большакова** // Искусственный интеллект: теоретические аспекты, практическое применение : материалы Донецкого международного научного круглого стола, Донецк, 30 мая 2024 года. – Донецк: ФГБНУ "Институт проблем искусственного интеллекта", 2024.– С. 11–15. – EDN FBZFNY.

13) **Большакова, С. А.** К вопросу о снятии омонимии "предикатив - предложная группа" / **С. А. Большакова** // Искусственный интеллект: теоретические аспекты, практическое применение : материалы Донецкого международного научного круглого стола, Донецк, 24 мая 2023 года. – Донецк: Федеральное государственное бюджетное научное учреждение "Институт проблем искусственного интеллекта", 2023. – С. 25-28. – EDN FVCVTM.

14) **Большакова, С. А.** К вопросу о снятии омонимии в некоторых группах омонимов, включающих предикатив / **С. А. Большакова**, А. В. Ниценко, В. Ю. Шелепов // Искусственный интеллект: теоретические аспекты, практическое применение : Материалы Донецкого международного научного круглого стола, Донецк, 25 мая 2022 года. – Донецк: Государственное учреждение Институт проблем искусственного интеллекта, 2022. – С. 152-158. – EDN VWSBVF.

15) Ниценко, А. В. Исследование омонимии предикативных словосочетаний на основе национального корпуса русского языка / А. В. Ниценко, В. Ю. Шелепов, **С. А. Большакова** // Современные информационные технологии в образовании и научных исследованиях (СИТОНИ-2021) : Материалы VII Международной научно-технической конференции, Донецк, 23 ноября 2021 года / Под общей редакцией В.Н.

Павлыша. – Донецк: Донецкий национальный технический университет, 2021. – С. 510-514. – EDN JSNBFV.

16) **Большакова, С. А.** К вопросу об автоматическом снятии омонимии русских деепричастий / **С. А. Большакова** // Искусственный интеллект: теоретические аспекты, практическое применение : Материалы Донецкого международного научного круглого стола, Донецк, 27 мая 2021 года. – Донецк: Государственное учреждение Институт проблем искусственного интеллекта, 2021. – С. 120-123. – EDN EMNUWA.

17) **Большакова, С. А.** Об автоматизированных системах адаптации русскоязычных текстов / **С. А. Большакова** // Искусственный интеллект: теоретические аспекты, практическое применение : материалы Донецкого международного научного круглого стола, Донецк, 27 мая 2020 года. – г. Донецк: Государственное учреждение Институт проблем искусственного интеллекта, 2020. – С. 27-32. – EDN QASKGE.

Свидетельства о государственной регистрации программ для ЭВМ:

1) Свидетельство о государственной регистрации программы для ЭВМ № 2025611191 Российская Федерация. Экспериментальное программное обеспечение для морфологической разметки текста со снятием омонимии : № 2024693205 : заявлено 26.12.2024 : опубликовано 16.01.2025 / Шелепов В.Ю., Ниценко А.В., **Большакова С.А.** ; заявитель Федеральное государственное бюджетное научное учреждение «Институт проблем искусственного интеллекта».