Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное научное учреждение
«Институт проблем искусственного интеллекта»

На правах рукописи

Большакова Светлана Анатольевна

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)

Диссертация
на соискание ученой степени
кандидата технических наук

Научный руководитель
доктор физико-математических наук,
профессор Шелепов В. Ю.

г. Донецк – 2025

ОГЛАВЛЕНИЕ

ПЕРЕЧЕНЬ ОБОЗНАЧЕНИЙ И СОКРАЩЕНИЙ

В настоящей работе применяют следующие сокращения и обозначения

NLP – Natural Language Processsing

ЕЯ – естественный язык, естественно-языковый

ИИ – искусственный интеллект

МА – морфологический анализ

МИ – морфологическая информация

МС – морфологический словарь

НИР – научно-исследовательская работа

НКРЯ – национальный корпус русского языка

ЯМ – языковая модель

ВВЕДЕНИЕ

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 142 страницах, иллюстрируется 9 рисунками и содержит 19 таблиц.

РАЗДЕЛ 1
ОБЗОР СУЩЕСТВУЮЩИХ ТЕХНОЛОГИЙ И МЕТОДОВ РАБОТЫ С СИНОНИМИЕЙ И ОМОНИМИЕЙ

В аспекте любых NLP-задач естественный язык рассматривается как система, где каждая единица тесно связана с другими: «Современный русский литературный язык представляет собою сложную систему, части которой находятся в отношениях постоянной и необходимой взаимосвязи: один участок этой системы не существует без другого» [1].

В настоящем исследовании речь пойдет о таких явлениях языка, как *омонимия* и *синонимия*, которые рассматриваются на уровне не только лексическом, но и морфологическом и синтаксическом, т.е. будут представлены как системные свойства языка в целом.

Явление синонимии проявляется в том, что один и тот же смысл может быть выражен различными синтаксическими конструкциями и различными словами. Путем использования особенностей синонимии может выполняться упрощение (адаптация) текста – преобразование его к более простой и возможно более краткой форме. Упрощение языка – один из распространенных способов экономии речевых усилий и времени для быстрого и лучшего понимания сказанного, охвата большей целевой аудитории с разным уровнем владения языком, что позволяет максимально упростить языковое общение и создает наиболее комфортные условия для обмена текстовой информацией [2].

Явление омонимии в меньшей степени системно, поскольку связь омонимов чисто формальная и основана на полном совпадении формы. Однако, омонимы – отражение не только случайных совпадений, но и системности языка, в структуре которого, заложены зоны для подобных двойников. Тем более что омонимия охватывает не только лексический уровень, но и словообразовательный, и грамматический.

Разрешение многозначности является одной из важнейших задач автоматической обработки естественного языка. Для решения проблемы

омонимии существует несколько подходов, которые основаны на правилах, статистике и машинном обучении.

Для компьютерной обработки текстовой информации необходимо определить принадлежность каждой словоформы к парадигме определенной лексемы и ее грамматические признаки, т. е провести морфологический анализ. В данном разделе проанализированы методы морфологического анализа и внутреннего представления морфологического словаря. А также рассмотрено применение нейросетевых языковых моделей при решении задачи обработки естественного языка.

1.1 Использование синонимии при автоматизированной адаптации текста

Автоматическая адаптация текста – это процесс упрощения письменного текста с сохранением его смысла и структуры. Автоматизированная адаптация текста включает лексическое и синтаксическое упрощение текста. Лексическое упрощение заключается в замене сложных слов на более простые синонимы, а синтаксическое – в изменении структуры предложения для облегчения его понимания.

Синонимы – слова или словосочетания, различные по произношению и написанию, но имеющие схожее лексическое значение: бежать – мчаться, большой – огромный, стужа – холод [3].

Два и более лексических синонима образуют в языке определенную группу, которая иначе называется синонимическим рядом. Основное слово синонимического ряда, передающее наиболее общее понятие и являющееся нейтральным по употреблению, называется доминантой синонимического ряда. Остальные слова синонимического ряда выражают дополнительные оттенки.

Поскольку доминанта выражает понятие, свойственное всем словам, входящим в данный синонимический ряд, она обычно располагается в начале синонимического ряда.

С точки зрения постоянства состава слов синонимические ряды характеризуются незамкнутостью. В них возможны изменения и дополнения, обусловленные протекающим процессом развития всей лексической системы.

Детекция сложных слов в тексте для замены может происходить по нескольким критериям:

– Сложность слова. Для этого учитывают количество слогов (больше трёх) и частоту встречаемости [4].

– Связь с доменной тематикой. Например, «ирригация» можно заменить на «орошение» [5].

– Частота слова. Замена редких слов на более частотные синонимы применяется в системах лексического упрощения для снижения сложности текста [6].

При замене слов их синонимами необходимо учитывать ряд тонкостей языка, чтобы не допустить речевые ошибки. Синонимы могут отличаться не только оттенками значения и стилистической окраской, но и сочетаемостью с другими словами. Например, слова «серый» и «пасмурный» являются синонимами, но «серым» может быть и костюм, а «пасмурным» – только день.

При лексической адаптации в англоязычной среде популярным решением является применение семантической сети WordNet. Слова в WordNet сгруппированы по наборам когнитивных синонимов, которые называются синсетами. В каждый синсет входят слова, которые не просто схожи, а передают один и тот же смысл, подходящий для разных контекстов. Это значит, что слова не только «родственны», но и имеют схожие ассоциации, эмоциональные оттенки и даже подтексты. Лексическое упрощение может быть достигнуто через перефразирование и замену слов синонимами из словаря [7] или объяснения слов с использованием словарных определений [8].

Распространенной проблемой, связанной с заменой слова на синоним, является нарушение правил синтаксиса при замене. Это может произойти в связи с разными грамматическими характеристиками заменяемого слова и его

синонима. При разработке методов автоматического лексического упрощения текста путем синонимических замен необходимо соблюдать правила синтаксиса.

В настоящее время нет точного стандарта оценки качества автоматического упрощения текста [9]. Чтобы получить общее представление о характеристиках оригинальных и адаптированных текстов могут использоваться морфологические, лексические и синтаксические характеристики оригинальных и адаптированных текстов. Обычно используются такие критерии, как удобочитаемость текста и легкость восприятия целевой аудиторией. Оценка проводится с помощью автоматических метрик или экспертами. Легкость текста оценивается по следующим средним параметрам: количество слов в тексте, количество предложений, длина слова в слогах, длина слова, длина предложения, количество пунктуации на предложение.

В некоторых работах результат упрощения оценивается с помощью метрики BLEU. «BLEU – это показатель качества для систем вывода текста, который пытается измерить соответствие между результатами машинного перевода и человеческим переводом. Основная идея BLEU заключается в том, что чем ближе машинный перевод к профессиональному человеческому переводу, тем он лучше. Оценки BLEU отражают только то, как система работает с определенным набором исходных предложений и переводов, выбранных для теста. Поскольку выбранный перевод для каждого сегмента может быть не единственным правильным, часто можно получить плохие оценки хороших переводов. В результате оценки не всегда отражают реальную потенциальную производительность системы, особенно по содержанию, которое отличается от конкретного тестового материала» [10].

Использование стандартных метрик, пришедших из машинного перевода, не является оптимальным решением для оценки качества упрощения, т.к. они направлены на то, чтобы сравнивать решение с одним эталонным ответом. При упрощении текста можно хорошо решить задачу несколькими способами – оба будут хороши, но друг на друга совсем не похожи. Поэтому сравнивать все

варианты с одним эталоном некорректно. А если сравнивать не с одним эталоном, тогда непонятно, как оценивать.

Ни одна из этих метрик не является полностью подходящей для оценки упрощённых текстов. Типичное оценивание также может включать опрос экспертов на предмет определения правильности и полезности замен. Это не всегда оптимально, но в сочетании с несколькими разными метриками можно получить некоторое представление о том, насколько хорошо выполнено упрощение.

В настоящее время разрабатываются метрики специально для оценки упрощения текста [11]. Чтобы оценить качество произведенного упрощения текста выполняют вычисление разных лингвистических параметров: количество слов, длина слов, количество слогов и так далее.

Современные исследования автоматического упрощения текста охватывают различные направления.

В работе W.Coster и др. рассматривается снижение сложности предложений за счет включения более доступной лексики и структуры предложений. Авторами сформирован новый набор данных, который объединяет английскую Википедию с Simple English Википедии. Данные содержат полный спектр операций по упрощению, включая изменение формулировок, переупорядочение, вставку и удаление. Используется сопоставление оригинальных и упрощённых статей для создания параллельных корпусов. Полученный корпус был проверен с использованием системы машинного перевода Moses. Качество переводов было оценено с помощью метрики BLEU, которая показала, что использование набора упрощённых предложений обеспечивает лучшее качество перевода по сравнению с набором неупрощённых предложений [12].

В работе K. Woodsend и M. Lapata представлена управляемая данными модель, основанная на квазисинхронной грамматике, формализме, который может естественным образом фиксировать структурные несоответствия и сложные операции перезаписи. Данная грамматика создана на основе параллельного корпуса: оригинальные статьи из Wikipedia и их упрощённые варианты из

SimpleWiki. Статьи переписываются с использованием грамматик, позволяющих применять лексические и синтаксические упрощения, включая разбиение фраз [13].

Упрощение текста может выполнятся для различных целевых аудиторий. Исследование J. De Belder и др. направлено на адаптацию текстов для детей. Для синтаксического упрощения предложений предлагается разбивать их на части, а для лексического упрощения – заменять сложные слова более простыми синонимами. Эффективность этого подхода тестировалась для каждого компонента отдельно и глобально при автоматическом создании упрощённых новостных и энциклопедических статей. Использование языковой модели на этапе лексического упрощения позволило достичь лучших результатов по сравнению с базовым методом. Однако синтаксическое упрощение показало сложности с распознаванием некоторых явлений с помощью синтаксического анализатора и частые ошибки. Упрощенный текст менее сложный, чем оригинал, но не достаточно простой для маленьких детей [14, 15].

Другим возможным направлением адаптации является упрощение медицинских и научных текстов. Сложный жаргон часто делает научную работу менее доступной для широкой публики. Одной из стратегий предоставления информации о научных достижениях в доступной и увлекательной форме является использование более простых терминов вместо сложного жаргона. Чтобы помочь в этом процессе, Kim Y., Hullman J.R и Adar E. в работе [16] предлагают систему DeScipher для редактирования текста, которая подсказывает и ранжирует возможные упрощения сложной терминологии для журналиста во время написания статьи. DeScipher применяет правила упрощения, основанные на большой коллекции научных рефератов и связанных с ними авторских резюме, и учитывает текстовый контекст при составлении предложений журналисту.

Исследование Lu J и др. посвящено упрощению доступа к медицинской литературе. В работе предложена двухэтапная стратегия NaPSS «обобщить, а затем упростить», которая позволяет определить релевантный контент для упрощения, сохраняя при этом исходный поток повествования. При таком подходе сначала создаются справочные резюме с помощью сопоставления

предложений между оригинальным и упрощенным резюме. Эти резюме затем используются для обучения экстрактивного составителя резюме, который изучает наиболее релевантный контент, подлежащий упрощению. Затем, чтобы обеспечить последовательность изложения упрощенного текста, синтезируются вспомогательные подсказки, объединяющие ключевые фразы, полученные в результате синтаксического анализа исходного текста. Данная модель дает результаты, значительно превосходящие исходные данные seq2seq по английскому медицинскому корпусу, обеспечивая абсолютное улучшение лексического сходства на 3-4% и обеспечивая дополнительное улучшение показателя SARI на 1,1% в сочетании с исходными данными. Авторы также подчеркивают недостатки существующих методов оценки и вводят новые показатели, которые учитывают как лексическое, так и семантическое сходство высокого уровня. Эффективность предложенного подхода также подтверждает оценка, проведенная человеком на случайной выборке из набора тестов [17].

Большая часть существующих работ по упрощению текста ограничена входными данными на уровне предложений, а попытки итеративного применения этих подходов к упрощению на уровне документа не приводят к последовательному сохранению структуры документа. L. Cripwell в работе [18] предлагают использовать план упрощения. Авторы рассматривают план как последовательность меток, каждая из которых описывает одну из четырех операций упрощения на уровне предложений (копирование, перефразирование, разделение или удаление). Модель планирования помечает каждое предложение во входном документе, принимая во внимание как его контекст (окно с окружающими предложениями), так и его внутреннюю структуру (представление на уровне токенов). Эксперименты с двумя тестами упрощения (Newsela-auto и Wiki-auto) показывают, что данный подход превосходит базовые показатели при упрощении на уровне документов.

Работы H. Saggion и др. [19, 20] посвящены автоматическому упрощению текста для испанского языка. Модульная система Simplext предлагает синтаксическое и лексическое упрощение, которые основаны на анализе корпуса,

упрощенного вручную для людей с особыми потребностями. Автоматическая оценка результатов работы системы принимает во внимание взаимодействие между тремя различными модулями, посвященными различным аспектам упрощения. Одна из оценок основана на показателях удобочитаемости для испанского языка и показывает, что система способна снизить лексическую и синтаксическую сложность текстов. В большинстве случаев смысл предложения сохраняется. Работа представляет собой одну из первых систем автоматического упрощения текста для испанского языка, которая учитывает различные лингвистические аспекты, и сопоставима с достижениями в области автоматического упрощения текста на английском языке.

В работе K. Inui и др. представлены результаты исследования по упрощению текстов на японском языке для оказания помощи в чтении людям с врожденной глухотой. Для оценки удобочитаемости предложен новый подход, в рамках которого были проведены анкетные опросы для сбора данных об оценке удобочитаемости и использовался эмпирический метод, основанный на корпусах данных, для получения модели ранжирования удобочитаемости. Результаты опросов показывают потенциальное влияние упрощения текста на удобство чтения. Предложен усовершенствованный механизм лексико-структурного перефразирования. Были вручную обработали более тысячи правил упрощения, которые реализуют широкий спектр лексических и структурных перефразировок [7, с. 14].

Исследование В.Г. Сибирцевой и др. посвящено использованию аутентичных материалов лингвистического корпуса в рамках проекта «Создание электронного учебника русского языка как иностранного». Особое внимание уделено автоматической адаптации лингвистического материала. Разработанный продукт понижает сложность оригинальных текстов с точки зрения их синтаксической и морфологической структуры и словарного запаса [21, 22].

Проблеме подготовки адаптированных текстов для изучающих иностранный язык также посвящено исследование J. Burstein. Адаптация текста - это педагогическая практика, используемая для улучшения понимания

прочитанного и развития навыков владения языком у изучающих новый язык. Практика адаптации текста предполагает внесение преподавателем изменений в тексты, чтобы сделать их более понятными, учитывая уровень чтения учащегося. Адаптация преподавателя включает в себя краткое изложение текста, поддержку словарного запаса (например, предоставление синонимов) и перевод. Это трудоемкая, но крайне важная практика для преподавателей, поскольку часто бывает трудно найти тексты, подходящие для чтения на соответствующем уровне. Авторы разработали инструмент автоматической адаптации текста, который автоматически генерирует текстовые адаптации, аналогичные тем, которые могут создавать учителя. Синонимы для более низкочастотных (более сложных) слов слова выводятся с использованием статистически сгенерированной матрицы сходства слов. Разработанная система ориентирована на упрощение текстов при изучении английского для носителей испанского языка. При подборе синонимов отдается предпочтение когнатам - словам, имеющим одинаковое написание и значение в двух языках [23].

Для данного исследования также могут быть интересными работы, посвященные упрощению для языков романо-германской группы [24-40].

В настоящее время разработано и активно разрабатывается множество систем автоматического упрощения текста для разных целевых аудиторий и языков. Однако среди уже существующих систем автоматического упрощения текста славянские языки представлены недостаточно. Это делает данную работу актуальной и практически значимой.

1.2 Автоматическая обработка языка на морфологическом уровне

Для корректной работы систем автоматической обработки языка необходимо провести морфологический анализ, на этапе которого решается проблема морфологической омонимии, поскольку снятие омонимии является необходимым и важным этапом для качественного машинного анализа текстов и, в конечном итоге, понимания и извлечения знаний из них.

Цель морфологического анализа (МА) – определить принадлежность некоторой словоформы к парадигме определенной лексемы и грамматические признаки для этой словоформы – морфологическую информацию (МИ) для использования ее на последующих этапах обработки ЕЯ текста.

Так для существительных этими признаками будут: род, число, падеж и склонение, для прилагательных: род, число и падеж; для глаголов – время, лицо, число, спряжение, вид; для местоимений – число и лицо.

«Причины многообразия методов морфологической обработки текстов на русском языке заключаются в сложности его морфологии. Русский язык обладает рядом особенностей, которые осложняют его морфологическую обработку:

1. Супплетивизм – образование словоизменительной формы некоторого слова уникальным для языка образом (*идти – шел*).

2. Чередование или выпадение букв в основе (*собирать – собрать*).

3. Синонимия – образование словоформы более чем одним способом. Часто это встречается при образовании сравнительной степени прилагательных и наречий (*сильнее – сильней*), а также творительного падежа единственного числа некоторых существительных (*водой – водою*).

4. Омонимия – наличие у одной грамматической формы нескольких грамматических значений; например, словоформа *стекло* может быть существительным среднего рода, единственного числа, именительного или винительного падежа или формой глагола *стекать* изъявительного наклонения прошедшего времени совершенного вида.

5. Синтетические (простые: *делать*, *говорил*) и аналитические (сложные: *буду делать*, *говорил бы*) формы слов.

Перечисленные особенности (сложности) морфологии необходимо учитывать при разработке методов генерации и определения форм слов [41]».

1.3 Методы разрешения омонимии

Омонимы (от др.-греч. *homos* – одинаковый + *onyma* – имя) – слова и словосочетания, разные по значению, но одинаковые по звучанию и написанию: *заставить*[1] «загородить чем-л. поставленным» и *заставить*[2] «принудить кого-то что-то сделать».

Омонимы в меньшей степени системны, чем остальные единицы лексики, поскольку их связь чисто формальная, основанная на полном совпадении формы. В связи с этим некоторые учёные отказывают омонимам в системности вообще: «Асистемность омонимии проявляется в том, что значения слов-омонимов совершенно не связаны между собой, не имеют в своём составе никаких общих сем и поэтому вообще несопоставимы» [42]. Однако, омонимы – отражение не только случайных совпадений, но и системности языка, в структуре которого, заложены зоны для подобных двойников. Тем более что омонимия охватывает не только лексический уровень, но и словообразовательный, и грамматический.

Многозначность словоформ – «одна из природных особенностей естественного языка, способствующая качественному развитию словарного запаса, тем самым «экономящая» словесный материал. Разрешение многозначности (дизамбигуация или снятие омонимии) является одной из важнейших задач автоматической обработки естественного языка» [43]. Исследователи выделяют несколько типов многозначности естественного языка: морфологическую, синтаксическую и лексико-семантическую многозначности.

Для решения проблемы омонимии существует несколько подходов, которые основаны на правилах; статистике [44, 45, 46, 47] и машинном обучении; глубоком обучении.

1.3.1 Методы снятия омонимии, основанные на правилах

Для снятия омонимии могут применятся методы, основанные на правилах. С помощью экспертов определяют ряд зависимостей, позволяющие снимать отдельные случаи омонимии и формируется база правил, описывающая тонкости

языка, с опорой на лингвистические знания. Например, с помощью правил снимают разрешают омонимию для связанных пар слов. Если известно что в этих парах оба слова имеют одинаковое значение некоторой грамматической категории (например, часть речи или род), и одно из слов оказывается не омонимичным, тогда по известному значению категории для этого слова определяют значение для второго слова [48, 49].

В работе Зинькиной Ю. В., Пяткина Н. В. и Невзоровой О. А. предложено автоматическое разрешение функциональной омонимии в русском языке на основе метода контекстных правил. Для каждого типа функциональной омонимии разрабатывается обобщенное правило разрешения омонимии данного типа. Обобщенное правило представляет собой упорядоченную совокупность правил, записанных на специальном формальном языке. Каждое правило внутри совокупности фиксирует некоторый разрешающий контекст. Структура задает порядок применения правил, который базируется на оценке частотности контекстов. Данный метод базируется на синтаксических моделях. Он обладает более высокой точность, по сравнению с вероятностными методами [50].

Плюсами методов, основанных на правилах, является их простая интерпретируемость, небольшие требования к вычислительным ресурсам, высокая точность снятия омонимии и отсутствие необходимости для своего обучения размеченных текстовых корпусов.

Однако данные методы, ввиду сложности морфологии русского языка, для своей точной работы требуют ручной проверки базы правил специалистами на совместимость и непротиворечивость [51]. Такие методы не получили широкого распространения ввиду большой трудоемкости и больших затрат времени высококвалифицированных специалистов [52].

Существуют подходы к использующие правила, автоматически получаемые из корпуса текстов. Такие правила могут быть основаны на статистике встречаемости слов или синтаксических конструкций. В некоторых случаях применяют лексико-семантическая онтологическая базу знаний типа WordNet [53].

Методы снятия омонимии, основанные на правилах имеют ряд недостатков: сложность задания правил, в том числе для языков со свободным порядком слов,

необходимость многократного выполнения анализа для каждой омоформы. Поэтому чаще встречаются морфологические процессоры, основанные на статистических методах и машинном обучении.

1.3.2 Статистические методы и методы машинного обучения для снятия омонимии

Для работы статистических методов и методов машинного обучения требуются размеченные корпуса со снятой омонимией большого объема [54]. Ручная морфологическая разметка объемного корпуса весьма сложный и долгий процесс. Часто для разметки текстов используют специальные автоматические разметчики (например, Pymorphy2, Mystem). Данные морфологические анализаторы, как правило, приписывают слову не единственный разбор, а все теоретически возможные.

Определение статистики различных интерпретаций на основе корпуса – это самый простой метод устранения омонимии. В размеченном корпусе без омонимии происходит вычисление апостериорных вероятностей каждого из вариантов разбора по следующей формуле 1.1:

$$P(w|t) = \frac{Fr(w,t) + 1}{Fr(w) + |R(w)|} \quad (1.1)$$

В приведенной формуле, $Fr(w)$ – количество раз, которое словоформа w встретилась в корпусе, а $Fr(w, t)$ – количество раз, которое эта словоформа встретилось с тегом t. $|R(w)|$ – число разборов, полученных от анализатора для словоформы w [49, с. 54].

Методы, учитывающие контекст слова, обладают более высокой точностью снятия омонимии. Модель, анализирующая слово и n его соседей, называется n-грамм. Оптимальный результат показывает триграммная модель. Она обладает более высокими показателями точности снятия омонимии, чем уни- и биграммные модели, но в то же время не требует столько места как четырехграммная модель [55].

«В рассмотренных вариантах происходит попытка предсказать текущее слово по предыдущим. Однако на практике может оказаться, что текущее слово связано с несколькими последующими. Таким образом, необходимо выделять не просто слово по его контексту, а найти максимум вероятности для всего предложения. Из-за этого, скорость работы метода снятия омонимии растёт по экспоненте от длины предложения, а само снятие омонимии сводится к задаче поиска оптимального решения [49, с.73]».

Для языков с жестким порядком слов в предложении разрешение морфологической омонимии сводится к проблеме определения части речи слова (POS-теггинг). Для этого используются алгоритмы, основанные на статистических моделях, учитывающих вероятность появления тега той или иной части речи в данном контексте. Например, для английского языка подобные методы работают с точностью не менее 96 %.

«Для русского языка в виду свободного порядка слов в предложении точность таких алгоритмов намного меньше. Во-первых, морфологическая омонимия в русском языке, не сводится к омонимии частей речи, а охватывает множество различных грамматических признаков. Во-вторых, в английском языке порядок слов фиксированный. Это позволяет опираться только на локальный контекст слова (соседние слова) без учета дальних зависимостей. Поэтому для морфологической дизамбигуации в английском языке можно успешно использовать алгоритмы, основанные на Марковских моделях и учитывающие зависимость каждого набора тегов только от одного элемента контекста – непосредственно предшествующего ему набора тегов. В русском языке количество возможных контекстов из-за этого увеличивается и эффективность обучения простой модели, основанной на локальных зависимостях, снижается. Поэтому, наряду с Марковскими моделями, для снятия морфологической омонимии в русском языке используются более сложные статистические модели или гибридные системы, в которых статистика дополняется набором правил [50, с. 199]».

Помимо статистических методов, для снятия омонимии используются методы машинного обучения.

«Снятие омонимии без учета контекста в морфологическом анализаторе MyStem от Яндекс происходит благодаря обучению наивного баесовского классификатора на размеченном корпусе со снятой омонимией. Вероятность принадлежности неизвестного слова $word$, имеющего основу $stem$ и окончание $flex$, к парадигме $para$ рассчитывается по формуле 1.2:

$$P(para|word) = \frac{P(word|para)\cdot P(para)}{P(word)} = \frac{P(stem|para)\cdot P(flex|para)\cdot P(para)}{P(word)} \qquad (1.2)$$

При этом предполагается, что $stem$ и $flex$ являются независимыми случайными величинами [49, с. 56]». А правила определения парадигмы слова основаны на словаре Зализняка [56].

В морфоанализаторе Pymorphy2 реализовано бесконтекстное снятие омонимии на основе статистики словоупотреблений в размеченном текстовом корпусе Open Corpora. Данный процессор может анализировать несловарные слова на основе предсказателя. Морфоанализатор Pymorphy2 возвращает все допустимые варианты разбора. Для снятия омонимии у каждого разбора есть параметр score - оценка вероятности того, что данный разбор правильный.

Условная вероятность score оценивается на основе корпуса OpenCorpora: ищутся все неоднозначные слова со снятой неоднозначностью, для каждого слова считается, сколько раз ему был сопоставлен данный тег, и на основе этих частот вычисляется условная вероятность тега (с использованием сглаживания Лапласа). Для тех слов, для которых такой оценки нет, вероятность score либо считается равномерной (для словарных слов), либо оценивается на основе эмпирических правил (для несловарных слов). Разбор, выбранный на основе оценки score, верен примерно в 79% случаев.

Оценки score помогают улучшить разбор, но их недостаточно для надежного снятия неоднозначности по следующим причинам:

– pymorphy2 работает только на уровне отдельных слов и не учитывает контекст;

– условная вероятность score оценена на основе сбалансированного набора текстов; в специализированных текстах вероятности могут быть другими.

Примером использования методов машинного обучения является работа Хельмута Шмида, в которой описывает метод морфологического анализа и снятия омонимии с использованием дерева решений. Метод использует бинарное дерево решений, которое состоит из корневых узлов, задающих вопросы о контексте вокруг слова, и листьев, взвешенных слов с определёнными вероятностями для каждой части речи. TreeTagger не требует большого размеченного корпуса для обучения, т.к. использует готовые словари и деревья решений для отслеживания контекста слов. Метод показывает большую скорость работы и точность разметки 84% [57].

Одним из популярных методов снятия омонимии, основанных на обучении по претендентам, является метод опорных векторов (SVM) при котором каждую омоформу необходимо отнести к тому или иному классу. Для снятия омонимии используется такое число классов, которое соответствует грамматическим характеристикам омонимичных слов. В работе А. А. Порохнина показано, что скрытая марковская модель для снятия омонимии в текстах на русском языке работает лучше метода опорных векторов, в отличии от английского языка [58].

Статистические методы и методы машинного обучения для снятия омонимии демонстрируют довольно высокую точность определения морфологических характеристик, которая вычисляется как отношение количества словоформ с верно определёнными характеристиками к общему количеству анализируемых словоформ. К недостаткам можно отнести необходимость формирования объемных размеченных текстовых корпусов со снятой омонимией, а также более низкие показатели точности анализа для флективных языков со свободным порядком слов, к которым относится русский язык.

Также существенным недостатком стандартных подходов МА является то, что они не рассматривают словосочетания как единую словарную единицу, в то время как в русском языке часто встречаются неделимые словосочетания, которые могут быть наречием или предикативом. Например, в охотку, на выданье и т. д.

При автоматическом снятии омонимии предикативов осложняющим фактором является недостаток размеченных корпусов со снятой омонимией. Так в самом авторитетном из корпусов русского языка, Национальном корпусе русского языка (НКРЯ), к подкорпусу с омонимией, снятой вручную, относится только 1% от общего числа вхождений, а в подкорпусе с автоматически снятой омонимией встречается невероятно количество ошибок при разметке.

1.4 Нейросетевые языковые модели

На сегодняшний день развитие глубоких нейронных сетей и языковых моделей (ЯМ) на их основе эффективно применяется для решения большинства NLP-задач. ЯМ способны не только анализировать текст, но и генерировать новый, а также классифицировать текст по различным параметрам.

«Задачей языкового моделирования является определение вероятности последовательности слов $w = (w_1, w_2, \ldots, w_m)$. Методики глубокого обучения и применение рекуррентных нейросетей для обработки текстов существенно улучшают качество ЯМ за счет учета контекста и отсутствия ограничений на использование только *n* предыдущих слов.» [59, с.26]

В рекуррентных нейронных сетях (RNN) связи между элементами образуют направленную последовательность. Благодаря этому появляется возможность обрабатывать серии событий во времени или последовательные пространственные цепочки. «RNN создаются с помощью добавления «блоков памяти», что обеспечивает учёт состояний предыдущих циклов (вычисленных значений нейронов) в последующих циклах в любых позициях.» [59, с.27] Обратная связь даёт RNN способность запоминать информацию, что позволяет моделировать динамические процессы. Скрытый слой RNN хранит всю предыдущую историю, поэтому размер контекста не ограничен. RNN превосходят стандартные методы, за исключением их высокой вычислительной сложности при обучении [60].

Обучение хранению информации в течение длительных интервалов времени с помощью периодического обратного распространения занимает очень много времени, главным образом из-за недостаточного, затухающего обратного потока ошибок. В работе S. Hochreiter и J. Schmidhuber был предложен новый эффективный метод, основанный на градиенте, который называется долговременной кратковременной памятью (LSTM). С помощью сокращения градиента LSTM может научиться преодолевать минимальные временные задержки, обеспечивая постоянный поток ошибок с помощью каруселей постоянных ошибок в специальных блоках. Мультипликативные логические модули учатся открывать и закрывать доступ к постоянному потоку ошибок. По сравнению с рекуррентным обучением в реальном времени, обратным распространением во времени, рекуррентной каскадной корреляцией, сетями Элмана и разделением нейронных последовательностей на фрагменты, LSTM обеспечивает гораздо более успешные запуски и обучается намного быстрее. LSTM также решает сложные задачи с искусственным длительным запаздыванием, которые никогда не решались предыдущими рекуррентными сетевыми алгоритмами [61].

«Глубокое обучение и использование рекуррентных нейронных сетей для анализа текста значительно улучшают качество языковой модели благодаря учёту контекста и отсутствию ограничений на использование только предыдущих слов. Развитие нейросетевого подхода в области NLP привело к появлению моделей Sequence-to-sequence (seq2seq). Данная модель принимает на вход последовательность элементов (слов, символов, изображений и т. д.) и возвращает другую последовательность элементов [59, с.27-28]». Этот подход был предложен в работе I. Lourentzou и др. для нормализации текста в социальных сетях.

Базовая модель seq2seq состоит из энкодера и декодера. Энкодер преобразует входные данные в вектор, а декодер использует его для генерации выходных данных. Размер контекстного вектора задаётся при обучении модели и определяет количество скрытых нейронов в кодере RNN. Однако данная модель

seq2seq не могла обрабатывать длинные предложения из-за сложностей с контекстным вектором [62].

Следующим шагом стала новая архитектура, получившая название Transformer, которая в отличие от начальных моделей seq2seq, не использует RNN, в качестве стандартных архитектур для енкодера и декодера Transformer использует полносвязные слои, а также и механизм многослойного обучающего внимания (multi-head attention) в енкодере. Multi-head attention – новый слой, который дает возможность каждому входному вектору взаимодействовать с другими через механизм внимания, вместо передачи скрытого состояния как в RNN. Модель на порядок быстрее обучается и показывает более высокое качество машинного перевода, чем seq2seq с использованием RNN со слоем внимания.

Основные компоненты кодировщика и декодера включают механизмы внутреннего внимания и прямые слои распространения (Feed Forward Layers). Входные и выходные данные (целевые предложения) сначала преобразуются в n-мерное пространство. Важной особенностью модели является использование позиционного кодирования для разных слов. Поскольку модель не использует рекуррентные сети, которые могут запоминать последовательности слов, необходимо присвоить каждому слову относительное положение, так как порядок элементов влияет на последовательность. Эти позиции добавляются к встроенным представлениям (n-мерным векторам) каждого слова. Внутреннее внимание можно описать как функцию, которая отображает вход и набор пар «ключ-значение» на выход, где запрос, ключи, значения и выходные данные являются векторами. Выходные данные вычисляются как взвешенная сумма значений, где вес, присвоенный каждому значению, определяется функцией совместимости запроса с соответствующим ключом [63].

Наиболее популярной моделью при решении задачи обработки естественного языка является BERT [64]. Для решения задач, связанных с преобразованием одних текстов в другие, используют языковые модели T5 [65], которые также основаны на архитектуре Transformer. Удачный опыт применения

архитектуры Transformer для решения задач обработки текстов на русском языке показали модели T5 [66], RuGPT3 [67] и Mbart [68], что описано в работе [65, с. 230].

Перечисленные модели имеют значительные отличия. Так, в модели BERT используются только энкодеры, можно работать с текстом в прямом и обратном направлениях, с методом обучения и некоторыми другими небольшими деталями. Модели GPT [69], наоборот, используют стек декодеров, а также другой тип самовнимания. В модели T5 входящая и выходящая информация является строками, ее архитектура отличается изменением слоя нормализации, в котором выполнено удаление смещения слоя, а также размещения слоя за пределами остаточного пути. Кроме того, в T5 применяется другая схема кодирования позиции токенов.

Васильевым Д. Д. и Пятаевой А. В. проведены исследования применимости ЯМ на базе Transformer к задаче упрощения текста. Качество работы модели оценивалось с использованием интегральной характеристики с учетом специфических метрик оценки текстов, а также экспертами-лингвистами. В результате исследования был сделан вывод о недостаточном качестве работы алгоритма автоматической симплификации. Самой вероятной причиной этого является недостаточное количество обучающих данных [65, с. 232].

В 2021 г. на Международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог-21» проводилось соревнование RuSimpleSentEval, посвященное упрощению текстов на русском языке [70]. Специально для этого соревнования был подготовлен набор данных, насчитывающий около 3 тысяч сложных предложений, собранных из веб-энциклопедии Википедия. Их упрощения подготовлены работниками краудсорсинговой платформы Яндекс.Толока. Предполагалось решение задачи упрощения по аналогии с задачей машинного перевода: на входе – сложное предложение, на выходе его упрощенная версия. В качестве показателя успешности системы использовалась широко распространенная мера SARI. В соревновании приняли участие 14 команд. Суммарно было получено 350 вариантов решений, использующих различные языковые модели (mBART, GPT).

Полученные средние баллы SARI соответствуют ожиданиям и близки к результатам, для других языков. Однако, несмотря на высокие значения автоматической меры качества, тексты, генерируемые моделью, иногда полностью или частично теряли первоначальный смысл [71, 72, 73].

В работе [74] Васильевым Д. Д. и Пятаевой А. В. проведены исследования применимости больших языковых моделей к задаче упрощения текста. В результате был сделан вывод о недостаточном качестве работы модели по причине недостаточного количества обучающих данных.

Таким образом, основными недостатками больших языковых моделей применительно к лексической адаптации текста можно считать сложность их обучения и тонкой настройки, а также зачастую полную либо частичную утрату первоначального смысла в упрощенном тексте [75].

1.5 Выводы к разделу 1

Обзор технологий и методов компьютерной обработки текстовой информации, таких как автоматизированное упрощения текста и его обработки на морфологическом уровне позволил сделать следующие выводы.

1. Задачу лексического упрощения можно решить путём замены сложных слов на более простые синонимы. Часто при этом могут возникнуть ошибки, связанные с как с семантикой и стилем, так и с синтаксисом, поскольку заменяемое слово и его синоним могут принадлежать к разным грамматическим категориям. Это приводит к тому, что большинство доступных программных продуктов «синонимайзеров» можно применять, только используя эксперта-человека для определения выбора подходящего синонима и соблюдения правил синтаксиса. В связи с этим проблема правильного подбора синонимических замен является по прежнему весьма актуальной.

2. Анализ существующих решений для автоматического упрощения показал что оптимальные решения пока найдены далеко не для всех исследовательских задач в этой области, а существующие методы применимы к

сравнительно небольшому числу языков, что обуславливает актуальность и практическую ценность данной работы

3. Любой метод МА включает две части: декларативную, которая состоит из словарей, и процедурную, включающую алгоритмы. При этом морфологический словарь – важнейшая компонента морфологических процессоров. Эффективной структурой данных для представления морфологического словаря являются префиксные деревья, где значения МИ, хранимые в словаре, находятся сразу после словоформ.

4. Для решения проблемы омонимии существует несколько подходов, которые основаны на правилах; статистике и машинном обучении; глубоком обучении. Методы, использующие правила, разрешают омонимию в определённых случаях в зависимости от контекста и самих омонимичных слов. Плюсами этих методов являются их простая интерпретируемость, небольшие требования к вычислительным ресурсам, а также то, что для своего обучения они не требуют размеченных текстовых корпусов большого объема, в отличие от методов машинного обучения и статистических. Однако, ввиду сложности морфологии русского языка, для своей точной работы методам этого подхода необходима обширная база правил.

5. Современные ЯМ, представляющие собой глубокие нейросети на базе архитектуры Transformer, эффективно решают задачу снятия омонимии за счет учета контекста словоформы в предложении и улучшенного моделирования дальних зависимостей слов, но имеют огромное количество параметров, доходящее до нескольких миллиардов, что вызывает сложность в обучении.

6. В дальнейшем диссертационном исследовании представляется целесообразным 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 мифологического словаря.

РАЗДЕЛ 2
РАЗРАБОТКА АЛГОРИТМА ОПРЕДЕЛЕНИЯ МОРФОЛОГИЧЕСКИХ ПАРАМЕТРОВ РУССКОЯЗЫЧНЫХ ТЕКСТОВ

В данном разделе приведено описание основных структур сформированного тестового корпуса - словарь словоформ для лемматизации, его представление в виде префиксного дерева и алгоритм индексирования строк данного словаря. Также описана общая схема работы алгоритма лемматизации, определения морфологических параметров, что необходимо для дальнейшей морфологической обработки, связанной с разрешением омонимии на основе базы правил, которая использует словарь русских словоформ.

2.1 Описание основных структур словаря русских словоформ

Определение грамматических параметров и лемматизация составляющих предложение словоформ осуществляется путем поиска их в текстовом файле словаря русских словоформ М. А. Хагена «Полная парадигма. Морфология» [76]. Словарь содержит более 4 миллионов словоформ для более чем 130 тыс. лемм. Этот словарь организован как множество строк, объединенных в блоки, каждый из которых начинается леммой и образует полную парадигму слова [77]. Порядок лемм – алфавитный. Пример такого блока:

ехать | гл несов непер инф
ехать | гл несов непер инф
едемте | гл несов непер пов мн
ехал | гл несов непер прош ед муж
ехала | гл несов непер прош ед жен
ехало | гл несов непер прош ед ср (2.1)
ехали | гл несов непер прош мн
…
едучи | деепр несов непер наст

Вслед за разделительным знаком «|» приведена морфологическая информация о словоформе. Она выражается аббревиатурами и сокращениями, из которых пояснения требует лишь запись «2вид», относящаяся к глаголу. Она означает совпадение по форме глаголов совершенного и несовершенного видов. Следует отметить, что из словаря удален ряд деепричастных форм, не используемых в современном языке, а также несколько пополнили словарь.

В работе использовано представление множества всех словоформ этого словаря в виде префиксного дерева [78]. Это позволяет, несмотря на сверхбольшой объем словаря, почти мгновенно осуществлять в нем поиск всех словоформ, соответствующих заданной последовательности символов, и получать результат в виде последовательности строк вида:

ложка | сущ неод ед жен им (2.2)

ложка | сущ неод ед муж род.

В рамках данной работы такая последовательность называется группой.

Кроме того, поскольку в данной работе приходится систематически пользоваться лемматизацией, то разработан вариант словаря, где в каждой строке после словоформы добавлена ее лемма. Например, для начальной части блока (2.1) имеем следующую последовательность строк:

ехать | ехать | гл несов непер инф

едемте |ехать | гл несов непер пов мн

ехал |ехать | гл несов непер прош ед муж

ехала |ехать | гл несов непер прош ед жен

ехало |ехать | гл несов непер прош ед ср

ехали |ехать | гл несов непер прош мн

…….. [77, с. 39]

Ясно, что использование этого словаря сводит лемматизацию словоформы к поиску этой словоформы в словаре, и в результате лемматизация происходит с той же скоростью, что и поиск словоформы.

«Леммой для причастия договоримся считать не инфинитив соответствующего глагола, а форму этого причастия в именительном падеже единственного числа мужского рода. Тогда леммой для слова «искавшую» является слово «искавший», а леммой для слова «ищущим» является слово

«ищущий». В этом смысле причастия трактуются как самостоятельные части речи – аналоги прилагательных. Это достаточно естественно ввиду обилия у них словоформ и наличия процедуры склонения. Леммой же для деепричастия служит инфинитив глагола [79]».

Отметим также следующее. Парадигма прилагательного включает формы его превосходной степени. Например,

быстрый | прл ед муж им

быстрого | прл ед муж род

……

быстрейший | прл прев ед муж им

быстрейшего | прл прев ед муж род

……

наибыстрейший | прл прев ед муж им

наибыстрейшего | прл прев ед муж род

«В рамках данного исследования леммой для какой-либо словоформы превосходной степени прилагательного форму превосходной степени именительного падежа, мужского рода, единственного числа. Она определяется путем движения по списку от лемматизируемой словоформы в направлении снизу вверх [79, с. 79]».

В результате для каждого слова из анализируемого текста создается структура, определяющая его лемму и грамматические характеристики. Ее описание приведено в таблице 2.1. Если какие-то из характеристик не определены в словаре для данного слова (например, падеж для глагола или время для существительного) соответствующие поля остаются пустыми.

С помощью данного словаря можно выполнять не только морфологический анализ, но и синтез словоформы с заданными морфологическими характеристиками. Он сводится сначала к поиску в словаре леммы, а затем формы, которая входит в ту же парадигму, что и лемма, и имеет требуемые грамматические характеристики.

Таблица 2.1. – Структура, определяющая грамматические характеристики словоформы

Поле	Возможные значения	Описание
Лемма		Начальная форма слова
Часть речи	сущ, мест сущ, прл, мест прил, гл, прч, дееп, нар, мест нар, предик, союз, союзн сл, част, межд, (предл гр), (отриц),(соч)	Часть речи или категория словосочетания (предложная группа, отрицание, сочетание слов)
Падеж	им, род, дат, вин, тв, пр, счет, зват	Падеж указывается для существительных, местоимений-существительных, прилагательных, местоимений-прилагательных, причастий и числительных. Для предлогов указан падеж, который имеет управляемое слово.
Одушевленность	одуш, неод	Одушевленность указывается для существительных, местоимений-существительных, прилагательных, местоимений-прилагательных, причастий и числительных.
Степень сравнения	сравн, превосх	Степень сравнения для наречий и прилагательных.
Вид числительного	кол, поряд, собир	Вид числительного – количественное, порядковое или собирательное.
Переходность глагола	перех, непер, пер/не	Определяет, является ли глагол переходным, непереходным или переходным/непереходным
Форма глагола	инф, пов	Неопределенная форма или глагол повелительного наклонения
Возвратный глагол	Воз	Указывает, является ли глагол возвратным
Вид глагола	сов, несов, 2вид	Указывает вид глагола – совершенный, несовершенный, двувидовый
Время	наст, прош, буд	Время для глаголов, деепричастий и причастий
Род	муж, жен, ср, общ	Указывает род. Сокращение «общ» обозначает общий род (мужской и женский).
Число	ед, мн	Указывает число
Лицо	1-е, 2-е, 3-е	Указывает лицо для глаголов
Залог	Страд	Указывает страдательный залог для причастий
Краткость	Крат	Указывает краткую форму прилагательного или причастия

2.2. Представление множества словоформ в виде префиксного дерева

Для обеспечения быстрого поиска в данном словаре используется представление его в виде префиксного дерева, состоящего из множества узлов. В каждом узле дерева хранится метка – один из символов алфавита $A = \{a_1, a_2, \ldots, a_d\}$. Ключом, который соответствует некоторому узлу, является путь от корня дерева до узла, а точнее строка $c_1c_2\ldots c_m$ составленная из меток узлов, повстречавшихся на этом пути. Если соответствующая строка есть в словаре, то с узлом ассоциируется индекс, являющийся ее порядковыми номером в словаре. В словаре может быть несколько словоформ-омонимов с одинаковым написанием, но отличающихся по смыслу или морфологической информацией. Поэтому вместо одного упомянутого индекса может присутствовать некоторый список таких индексов. Наличие одного или нескольких индексов указывает на то, что узел является конечным, то есть соответствует концу некоторого слова. В противном случае узел является лишь промежуточным по дороге в какой-либо другой, который является конечным. Корень дерева, очевидно, соответствует пустому ключу. Для каждого узла также известно множество дочерних узлов следующего нижнего уровня (смежные узлы – потомки). Количество их может меняться от *0* до *d* по числу символов алфавита.

Хранение дерева в памяти осуществляется с помощью списка всех его узлов и списка номеров смежных узлов-потомков для каждого из узлов. Для узлов, не имеющих ни одного потомка, список смежных узлов пуст. Если в дереве имеется *L* узлов, пронумерованных 0,…,*L-1*, то в памяти будет храниться *L* списков потомков, собранных в главный список.

0: {1, 5, 9, 15}

1: {2, 4}

2: {3}

…

L-1: {}

Порядковый номер каждого списка смежных узлов соответствует номеру данного узла в списке.

В списке узлов хранятся данные для каждого узла, состоящие из двух полей – символа алфавита и списка индексов строк (для промежуточных узлов последний список будет пустым).

0: [' ',{}]

1: ['а',{0,1,2}]

2: ['б',{}]

3: ['а',{3}]

4: ['ж',{}]

…….

Создание такого дерева для большого словаря может занять достаточно длительное время, поэтому целесообразно сохранить полученную структуру в файле, чтобы можно было быстро получить к ней доступ при следующих сеансах работы. В начале файла записывается количество узлов *L*, затем для каждого узла последовательно записывается количество потомков и массив индексов узлов-потомков в двоичном виде. После этого записывается массив данных узлов. Загрузка такого файла значительно менее затратна по времени, чем создание дерева «с нуля» [80].

Файл данных имеет имя, совпадающее с именем файла-словаря и расширение «.dat». Этот файл создается один раз, и в дальнейшем используется только для чтения. Заново создавать его необходимо только в том случае, если нужно перестроить дерево после редактирования текстового файла [81].

Алгоритм поиска омонимов с использованием дерева можно описать следующим образом. Пусть есть слово или словосочетание, для которого необходимо найти все возможные совпадающие по написанию формы (омонимы) в словаре, представленном в виде дерева. Будем спускаться из корня дерева на нижние уровни, каждый раз переходя в узел, чей символ совпадает с очередным символом строки. После того как обработаны все символы строки, узел, в котором остановился спуск и будет искомым узлом. Индекс, ассоциированный с этим узлом, есть номер строки в словаре. Если в процессе спуска не нашлось узла с символом, соответствующим очередному символу строки, или спуск остановился на промежуточной вершине (с пустым списком индексов), то искомый ключ

отсутствует в дереве, а строка – в словаре. Другими словами, считается, что слово есть в словаре, если можно построить путь от вершины, в котором содержатся все символы данного слова в правильном порядке, и в последнем узле которого есть не пустой список индексов.

Таким образом, для нахождения всех омонимов заданного слова нужно произвести поиск всех совпадающих по написанию словоформ по вышеописанному алгоритму. Если список индексов для данной словоформы содержит более одного индекса (т.е. таких словоформ в словаре несколько), то она имеет омонимы. [77, с. 44]

2.3 Алгоритм индексирования строк морфологического словаря

Так как текстовый файл, содержащий морфологический словарь, имеет значительный объем – более 4 млн. строк-записей, его полное чтение в память представляется нецелесообразным. В то же время, все строки имеют переменную длину, поэтому непосредственное позиционирование в файле на строку с нужным индексом невозможно без считывания всех предыдущих строк. Учитывая, что данный файл используется только для чтения информации и не модифицируется программой, можно обеспечить быстрый доступ к его записям с помощью статического массива файловых указателей, который можно хранить в отдельном двоичном файле. Указатель представляет собой 32-разрядное целое число, определяющее смещение (в байтах) начала строки относительно начала файла, посредством которого можно сразу перейти к считыванию строки с заданным индексом без необходимости считывать все предыдущие. Создание файла указателей происходит при однократном построчном чтении файла словаря по следующему алгоритму:

1) Запись 0 в файл указателей (смещение, указывающее на начало первой строки).
2) Если не конец файла, чтение очередной *i*-й строки, иначе переход к шагу 4.
3) Запись текущего смещения в файл указателей. Переход к шагу 2.
4) Конец алгоритма.

Таким образом, содержимое файла указателей представляет собой статический целочисленный массив размера *N*, где *N* – количество строк в текстовом файле. С его помощью можно легко установить соответствие между порядковым номером строки и смещением, по которому эта строка находится в текстовом файле.

Таблица 2.2 – Пример индексирования строк текстового файла

Индекс строки	Указатель	Текстовый файл
0	0	абажур | сущ неод ед муж им\n
1	28	абажура | сущ неод ед муж род\n
2	58	абажуру | сущ неод ед муж дат\n
3	88	абажур | сущ неод ед муж вин\n
4	119	абажуром | сущ неод ед муж тв\n
…	…	…

Чтобы считать из текстового фала строку с заданным индексом *i,* вначале из файла указателей считывается указатель с данным индексом, затем внутри текстового файла осуществляется переход по этому указателю и чтение строки в память (от заданного смещения до символа конца строки «\n»). При этом не требуется полностью считывать в память файл указателей и файл словаря, что существенно экономит время. Файл указателей имеет имя, совпадающее с именем текстового файла и расширение «.ind_txt». Он создается однократно, и в дальнейшем используется только для чтения. Заново создавать его необходимо только в том случае, если в текстовый файл внесены какие-либо изменения.

2.4 Основные этапы работы алгоритма лемматизации и определения морфологических параметров русскоязычных текстов

Входными данными является строка символов, содержащая анализируемый текст на русском языке. Текст может вводиться с клавиатуры или загружаться из текстового файла.

Блок-схема алгоритма функционирования метода снятия омонимии представлена на рисунке 2.1. «На первом этапе обработки текста производится его сегментация, т. е. выделение в тексте предложений и словоформ, точнее токенов (т. к. в тексте могут быть не только слова, но и знаки препинания, сочетания слов и т.д.). В результате исходный текст преобразуется в цепочку токенов. Это позволяет получить текст, пригодный для корректного проведения морфологического анализа: получения начальной формы заданного слова (леммы), а также морфологических параметров [79, с.78]».

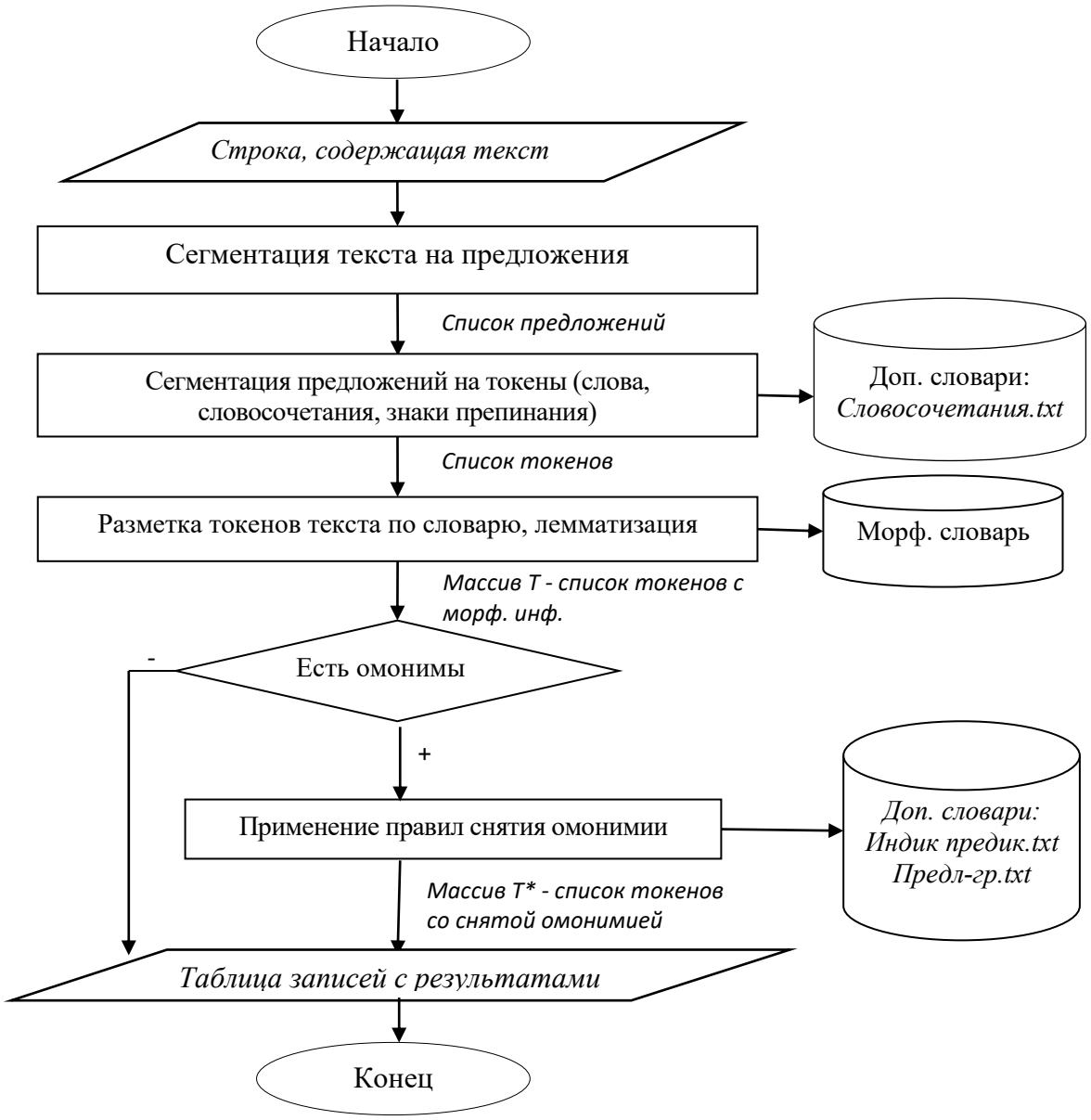

Рисунок 2.1 – Блок-схема алгоритма функционирования метода снятия омонимии

Токен – последовательность символов некоторого алфавита. В рамках данной работы используются следующие алфавиты: алфавит русских символов; алфавит знаков препинания.

Далее для каждого токена, который является словом или словосочетанием, производится поиск в словаре всех возможных вариантов разбора и соответствующих грамматических параметров, а также определение начальной формы для каждого из вариантов. К полученному списку токенов с грамматической информацией для каждого предложения текста применяются правила снятия неоднозначности (описаны в разделе 3), с помощью которых из нескольких вариантов разбора слова выбирается один. В результате выбранные варианты разбора помечаются специальной пометкой «!».

Рассмотрим более подробно некоторые этапы.

1) На входе строка T, содержащая текст.

Сегментация строки T на предложения. Разделители – знаки препинания «.!?». Результат – список предложений S размера N_s, приведен в таблице 2.3.

Таблица 2.3 Результат сегментации строки T на предложения

S[0]	S[1]	S[2]	…	$S[N_s-1]$

2) Сегментация предложений на токены (слова, словосочетания, знаки препинания). Получение списка токенов для каждого предложения.

Таблица 2.4 – Сегментация предложений на токены

Предложение	Список токенов			
S[0]:	T[0]	T[1]	…	$T[N_{s0} - 1]$
…	…	…	…	…
$S[N_s]$	T[0]	T[1]	…	$T[N_{sN} - 1]$

Сегментация производится разрезанием каждой из строк S[i] по пробелам, за исключением некоторых заранее определенных словосочетаний, которые не разделяются, а воспринимаются как единая морфологическая единица. Список данных словосочетаний приведен в дополнительном файле *Словосочетания.txt*.

3) Разметка каждого токена текста по словарю и определение начальных форм. Разметка представляет собой одну или несколько (если слово имеет омонимы) записей, состоящую из стороковых полей, представленных в таблице 2.5.

Таблица 2.5 – Структура, определяющая грамматические характеристики словоформы

Поле	Идентификатор	Возможные значения
Лемма	lem	начальная форма слова (лемма)
Часть речи	pos	сущ, мест сущ, прл, мест прил, гл, прч, деепр, нар, мест нар, предик, союз, союзн сл, част, межд, (предл гр), (отриц),(соч)
Падеж	case	им, род, дат, вин, тв, пр, счет, зват
Одушевленность	anim	одуш, неод
Степень сравнения	degree	сравн, превосх
Вид числительного	num_form	кол, поряд, собир
Переходность глагола	verb_trans	перех, непер, пер/не
Форма глагола	verb_form	инф, пов
Возвратный глагол	verb_refl	Воз
Вид глагола	verb_type	сов, несов, 2вид
Время	tense	наст, прош, буд
Род	gender	муж, жен, ср, общ
Число	number	ед, мн
Лицо	person	1-е, 2-е, 3-е
Залог	voice	Страд
Краткость	short	Крат

Выходными данными работы алгоритма лемматизации и определения морфологических параметров русскоязычных текстов является массив *T*, каждый элемент которого является массивом записей с грамматической информацией и леммой, пример приведен в таблице 2.6. Данный массив являются входными данными для метода снятия омонимии, описанного в разделе 3.

Таблица 2.6 – Результат грамматической разметки для предложения

T[0]:	0	lem	pos	case	anim	degr	numform	…
	1	lem	pos	case	anim	degr	numform	…
	…	…						
…								
T[1]:	0	lem	pos	case	anim	degr	numform	…
	1	lem	pos	case	anim	degr	numform	…
	…	…						
…								

2.5 Выводы к разделу 2

В разделе 2 решена задача разработки алгоритма определения морфологических параметров словоформ и лемматизации:

1) На базе словаря М.А. Хагена сформирован словарь словоформ русского языка, далее обозначаемый МС, содержащий более 4 млн. словоформ для более 130 тыс. лемм. В каждой строке словаря хранится словоформа, МИ и ее лемма.

2) Осуществлено представление множества всех словоформ морфологического словаря в виде префиксного дерева. Для поиска всех вхождений словоформы в словарь используется алгоритм индексирования его строк, облегчающий поиск. Это позволяет, несмотря на сверхбольшой объем словаря, почти мгновенно осуществлять в нем поиск всех словоформ, соответствующих заданной последовательности символов.

3) В каждую строку МС добавлена лемма. Это сводит лемматизацию к поиску всех вхождений словоформы в словарь, в результате чего лемматизация происходит за один проход с той же скоростью, что и поиск словоформы.

РАЗДЕЛ 3
РАЗРАБОТКА МЕТОДОВ И АЛГОРИТМОВ СНЯТИЯ ОМОНИМИИ В РУССКОЯЗЫЧНЫХ ТЕКСТАХ

Существенным недостатком существующих подходов МА, рассмотренных в разделе 1, является то что они не рассматривают словосочетания как единую словарную единицу, в то время как в русском языке часто встречаются неделимые словосочетания, выполняющие роль наречия или предикатива. Еще одной проблемой, связанной с автоматическим разрешением частеречной омонимии предикативов, является недостаток размеченных корпусов со снятой омонимией, о чем свидетельствует тот факт, что в самом авторитетном из корпусов русского языка, НКРЯ, к подкорпусу с вручную снятой омонимией относится только 1% от общего числа вхождений. Кроме того, декларативные и процедурные методы имеют низкую точность разрешения омонимии среди деепричастий и наречий. Это приводит к тому, что в подкорпусе с автоматически снятой омонимией встречается большое количество ошибок.

В связи с вышесказанным, в ходе выполнения данного диссертационного исследования внимание должно быть сконцентрировано на повышении точности методов снятия частеречной омонимии для предикативов и предикативных словосочетаний, деепричастий, а также групп наречие-существительное.

В данном разделе решается задача разработки алгоритмов снятия некоторых случаев частеречной омонимии. Приведено описание методов автоматического разрешения омонимии на основе гибридного подхода, использующего как декларативные знания в виде словарей, так и базу продукционных правил, что позволяет снять частеречную омонимию предикативов и предикативных словосочетаний, деепричастий, групп наречие-существительное. В разделе приведено сравнение эффективности предложенного метода снятия омонимии с существующими решениями.

3.1 Основные этапы работы метода снятия омонимии

На предварительном этапе обработки текста, описанном в подразделе 2.4, производится сегментация и токенизация текста. В результате исходный текст преобразуется в цепочку токенов, для каждого из которых производится поиск в словаре всех возможных вариантов разбора: соответствующих грамматических параметров и его лемма (описано в разделе 2).

Входными данными метода снятия омонимии является результат грамматической разметки для предложения - массив *T*, каждый элемент которого является массивом записей с грамматической информацией, пример приведен в таблице 2.6. К полученному списку токенов для каждого предложения текста применяются правила снятия неоднозначности, описаны в п. 3.2-3.6, с помощью которых из нескольких вариантов разбора слова выбирается один. В результате выбранные варианты разбора помечаются специальной пометкой «!». Выходными данными является массив T*, содержащий список токенов со снятой омонимией (пометка «!»). В случае если метод снятия омонимии не может однозначно идентифицировать омоним массив *T** равен массиву *T*. Блок-схема алгоритма применения правил снятия омонимии представлена на рисунке 3.1.

На рисунке 3.1 переменная hom_type может принимать целочисленное от 1 до 8 и обозначает вид частеречной омонимии, а именно:

1. Омонимия предикатив-наречие-краткое прилагательное в случае единственного кандидата на предикатив (hom_type = 1).
2. Омонимия предикативных неделимых словосочетаний (hom_type = 2).
3. Омонимия словосочетаний с отрицанием (hom_type = 3).
4. Омонимия предикативных разделимых словосочетаний, не являющихся предложными группами (hom_type = 4).
5. Омонимия предикативных словосочетаний, которые могут быть предложными группами (homonym type = 5), для снятия которой наряду с правилами используется размеченный словарь словосочетаний Предл гр.txt. Этот словарь создан в результате работы с НКРЯ и содержит текстовые записи, предложных группами, их управляющих групп, а также разметки, определяющий, должно ли словосочетание интерпретироваться как единое целое, или как отдельные слова.

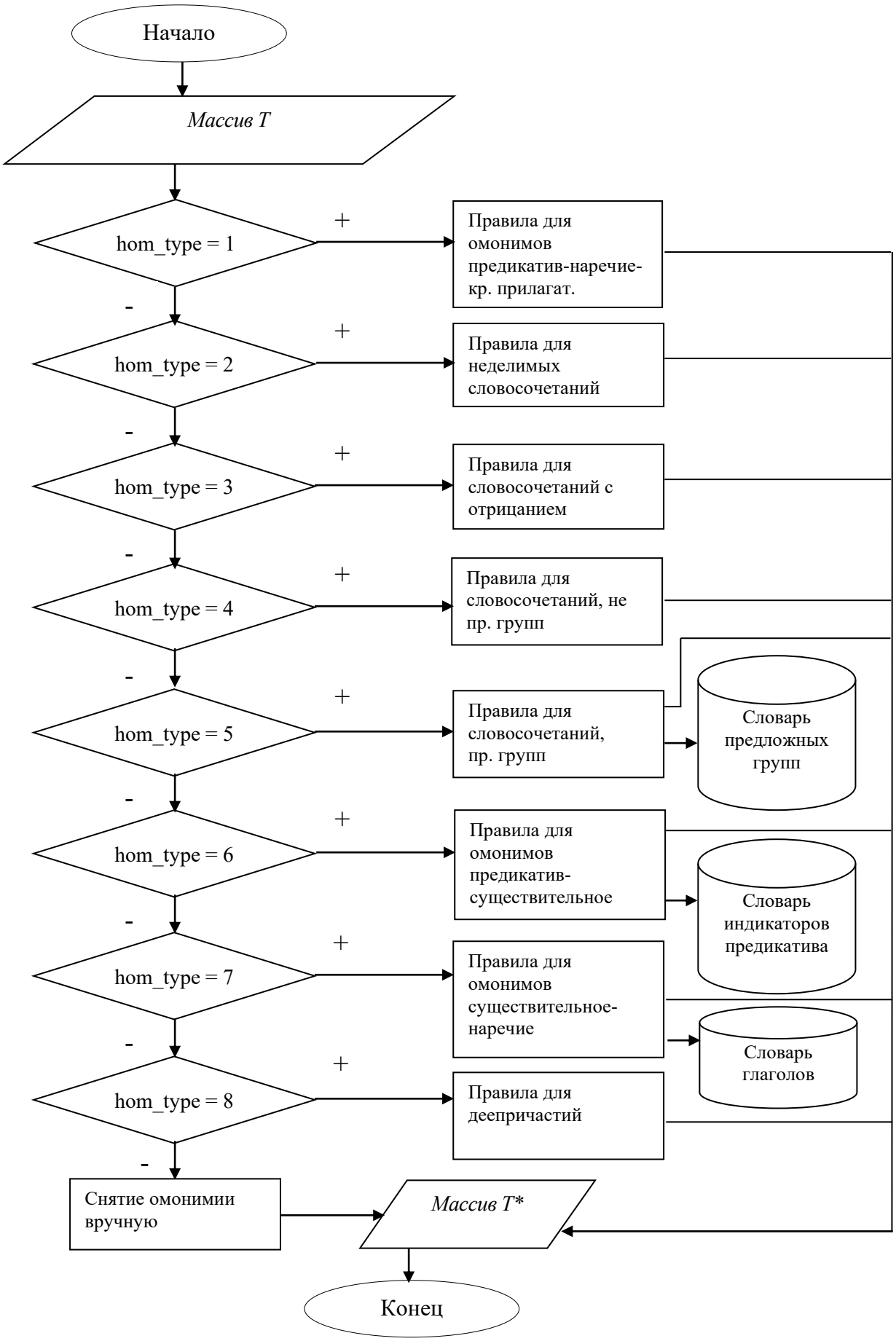

Рисунок 3.1 – Блок-схема алгоритма применения правил снятия омонимии

6. Омонимия предикатив-существительное (hom_type = 6), для снятия которой наряду с правилами используется словарь индикаторов предикатива, который представляет собой текстовый файл Индик предик.txt.

7. Омонимия наречие-существительное (hom_type = 7), для снятия которой наряду с правилами используется словарь глаголов Глаг-нар.txt, употребляемых с наречием или существительным, и список предлогов.

8. Омонимия деепричастий (hom_type = 8), для снятия которой используются перечни омонимов, сформированные с помощью разработанного морфологического словаря.

Предикатив – сравнительно недавно введенная в лингвистический обиход часть речи, связанная с функцией сказуемого в предложении. Наиболее частым в предложении является глагольное сказуемое. Вместе с тем сказуемое может выражаться другими частями речи: существительным, прилагательным (в частности, кратким прилагательным) и так далее. Позволим себе называть подобную реализацию сказуемого традиционной. Однако в последний перечень было бы неестественно включать, например, наречие, ибо наречие выражает дополнительную характеристику действия или качества, выраженного чаще всего глагольной формой или прилагательным. Поэтому для описания сказуемого в предложении типа «Мне холодно» была введена новая часть речи, которая обозначалась в русской лингвистике как «категория состояния». Позднее соответствующее понятие было расширено, а его название было заменено заимствованным у чешских лингвистов термином «предикатив» [82].

Существенным недостатком стандартных подходов МА является то что они не рассматривают словосочетания как единую словарную единицу, в то время как в русском языке часто встречаются неделимые словосочетания, которые могут быть наречием или предикативом. Например, в охотку, на выданье и т. д.

При разработке методов снятия частеречной омонимии решено сконцентрировать внимание на предикативах и предикативных словосочетаниях, деепричастиях, группах наречие-существительное, т. к. стандартные подходы

обеспечивают низкую точность разрешения омонимии среди данных групп омонимов.

При автоматическом снятии омонимии предикативов осложняющим фактором является недостаток размеченных корпусов со снятой омонимией. Так в самом авторитетном из корпусов русского языка, НКРЯ, к подкорпусу с вручную снятой омонимией относится только 1% от общего числа вхождений, а в подкорпусе с автоматически снятой омонимией встречается большое количество ошибок при разметке.

Разработанный метод снятия омонимии анализирует правильные русские предложения с правильно поставленными знаками препинания. Опыт показывает, что 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 $T[i_{end}] \in P$, $i_{end} > i$.

Возможны случаи, когда предложение может нести более одного смысла. Например, фраза *Выгодно продать излишки* может означать оценку факта продажи излишков, и тогда *выгодно* – предикатив. Кроме того, она может означать указание (приказ) о выгодной продаже, и тогда *выгодно* – наречие. Реальный смысл фразы может знать только говорящий или, возможно, его можно определить из более широкого контекста. При наличии только этой фразы автоматическое снятие омонимии невозможно. В случае, когда однозначный выбор омонима компьютером невозможен, по умолчанию меткой «!» помечается наиболее частый вариант МИ.

В случае, когда на отрезке два или более кандидата на предикатив, вопрос о том, как они распределяются по частям речи, требует отдельного исследования.

Этот вопрос заведомо не является простым хотя бы потому, что для двух кандидатов потенциально возможны 682^2 комбинаций.

В подразделах 3.2-3.6 подробно описан метод снятия частиречной омонимии, основанный на правилах, для предикативов, деепричастий, а также групп наречие-существительное.

Программная реализация методов и алгоритмов снятия омонимии имеет несколько режимов для удобства изучения различных аспектов.

На рисунке 3.2 представлено основное окно программы, реализующей разрабатываемые методы и алгоритмы автоматического снятия омонимии. Анализируемое предложение вводится в поле 1. При нажатии кнопки «Таблица» в поле 2 выводится последовательность групп омонимов, определенных с помощью морфологического словаря. В каждой из них программа отмечает выбранный ею омоним восклицательным знаком. В поле 3 автоматически создается таблица с тремя столбцами. Во втором столбце выводятся сверху вниз слова предложения, а в первом для каждого слова записывается слово, его подчиняющее. В третьем столбце выводится грамматическая информация для слова из второго столбца. Эта таблица определяет подчинительное дерево, которое графически отображается в поле 4 с использованием стандартного элемента управления «древовидный список». Настоящая работа содержит ряд результатов о снятии омонимии, но, разумеется, не закрывает проблему в целом. Поэтому построение подчинительного дерева также не является законченным результатом и не выносится на защиту.

Возможны случаи, когда предложение может нести более одного смысла. В этом случае однозначный выбор омонима компьютером невозможен. По умолчанию восклицательный знак ставится, учитывая наиболее частый вариант.

При необходимости можно скорректировать выбор программы вручную, щелкнув мышкой на нужной строке, и получить новый результат работы программы, нажав кнопку «C» (correct).

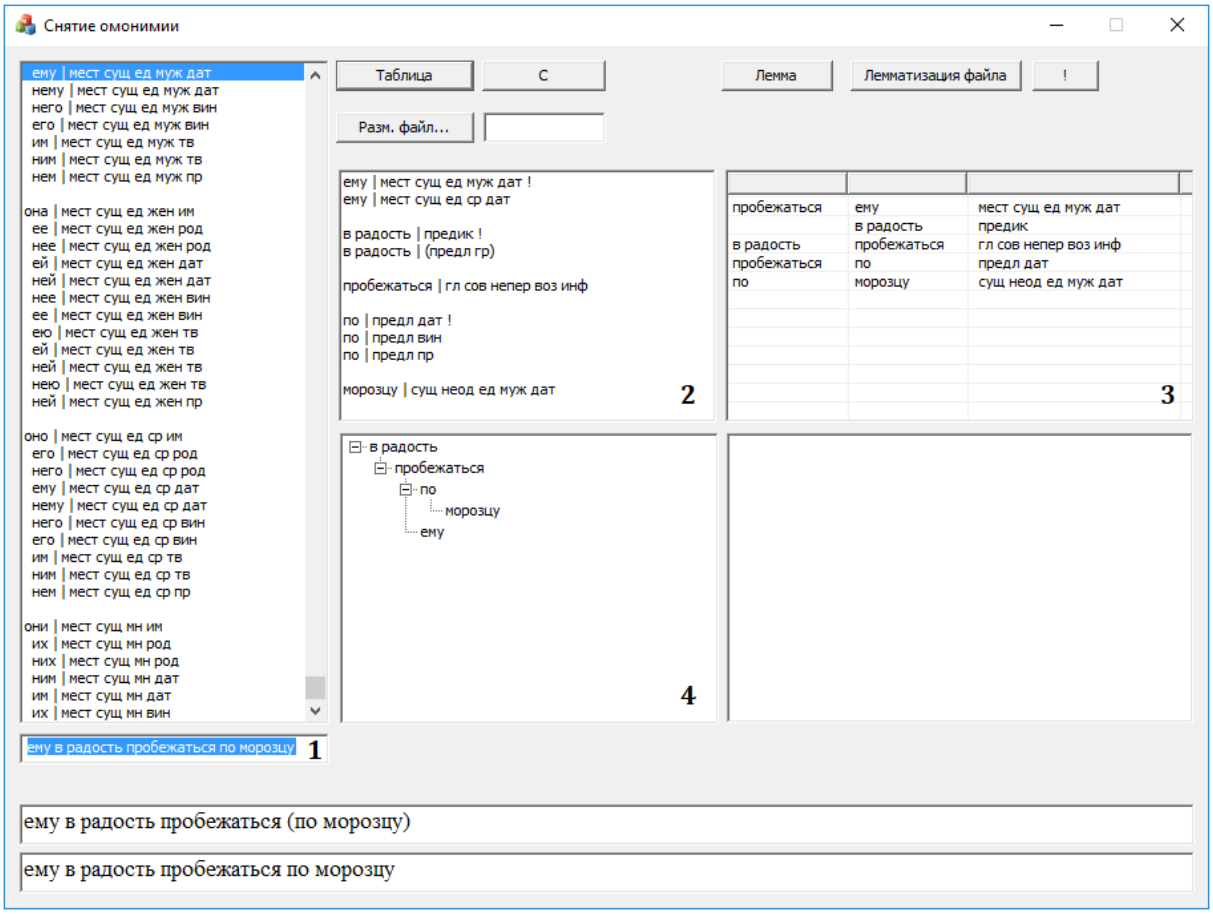

Рисунок 3.2 – Основное окно программы для снятия омонимии

Кнопка «Лемма» служит для определения начальной формы слова или слов, которые вводятся в поле 1. Начальные формы слов выводятся последовательно через пробел в нижнем текстовом поле. Если слово имеет омонимы, то соответствующие разным омонимам леммы приводятся, разделяясь вертикальной чертой.

Кнопка «Лемматизация файла» используется для определения начальных форм слов текста, содержащегося в текстовом файле, который указывает пользователь. Результат выводится в текстовый файл с именем «out.txt».

Кнопка «!» служит для обновления информации из текстовых управляющих файлов, используемых программой, после того как пользователь вносит в них изменения.

Кнопка «Разметить файл» предназначена для частеречной разметки заданного слова или словосочетания в предложениях, отобранных из текстового

корпуса. Там, где заданное слово встречается в тексте, программа выполнит автоматическую разметку по частям речи со снятием омонимии и в скобках выведет результат. С ее помощью можно проверить корректность определения части речи для заданного слова или словосочетания и найти ошибки. Результат разметки выводится в текстовый файл с именем «out.txt».

3.2 Метод снятия омонимии предикатив-наречие-краткое прилагательное в случае единственного кандидата на предикатив

В работе используется МС – сформированный автором морфологический словарь русских словоформ на базе словаря М.А. Хагена, описанный в подразделе 2.1, который содержит 1308 слов и словосочетаний, отмеченных как предикатив. Из них только 104 не имеют омонимов. Наибольшее число предикативов (числом 682) имеют омонимы в виде наречий и (или) кратких прилагательных. Настоящий подраздел посвящен проблеме снятия омонимии именно в классе предикативов, омонимичных наречиям и (или) кратким прилагательным.

Нижеизложенное в данном подразделе относится к отрезку между двумя соседними знаками препинания, на котором есть ***единственный*** кандидат на предикатив, выраженный одним словом [83].

Если сказуемое выражено традиционным способом, то кандидат на предикатив, наречие или краткое прилагательное не является предикативом [84]. Вот правила, связанные с такой ситуацией.

1. Пусть есть некоторый кандидат на предикатив, который может быть также кратким прилагательным. Пусть есть некоторый кандидат на предикатив – слово *A*, который может быть также кратким прилагательным. Если на отрезке $(T[i_{start}],\ T[i_{end}])$ есть существительное *B* (мест.-сущ.) среднего рода в именительном падеже (подлежащее), то кандидат – краткое прилагательное.

Пример: В результате анализа предложения *Его утверждение антинаучно* метод, использующий это правило, в качестве части речи словоформы *антинаучно* вернет прилагательное [85].

Запишем данное правило в формальном виде.

Если для *A* существуют $j_{пред.}$ и $j_{прл.}$ для которых выполняется условие:

$A[j_{пред.}].pos = «предик» \&\& A[j_{прл.}].pos == «прл» \&\& A[j_{прл}].short == «крат»$

и существует $B[j_{сущ.}]$, для которого выполняется условие

$B[j_{сущ.}].pos == «сущ» \&\& B[j_{сущ.}].case == «им» \&\& B[j_{сущ.}].gender == «ср»$

то $(A[j_{пред.}] \mid A[j_{прл.}]) \rightarrow A[j_{прл.}]$

2. Если на отрезке есть существительное (местоимение-существительное) в именительном. падеже, отсутствует глагольное сказуемое, а последнее выражено прилагательным (тоже в им. падеже), то кандидат на предикатив и наречие является наречием.

Пример: Пингвин невероятно(нар) красивый.

Запишем данное правило в формальном виде. Если на отрезке $(T[i_{start}], T[i_{end}])$ есть существительное *B* (местоимение-существительное) в именительном. падеже, отсутствует глагольное сказуемое, а последнее выражено прилагательным *C* (тоже в им. падеже), то кандидат на предикатив и наречие *A* является наречием [86].

Если для *A* существуют $j_{пред.}$ и $j_{нар.}$ для которых

$A[j_{пред.}].pos = «предик» \&\& A[j_{нар.}].pos == «нар»$

и существует $B[j_{сущ.}]$ для которого

$B[j_{сущ.}].pos == «сущ» \&\& B[j_{сущ.}].case == «им»$

и существует $C[j_{прл.}]$ для которого

$C[j_{прл.}].pos == «прл»$

то $(A[j_{пред.}] \mid A[j_{нар.}]) \rightarrow A[j_{нар.}]$

3. Кандидат на предикатив и наречие в пределах предложной группы является наречием

Пример(императив)*:Писать о глубоко*(нар) *важных явлениях!*

Запишем данное правило в формальном виде. Если на отрезке $(T[i_{start}], T[i_{end}])$ есть перед кандидатом стоит предлог T[k-1], то кандидат на предикатив и наречие T[k] является наречием.

Если для T[k] существуют $j_{пред.}$ и $j_{нар.}$ для которых

$T[j_{пред.}].pos$ = «предик» && $T[j_{нар.}].pos$ == «нар»

и существует T[k-1] для которого

T[k-1].*pos* == «предл»

то ($T[j_{пред.}]$ | $T[j_{нар.}]$) → $T[j_{нар.}]$

4. Если на отрезке есть глагол в личной форме или повелительном наклонении, то кандидат на наречие и предикатив есть наречие.

Примеры: Они легкомысленно(нар) *отказались от предложения. Мягко(нар) нажмите на педаль газа.*

5. То же относится к причастию и деепричастию.

Примеры: Это человек, интересно (нар) *рассказывающий о прошлом. Он вошел, комично*(нар) *прихрамывая.*

Значит последние два правила можно объединить следующим образом:

4-5. Если на отрезке есть глагольная форма, отличная от инфинитива, то кандидат на наречие и предикатив есть наречие.

Запишем объединённое правило 4-5 в формальном виде. Если на отрезке $(T[i_{start}], T[i_{end}])$ есть глагол *B* в личной форме или повелительном наклонении, то слово-кандидат *A* на наречие и предикатив есть наречие. То же относится к причастию и деепричастию.

Если для *A* существуют $j_{пред.}$ и $j_{нар.}$ для которых

$A[j_{пред.}].pos$ = «предик» && $A[j_{нар.}].pos$ == «нар»

и существует $B[j_{гл.}]$ || $B[j_{прч.}]$ для которого

$(B[j_{гл.}].pos$ = «гл» && $B[j_{гл.}].verb\_form$!= «инф») || $(B[j_{прч.}].pos$ = «прч»)|| $(B[j_{деепр.}].pos$ = «деепр»)

то ($A[j_{пред.}]$ | $A[j_{нар.}]$) → $A[j_{нар.}]$

6. Если вышеприведенные условия не выполняются, то обсуждаемый кандидат на предикатив действительно является предикативом за исключением случаев, оговоренных в пунктах 8 и 9.

Запишем правило 6 в формальном виде. Если на отрезке $(T[i_{start}], T[i_{end}])$ есть глагол *B* в личной форме или повелительном наклонении, то слово-кандидат *A* на наречие и предикатив есть наречие.

Если для *A* существуют $j_{пред.}$ и $j_{нар.}$ для которых

$A[j_{пред.}].pos = «предик» \&\& A[j_{нар.}].pos == «нар»$

и не существует $B[j_{гл.}] \,||\, B[j_{прч.}] \,||\, B[j_{предл.}] \,||\, B[j_{сущ.}] \,\&\&C[j_{прл.}]$для который

$(B[j_{гл.}].pos = «гл» \&\& B[j_{гл.}].verb\_form != «инф») \,||\, (B[j_{прч.}].pos = «прч») \,||\, (B[j_{деепр.}].pos = «деепр») \,||\, (B[j_{предл.}].pos == «предл») \,||\, (B[j_{сущ.}].pos == «сущ» \&\& B[j_{сущ.}].case == «им» \&\& C[j_{прл.}].pos == «прл») \,||\, (B[j_{сущ.}].pos == «сущ» \&\& B[j_{сущ.}].case == «им» \&\& B[j_{сущ.}].gender == «ср»)$

то $(A[j_{пред.}] \mid A[j_{нар.}]) \rightarrow A[j_{пред.}]$

7. Наличие вспомогательных глагольных словоформ *БЫЛО, БЫВАЕТ, БУДЕТ, СТАЛО, СТАНОВИТСЯ, СТАНЕТ* не превращает предикатив в наречие, если отрезок не содержит существительного (местоимения-существительного) в именительном падеже (подлежащего)

Пример: Ему будет холодно (предик).

Запишем правило 7 в формальном виде. Если на отрезке $(T[i_{start}], T[i_{end}])$ есть вспомогательные глагольные словоформы *БЫЛО, БЫВАЕТ, БУДЕТ, СТАЛО, СТАНОВИТСЯ, СТАНЕТ B* и отрезок не содержит существительного (местоимения-существительного) в именительном падеже (подлежащего) *C*, то слово-кандидат *A* на наречие и предикатив есть предикатив.

Если для *A* существуют $j_{пред.}$ и $j_{нар.}$ для которых

$A[j_{пред.}].pos = «предик» \&\& A[j_{нар.}].pos == «нар»$

и существует $B[j_{гл.}]$ для которого

$B[j_{гл.}].pos = «гл» \&\& B[j_{гл.}].lem \in \{$БЫЛО, БЫВАЕТ, БУДЕТ, СТАЛО, СТАНОВИТСЯ, СТАНЕТ $\}$

и не существует $C[j_{сущ.}]$ для которого

$C[j_{сущ.}].pos == «сущ» \&\& C[j_{сущ.}].case == «им»$

то $(A[j_{пред.}] \mid A[j_{нар.}]) \rightarrow A[j_{пред.}]$

8. При наличии подлежащего появление приведенных вспомогательных глаголов превращает кандидата на наречие и предикатив в наречие.

Пример: Он будет холодно(нар) *отвечать на вопросы.*

Если для *A* существуют $j_{пред.}$ и $j_{нар.}$ для которых

$A[j_{пред.}].pos$ = «предик» && $A[j_{нар.}].pos$ == «нар»

и существует $B[j_{гл.}]$ для которого

$B[j_{гл.}].pos$ = «гл» && $B[j_{гл.}].lem$ ∈ {БЫЛО, БЫВАЕТ, БУДЕТ, СТАЛО, СТАНОВИТСЯ, СТАНЕТ }

и существует $C[j_{сущ.}]$ для которого

$C[j_{сущ.}].pos$ == «сущ» && $C[j_{сущ.}].case$ == «им»

то ($A[j_{пред.}]$ | $A[j_{нар.}]$) → $A[j_{нар.}]$

9. Если на отрезке из глагольных форм есть только инфинитив, то кандидат на наречие и предикатив в большинстве случаев является предикативом.

Пример: Ему холодно(предик) *стоять в карауле.*

Здесь слово «холодно» не является характеристикой действия «стоять», которая была бы наречием, а описывает состояние того, кто стоит.

Запишем правило 9 в формальном виде. Если на отрезке $(T[i_{start}], T[i_{end}])$ нет личной формы глагола B, то слово-кандидат A на наречие и предикатив есть предикатив.

Если для A существуют $j_{пред.}$ и $j_{нар.}$ для которых

$A[j_{пред.}].pos$ = «предик» && $A[j_{нар.}].pos$ == «нар»

и не существует $B[j_{гл.}]$ для которого

$(B[j_{гл.}].pos$ = «гл» && $B[j_{гл.}].verb\_form$!= «инф»)

то ($A[j_{пред.}]$ | $A[j_{нар.}]$) → $A[j_{пред.}]$

Для проведения численных исследований точности разработанного метода в НКРЯ в подкорпусе с вручную снятой омонимией было отобрано 40 случайных фрагментов текста, содержащих кандидата на омонимию предикатив-наречие-краткое прилагательное. В таблице А.1 (Приложение А) приведены результаты разрешения омонимии с помощью разработанного метода в случае единственного кандидата на предикатив, а также проставленный морфологический параметр в НКРЯ, соответствующий части речи. Исследуемое слово выделено в таблице заглавными буквами. Как видно из таблицы, авторский метод дал две ошибки из 40 предложений (предложения 15 и 30), что соответствует точности 95%.

3.3 Метод снятия омонимии предикативных словосочетаний

В предыдущем подразделе рассмотрен случай, когда предикатив выражен одним словом. В данном подразделе описан анализ предикативов, выраженных словосочетанием, но 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.

Для того чтобы разъяснить суть возникающих здесь проблем, 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 используемого МС.

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 [87].

3.3.1 Правила автоматического снятия омонимии предикативных словосочетаний, не являющихся предложными группами

Отметим, что 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».*

Для того, чтобы метод мог в случае необходимости автоматически разбивать словосочетания на отдельные слова, их нужно в морфологическом словаре русских словоформ снабдить соответствующими метками.

В связи со сказанным, представляется целесообразным ввести упомянутые словосочетания-омонимы в морфологический словарь с метками (предл гр), (соч), (отриц). Тогда метод при однозначном выборе этих омонимов будет автоматически разбивать их на отдельные слова [88].

При отсутствии однозначности метод без разбиения словосочетания на отдельные слова поставит омониму метку «!».

Изложим правила снятия омонимии для предикативных сочетаний, не являющихся предложными группами.

1. Правила снятия омонимии предикативных неделимых словосочетаний. Словосочетания:

в ажуре, в долгу, в охотку, в самый раз, в тягость, в ходу, гуд бай, как без рук, на выданье, на заглядение, на слуху, на сносях, не вправе, не для чего, не до смеха, не до шуток, не замай, не к спеху, не надо, не надобно, не под силу, не под стать, не поздоровится, не тут-то было, не худо бы, ни гугу, ни при чем, ни шиша, по нутру, под силу, под стать, пустяк пустяком

фигурируют в МС как цельные единицы, разработанный метод также не разделяет их на отдельные слова.

Для данных словосочетаний группа омонимов будет содержать предикатив и омонимы из числа следующих: наречие, междометие и вводное словосочетание. В работе предложены следующие правила, определяющие выбор части речи омонима:

1.1 Если предложение состоит только из указанного словосочетания, то для него выбирается значение предикатива.

1.2. Если словосочетание выделено запятыми, то это вводное словосочетание.

1.3. Если словосочетание выделено восклицательным знаком, то это междометие.

1.4. Если на отрезке есть глагольная форма, отличая от инфинитива, то для словосочетания выбирается значение наречия. Оно подчиняется упомянутой форме.

1.5. Если в предложении из глагольных форм есть только инфинитив, то для словосочетания выбирается значение предикатива.

1.6. Если на отрезке нет глагольных форм, то в рассматриваемой группе выбирается предикатив.

2. Правило снятия омонимии предикативных словосочетаний с отрицанием. Словосочетания:

не грех; *не дело*; *не беда*; *не впрок*; *не лучше*; *не мудрено*; *не разгонишься*; *не редкость*; *не резон*; *не смешно*; *не удивительно*; *не худо*; *не чета*; *не в ходу*; *не к лицу*; *не к масти*; *не к месту*; *не по душе*; *не по мне*; *не по нутру*; *не по силам*; *не при деньгах*; *не до того*

помечены в МС как ***отр***, не рассматривается как цельная единица, если есть индикаторы разделения. В качестве индикаторов разделения выступают:

– наличие противопоставления, начинающееся противительными союзами «*а*» «но», «да», «зато», «однако», «же» (*Это не худо, а хорошо*);

– наличие запятой после словосочетания (*Он остался не при деньгах, только при славе*);

– для словосочетания *не прочь* наличие в предложении дательного падежа существительного (местоимения-существительного), предлога «*без*» или наречий «*зря*», «*впустую*», «*напрасно*», «*понапрасну*» (*Не прочь другому горя*, *Не прочь результата зря*).

2.1 Если есть индикаторы разделения, то это – не предикативное словосочетание с отрицанием.

3. Правила снятия омонимии предикативных разделимых словосочетаний, которые не являются предложными группами. Сочетания

куда там, куда тут, вероятнее всего, все равно, и все тут, как сказать, может быть, проще всего, так себе

введены в МС с пометкой ***соч***.

3.1. Словосочетания *куда там* и *куда тут* делятся на отдельные слова, если предложение вопросительное и в нем есть одно из следующих слов: *встать, стать, становиться, идти, пойти, зайти, ехать, поехать, заехать, скакать, летать, лететь, полететь, лечь, прилечь, залечь, класть, положить, вкладывать, вложить, заложить, разложить, расположить, ставить, поставить, писать, написать, вписать, вписывать.* С этой целью словосочетание включено в файл «*Предл гр.txt*» (см. следующий подраздел) и слова-индикаторы в нем помечены знаком «?». Данное *с*ловосочетание интерпретируется как предикатив, если дальше есть инфинитив глагола, либо как частица, если инфинитива нет.

3.2. Словосочетание *вероятнее всего* делится на отдельные слова, если после него идет одно из слов *добиться, достигнуть* и дальше слова *при условии,* деепричастный оборот или оборот, начинающийся или союзом *если* или союзом *когда*. В противном случае это предикатив.

3.3. Словосочетание *все равно* делится на отдельные слова, если после него есть числительное в дат. падеже. В противном случае это – предикатив при наличии на отрезке из глагольных форм только инфинитива, и частица при наличии других глагольных форм или полном их отсутствии.

3.4. Словосочетание *и все тут* является предикативом, если после него стоит знак препинания, иначе оно делится на отдельные слова.

3.5. Словосочетание *как сказать* является предикативом в утвердительном предложении, а также в ответе после вопросительного предложения. На отдельные слова оно делится при наличии далее текста в кавычках (*Как сказать по-английски «давно»?*).

3.6. Словосочетание *может быть* делится на отдельные слова, если справа соседствует существительное (местоимение-существительное) или прилагательное (местоимение-прилагательное) или причастие в творительном падеже (*Это может быть существительным. Это может быть красивым. Это может быть решенным.*). При выделении словосочетания запятыми оно является вводным.

3.7. Словосочетание *проще всего* делится на отдельные слова, если после него идет одно из слов *остального, другого, прочего*. В противном случае – это предикатив.

3.8. Словосочетание *так себе* делится на отдельные слова, если после него нет знака препинания (*Так себе и запиши*.). В противном случае при наличии в предложении глагола в личной форме словосочетание является наречием, при отсутствии – предикативом.

3.9. В случае выделения запятыми перечисленных в третьей группе правил словосочетаний, они являются вводным словом.

Для проведения численных исследований точности разработанного метода с помощью программной реализации в НКРЯ было отобрано по 20 случайных фрагментов текста, содержащих кандидатов на омонимию

– предикативных неделимых словосочетаний.

– предикативных словосочетаний с отрицанием.

– предикативных словосочетаний и разделимых словосочетаний, которые не являются предложными группами.

К сожалению НКРЯ не поддерживает возможность разметки словосочетания как единой словарной единицы. В связи с этим данные фрагменты были предварительно размечены вручную, т.е. создана эталонная разметка (ЭР).

В таблицах Приложения А (таблицы А.2-А.4) приведены результаты работы предложенного метода, основанного на правилах снятия омонимии, для предикативных словосочетаний. В этих таблицах анализируемое словосочетание выделено заглавными буквами.

В таблице А.2 содержатся результаты работы метода снятия омонимии для случая предикативных неделимых словосочетаний. Как видно из таблицы А.2, авторский метод дал одну ошибку из 20 предложений (предложение 1), что соответствует точности 95%.

В таблице А.3 содержатся результаты работы метода снятия омонимии предикативных словосочетаний с отрицанием. Как видно из таблицы А.3,

авторский метод дал две ошибки из 20 предложений (предложения 3 и 6), что соответствует точности 90%.

В таблице А.4 содержатся результаты работы метода снятия омонимии предикативных словосочетаний и разделимых словосочетаний, которые не являются предложными группами. На тестовых предложениях метод ни разу не ошибся, что говорит о полноте базы правил для соответствующих словосочетаний.

3.3.2 Лексико-синтаксический алгоритм снятия омонимии словосочетаний, которые могут быть предложными группами

Алгоритм работает с управляющим файлом *Предл гр.txt*, который создан в результате работы с НКРЯ и содержит текстовые записи, определяющие, должно ли словосочетание интерпретироваться как единое целое, или как отдельные слова. Файл представляет собой размеченный словарь словосочетаний, которые могут быть предложными группами, и их управляющих групп. Фрагмент словаря приведен ниже.

в новинку !(предл гр)
вглядываться
вглядеться
вкладывать
вкладываться
вложить
вложиться
всматриваться
всмотреться

в силе !
договор
сохраняться
сохраниться

Управляющие группы для словосочетаний разделены пробельными строками. Слова, расположенные ниже строк с восклицательным знаком, стоят в

начальной форме (леммы). Итоговый объем словаря *Предл гр.txt* составляет более 1300 управляющих групп для различных словосочетаний [89].

Алгоритм снятия омонимии проводит поиск в анализируемом предложении словосочетаний, отмеченных в словаре *Предл гр.txt* восклицательным знаком. Если словосочетание обнаружено, то осуществляется поиск в предложении словоформы какой-либо из нижестоящих лемм управляющей группы этого словосочетания из словаря. Для обработки метки «!» разработаны два правила.

Правило 1. Если после знака «!», стоящего за словосочетанием, есть метка (предл гр), то результаты поиска в предложении словоформы какой-либо из нижестоящих лемм обрабатываются следующим образом:

– поиск не дал результатов, тогда словосочетание интерпретируется как предикатив.

– поиск успешен, тогда словосочетание интерпретируется как предложная группа, алгоритм разобьет словосочетание и вернет список токенов со снятой омонимией: предлог и существительное в отдельных токенах.

Примеры:

1) в результате анализа предложения *Для девушки это было в новинку* предложенный алгоритм, использующий словарь предложных групп, интерпретирует словосочетание *в новинку* как неразделимое и являющееся предикативом, поскольку поиск в предложении словоформ указанных в словаре нижестоящих лемм не дал результатов.

2) в результате анализа предложения *Он решился-таки вложиться в новинку* предложенный алгоритм, использующий словарь предложных групп, интерпретирует словосочетание *в новинку* как предложную группу, разделит его и вернет части речи составляющих его токенов, поскольку поиск в предложении словоформ указанных в словаре нижестоящих лемм дал результат – *вложиться.*

Правило 2. Если после знака «!», стоящего за словосочетанием, нет меток, то результаты поиска в предложении словоформы какой-либо из нижестоящих лемм обрабатываются следующим образом:

– поиск не дал результатов, тогда словосочетание интерпретируется как предложная группа, алгоритм разобьет словосочетание и вернет список токенов со снятой омонимией: предлог и существительное в отдельных токенах.

– поиск успешен, тогда словосочетание интерпретируется как предикатив.

Примеры:

3) в результате анализа предложения *Договор сохраняется в силе* предложенный алгоритм, использующий словарь предложных групп, интерпретирует словосочетание *в силе* как неразделимое и являющееся предикативом, поскольку поиск в предложении словоформ указанных в словаре нижестоящих лемм вернул результат – *сохраниться.*

4) в результате анализа предложения *Мы уверены в силе армии* предложенный алгоритм, использующий словарь предложных групп, интерпретирует словосочетание *в силе* как предложную группу, разделит его и вернет части речи составляющих его токенов, поскольку поиск в предложении словоформ указанных в словаре нижестоящих лемм не дал результатов.

Для корректной обработки управляющих групп алгоритмом снятия омонимии введен ряд меток, которые содержат управляющие группы, приведенные в таблице 3.1.

Таблица 3.1 – Метки в управляющем файле «Предл гр.txt» для алгоритма снятия омонимии

Обозначение метки	**Значение метки и правила**	**Пример употребления метки**
[*часть речи*]	если данное слово встречается в отрезке, то словосочетание интерпретируется как указанная в квадратных скобках часть речи или тип словосочетания.	***в силах !(предл гр)*** вбирать [предик] власть [предик]

Продолжение таблицы 3.1 – Метки в управляющем файле «Предл гр.txt» для алгоритма снятия омонимии

Обозначение метки	Значение метки и правила	Пример употребления метки
(*часть речи 1 – часть речи 2*)	если данное слово встречается в отрезке, то интерпретация словосочетания может быть двоякой. Первоначально программа выбирает часть речи, указанную в скобках перед тире. При этом, если выбирается предложная группа, то она не будет разделена на отдельные слова и пользователь сможет при необходимости изменить результат выбора в окне программы.	***в силах !(предл гр)*** укрепляться (предл гр-предик) укрепиться (предл гр-предик)
(-1)	общее правило выбора работает только в случае, если указанное слово непосредственно предшествует рассматриваемому словосочетанию.	***в годах !(предл гр)*** ни (-1)
(1)	общее правило выбора работает, если указанное слово следует непосредственно за рассматриваемым словосочетанием.	***в ударе !(предл гр)*** же (1)
| *падеж*	правило работает в случае, когда на анализируемом отрезке после словосочетания есть существительное или местоимение-существительное в указанном падеже. Если после падежа стоит (1), то данное слово должно непосредственно следовать за словосочетанием.	***в сборе !(предл гр)*** | род ***в ударе !(предл гр)*** | тв (1)
| ***падеж**	правило работает в случае, когда на анализируемом отрезке перед рассматриваемым словосочетанием есть существительное или местоимение-существительное в указанном падеже.	***в радость !(предл гр)*** | *дат
| *падеж^*	правило работает в случае, когда на анализируемом отрезке после словосочетания есть существительное или местоимение-существительное, имеющее омоним в указанном падеже.	| ***в годах !(предл гр)*** род^
| *слово+падеж*	правило выбора работает, если на отрезке есть указанное слово, за которым следует существительное или местоимение-существительное в указанном падеже.	| ***в сборе !(предл гр)*** с+тв | со+тв
0 (предл гр-предик)	означает, что если глагол с такой меткой встречается в анализируемом отрезке в инфинитиве, то интерпретация словосочетания зависит от контекста и может быть двоякой. По умолчанию выбирается значение предложной группы. Если глагол встретился не в инфинитиве, то словосочетание однозначно интерпретируется как предложная группа.	***в моде !(предл гр)*** занимать 0 (предл гр-предик) играть 0 (предл гр-предик)

Продолжение таблицы 3.1 – Метки в управляющем файле «Предл гр.txt» для алгоритма снятия омонимии

Обозначение метки	**Значение метки и правила**	**Пример употребления метки**
им (предл гр-предик)	работает аналогичным образом, если в анализируемом отрезке есть существительное или местоимение-существительное в именительном падеже.	***в моде !(предл гр)*** взгляд им (предл гр-предик)
им (предик-предл гр)	аналог метки «им (предл гр-предик)», отдающий при именительном падеже приоритет предикативу.	***в моде !(предл гр)*** аксессуар им (предик-предл гр)
(-1) (предик-предл гр)	аналог предыдущей метки, работающий со словом, если оно непосредственно предшествует рассматриваемому словосочетанию.	***в годах !(предл гр)*** военный (-1) [предик]
+,	правило выбора работает, если за рассматриваемым словосочетанием стоит запятая, и далее следуют слова, указанные после запятой.	***в ударе !(предл гр)*** +, в бою
#1гл	правило выбора работает, если глагол с такой меткой является единственным в анализируемом отрезке.	
&	Разделитель, который ставится в словосочетание (слова приводятся без лемматизации) в случае необходимости применять правила не для отдельных слов, а для словосочетаний. В этом случае общее правило выбора будет работать, если в тексте будет присутствовать сочетание первого слова с любым из слов, перечисленных после разделителя.	***в годах !(предл гр)*** от&рождества не&таких&бы&и&больших ***в ударе !(предл гр)*** по&оргкомитету,москве,мячу
<список дисциплин>	Метка для управляющей группы словосочетания *в курсе*. Метод обращается к файлу *Список дисциплин.txt.* Если в нем содержится лемма, словоформа которой содержится в анализируемом отрезке текста между двумя знаками препинания, то словосочетание есть предложная группа. Файл пополняемый, его содержимое указано в Приложении Б.	

Помимо словаря *Предл гр.txt* при снятии омонимии словосочетаний, которые могут быть предикативами, в ряде случаев используются правила, разработанные для конкретных словосочетаний, которые могут быть только омонимами предикатив-предложная группа. Правила разработаны для случаев, нерегулируемых метками. Приведем эти правила.

в годах !(предл гр)

возраст	осознаться	совершаться
даже (предл гр-предик)	отдаваться	срок
дело	от&рождества	стиль
военный (-1) [предик]	ошибка	суть
время	продолжительность	трчать
какой (предик-предл-гр)	различие	тот&же (предл гр-предик)
лишь (предл гр-предик)	разница	угнаться [предик]
нежели	разрыв	черт&ли
не&таких&бы&и&больших	расторговаться	а&не
ни (-1)	розница	
	| род^	

В случаях, перечисленных в файле «Предл гр.txt», «в годах» - предложная группа. В остальных случаях:

1) если на отрезке с «в годах» есть глагольная форма (кроме форм глагола «быть»), то это предложная группа. Исключения:

– если перед «в годах» стоит одушевленное существительное, либо одушевленное существительное и наречие, то это предикатив;

– «заметно/совсем/уже в годах» – предикатив;

2) если на отрезке с «в годах» нет глагольной формы (либо есть только форма глагола «быть»), то это предикатив. Исключения:

– если после «в годах» стоит прилагательное в предложном падеже, то это предл. гр.;

– если после «в годах» стоит число, записанное цифрами, то это предл. гр.;

– если перед «в годах» стоит тире, то выбирается предл. гр. без разделения на слова (возможна как предложная группа, так и предикатив).

в диковинку !(предл гр)

вглядываться	вложить	всматриваться
вглядеться	вкладываться	всмотреться
вкладывать	вложиться	

в курсе !(предл гр)

<список дисциплин>	быть (-1) (предик-предл гр)	гений [предик]

глава
заправила (предик-предл гр)
звук
идея (-1)
излагать
изложить
изучать
изучить
использовать
использоваться
к (1)
лекция
монета
некто
обучение
падать
упадать
упасть
подниматься
подход
причина
пройти
проходить
разбираться
разобраться
расхваливать
расхвалить
революция (предл гр-предик)
речь
сообщение
| род (предик-предл гр)
ставить
поставить
уделять
уделить
уделяться
участие
учить
научить
обучать
обучить
обучаться
обучиться
читаться
чтение

Если после «в курсе» стоит слово, которое может быть только в предложном падеже, то «в курсе» - предложная группа (*в курсе нашем последует весьма выгодная перемена*).

В словосочетании «*в курсе моей истории*» для «в курсе» выбирается предикатив.

Если после «в курсе» стоит запятая, за которой следует причастие в предложном падеже, либо наречие и причастие в предложном падеже, то для в курсе выбирается предложная группа (Пример: Т*ак, в курсе, сейчас экзаменующемся, преобладает "Николай", и это обстоятельство дает курсу отпечаток самолюбивый и рассудочный*).

Запись <список дисциплин> означает, что метод должен обратиться к файлу *Список дисциплин.txt* (Приложение Б). Если он найдет в нем слово, словоформа которого содержится в анализируемом отрезке текста между двумя знаками препинания, то словосочетание есть предложная группа [90].

в моде !(предл гр)

авторитет
аксессуар им (предик-предл гр)
в & дизайне
взгляд им (предл гр-предик)
видеть
возможный
год
граница
дело им (предл гр-предик)
дизайн им (предл гр-предик)
есть & что-то, что-нибудь, кое-что
занимать 0 (предл гр-предик)

играть 0 (предл гр-предик)
излишество им (предл гр-предик)
изменение им (предл гр-предик)
исчезать 0 (предл гр-предик)
как и
карьера им (предл гр-предик)
линия им (предл гр-предик)
место им (предл гр-предик)
мех
на
направление им (предл гр-предик)
настроение им (предл гр-предик)
новое им (предл гр-предик)
ортодокс им (предл гр-предик)
перемена им (предл гр-предик)
поддержка им (предл гр-предик)
подражать 0 (предл гр-предик)
политический
понимать 0 (предл гр-предик)
предпочтение им (предл гр-предик)
предпочтение им (предл гр-предик)
преобразование им (предл гр-предик)
принцип им (предл гр-предик)
прогнозировать 0 (предл гр-предик)
прогнозироваться
производить 0 (предл гр-предик)
произвести 0 (предл гр-предик)
простота им (предл гр-предик)
разбираться 0 (предл гр-предик)
революция им (предл гр-предик)
роль 0 (предл гр-предик)
соединять 0 (предл гр-предик)
соединить 0 (предл гр-предик)
создавать 0 (предл гр-предик)
создать 0 (предл гр-предик)
стиль им (предл гр-предик)
тенденция им (предл гр-предик)
трансформация им (предл гр-предик)
сказочка
эксперт им (предл гр-предик)
/"особенно в моде" и ничего больше/

Если сразу за «в моде» на отрезке стоит согласованное прилагательное в предложном падеже, то «в моде» – предложная группа. (Пример: *Последний год характерен полной унификацией и уничтожением границ не только в моде политической*). На месте прилагательного может стоять причастие(через запятую).

Если сразу за «в моде» на отрезке стоит предлог «на», то «в моде» предложная группа. (Пример: *В моде на обувь происходят изменения.*)

Если анализируемый отрезок текста содержит только словосочетание «*особенно в моде*», то «в моде» интерпретируется как предложная группа.

в новинку !(предл гр)

вглядываться	вложить	всматриваться
вглядеться	вкладываться	всмотреться
вкладывать	вложиться	

в сборе !(предл гр)

свобода
заключаться
помеха
мешать
помешать
помощь
помогать

помочь
соотношение
состоять
участие
неучастие
участвовать
вокруг [предик]
| род
| с+тв
| со+тв

найти (предик-предл гр)
наличные (предл гр-предик)
существо [предик]
заставать [предик]
застать [предик]

в силах !(предл гр)

вбирать [предик]
власть [предик]
вобрать [предик]
делать [предик]
если&буду [предик]
сделать [предик]
недостаток
паритет
перевес
преимущество
превосходить

превосходство
предотвратить [предик]
предотвращать [предик]
провозглашать [предик]
провозгласить [предик]
разница
себя [предик]
создавать [предик]
создать [предик]
уверенный
уверенность
| род (1) (предик-предл гр)

укрепляться (предл гр-предик)
укрепиться (предл гр-предик)
участие
уберегать [предик]
уберечь [предик]
это [предик]
+,которых

Если после «в силах» стоит существительное, которое может быть в родительном падеже, то для него выбирается именно он (*в силах безопасности и пограничных частях*). Исключением является слово «*были*», которое в этом случае всегда интерпретируется как глагол (*мы в силах были бы решить все наши проблемы*).

Если на отрезке с «в силах» есть слово «помочь», то это всегда глагол, а не существительное.

Если сразу за «в силах» есть кандидат на прилагательное (местоимение-прилагательное), в предложном падеже, то выбирается именно он, и «в силах»- предложная группа (*все , что было в силах человеческих, для вас делается*).

Если в пределах отрезка есть глагол, не входщий в группу, то в случае, когда он в инфинитиве, «в силах» - предикатив, в случе, когда это другая глагольная форма «в силах» - предл. группа. Если есть 2 глагола, один из которых в инфинитиве, то «в силах»- предикатив.

Если на отрезке есть слово, входящее в группу, то «в силах» - предл. группа (на предыдущее правило внимание не обращаем). Если такого слова нет, но после «в силах» есть существительное в родительном падеже, то вначале используется предыдущее правило.

Если сразу за «в силах» идет «и» + сущ. в предл. падеже, то «в силах» - предл. гр. (*объективное зло в силах и стихиях природы*)

в ударе !(предл гр)

быстрота
двигаться
же (1)
заключаться
захватить
иметься
использование
клавиша
кровь
ошибиться
падать
плечо
практиковаться
присутствовать
развернуться
резкость
смысл
смычок
| тв (1)
срастаться
участвовать
участвующий
участие
четкость
по&оргкомитету,москве,мячу
+, в бою

в чести !(предл гр)

блюсти
дело&только
завидовать
мой (1)
нарицание
непреклонный
ни (-1)
отказать
отказывать
понижать
понимать
потребность
пребывать [предик]
прибыток
расти
сомневаться
| род (1)
убежденность
уверен&будучи
умаление
участие
явиться
+, которую

на сносе !(предл гр)

делаться
кладбище
настаивать

на счету !(предл гр)

банк (предл гр-предик)	лежать	остаться
деньги (предл гр-предик)	миллиард (предл гр-предик)	остаток
держать	миллион (предл гр-предик)	появляться
доллар (предл гр-предик)	накапливать	появиться
зависать	накопить	рубль (предл гр-предик)
зависнуть	находиться	скопиться
задерживать	не&густо	средство
задержать	нет	сумма
замораживать	нету	тысяча
заморозить	оставаться	у (предл гр-предик)
интересный	оставлять	хранить
капитал	оставить	храниться
копейка (предл гр-предик)	|род (предл гр-предик)	

Если перед или после «на счету» стоит числительное или цифры (сразу либо после глагола «быть») то выбирается предложная группа.

В словосочетании «*на счету была каждая марка*» для «на счету» выбирается предикатив [91].

по карману !(предл гр)

а	хлоп	спускаться
барабанить	хлопать	стукать
бедный	хлопнуть	стукнуть
бить	захлопать	стучать
вас&сегодня	похлопать	судить
врезать	похлопывать	тебя&сразу
гладить	проводить	трескать
погладить	провести	треснуть
глядеть	промахнуть	ударять
и&честь	с&боков	ударить
набирать	различия&положения	удар
набрать	ревнитель	хватить
наказывать	сеять	шарить
наказать	посеять	зашарить
	| род	

по плечу !(предл гр)

бить	провести (предл гр-предик)	хватить
бродить	проводить (предл гр-	хлестать
вдарить	предик)	хлестнуть
вести	прыгать (предл гр-предик)	хлопать
врезать	прыгнуть	хлопнуть
гладить	рассыпаться	хлопанье
поглаживать	садануть	хлопок
погладить	скользить	похлопать
двинуть	скользнуть	похлопывать
долбить	проскользить	похлопывание
долбануть	проскользнуть (предл гр-	прихлопывать (предл гр-
достаться	предик)	предик)
ерзать	слегка	прихлопнуть (предл гр-
заехать	стекать	предик)
и&ниже	стучать	хряснуть-
кто	стукнуть	хряскать
колотить	стукать	чиркнуть
лупить	теребить	чиркать
образовываться	потеребить	шарах
образоваться	трепать	шарахать
огреть	потрепать	шарахнуть
полоснуть	трескать	шлепать
постукать	треснуть	шлепнуть
постукивать	трогать	пошлепать
постучать	тронуть	пришлепывать (предл гр-
приходиться (предл гр-	ударять	предик)
предик)	ударить	пришлепнуть (предл гр-
прийтись	удар	предик)

Для проведения численных исследований точности разработанного метода в НКРЯ отобраны все предложения, содержащие кандидатов на омонимию предикативных словосочетаний и словосочетаний, которые могут быть предложными группами. Эти предложения размечены вручную (эталонная разметка), поскольку НКРЯ не поддерживает возможность разметки словосочетания как единой словарной единицы [92].

В таблице 3.2 приведена точность грамматической разметки для словосочетаний, которые могут быть предикативом или предложной группой.

Таблица 3.2 – Точность грамматической разметки для словосочетаний, которые могут быть предикативом или предложной группой

Словосочетание	Всего предложений	Размечены правильно	Точность разметки
в годах	410	406	99%
в диковинку	314	314	100%
в курсе	1663	1657	99%
в моде	673	669	99%
в новинку	146	146	100%
в сборе	783	782	99%
в силах	706	702	99%
в ударе	448	448	100%
в чести	287	277	97%
на сносе	14	14	100%
на счету	706	701	99%
по карману	420	420	100%
по плечу	1562	1559	99%
Всего	**8132**	**8095**	**99,5%**

Как видно из таблицы, предложенный метод дает высокую точность морфологической разметки, более 99%, что объясняется обширной базой правил для конкретных словосочетаний, охватывающих практически все возможные варианты снятия омонимии.

3.4 Правила и словари для снятия омонимии предикатив-существительное

Для снятия омонимии предикатив-существительное используется словарь индикаторов предикатива, который представляет собой текстовый файл *Индик предик.txt*, состоящий из групп, разделенных пробельными строками, следующего содержания:

беда, гибель, гроб, диво, жуть, караул, кошмар, смерть, смех, срам, страсть, страх, стыд, ужас, умора, чудо!

как, до чего, сколь, сколько, насколько

бабах, бум, стук, толк, тук, тюк, хлоп, щелк!
в, по

раз-два!
и

раз!
и

фа, фи!
какой, мерзкий, неинтересный, противный, скучный

Правила снятия омонимии на основе словаря следующие.

1. Если среди последовательностей, отмеченных восклицательным знаком, встречается слово анализируемого предложения, а также после него идет словоформа слова, следующего в группе за восклицательным знаком, то найденное слово выступает в предложении как предикатив.

Пример: в результате анализа предложения *Жуть сколько согнали техники* предложенный алгоритм, использующий словарь *Индик предик.txt*, интерпретирует слово *жуть* как предикативом, поскольку поиск в предложении словоформ указанных в словаре нижестоящих лемм дал результат – *сколько.*

2. Для слова *точка* предикатив определяется только в сочетании *и точка*.

3. Следующие слова будут определяться в предложении как предикатив, если предложение не содержит глагола или содержит его только в форме инфинитива:

благодать, время, горе, грех, далеко, досуг, капут, конец, красота, крышка, лафа, лень, мечта, могила, молодцом, мучение, навалом, недосуг, неохота, нож, отлично, отрада, охота, плохо, пора, пустяк, пустячок, раздолье, силен, скок,

счастье, тепло, тоска, треба, убожество, ура, фа, фи, фора, хана, хвала, хорошо, худо, черед, чур, швах, кап, каюк, куча, молчок, неправда, порядок, проклятие, прорва, рознь, сила, скрип, спасибо, стоп, темнота, теплынь, тик, тик-так, топ, хи-хи, чих

Примеры: *Они молчок. Нам охота искупаться.*

Следует обратить внимание на то, что в МС как предикативы и существительные фигурируют слова «*ага», «аминь», «ах», «бис».* Вывести правила автоматического снятия омонимии для них автору пока не удалось.

Для проведения численных исследований точности разработанного метода с помощью программной реализации в НКРЯ было отобрано 20 случайных фрагментов текста, содержащих кандидата на омонимию предикатив-существительное. Данные фрагменты были предварительно размечены вручную (эталонная разметка), поскольку НКРЯ не всегда поддерживает возможность разметки данных слов как предикатива.

В таблице А.5 (Приложение А) содержатся результаты работы метода снятия омонимии предикатив-существительное, где анализируемый омоним выделен заглавными буквами. Как видно из таблицы А.5, авторский метод дал две ошибки из 20 предложений (предложения 5 и 6), что соответствует точности 90%.

3.5 Правила и словари для снятия омонимии наречие-существительное

Для снятия омонимии наречие-существительное используется словарь глаголов, употребляемых с наречием или существительным. Этот словарь представляет собой текстовый файл *Глаг-нар.txt,* который содержит последовательность групп вида

авансом # (нар), брать, взыскать, внести, вносить, выдавать, выдать, выплатить, выполнить, делать, заплатить, засчитывать, начислять, платить, получать, получить.

авансом # (сущ), воспользоваться, довольствоваться, заинтересоваться, интересоваться, обеспечивать, обеспечить, поинтересоваться, пользоваться, пренебрегать, пренебречь, распоряжаться, распорядиться [93].

Т.е. каждая запись начинается омонимичным словом с меткой #, за которой стоит метка, указывающая указывается вариант его части речи (один их двух: (нар) или (сущ)), после чего следует список глаголов, которые могут употребляться с этим словом, имеющим такой морфологический параметр.

При создании словаря *Глаг-нар.txt* в качестве источника глаголов использовался НКРЯ [94], а также разработанный МС и словари [95, 96]. Итоговый объем сформированного словаря составляет более 1000 записей. Полное содержание файла приведено в Приложении В.

Правило снятия омонимии, использующее словарь *Глаг-нар.txt*: омониму присваивается та часть речи, которой соответствует метка того ряда, где находится глагол, содержащийся в анализируемом предложении.

Кроме словаря *Глаг-нар.txt* используется список предлогов

в комплекте с, в лад с, в погоне за, в ряду с, в связи с, в согласии с, в сообществе с, в соответствии с, в сопоставлении с, в соприкосновении с, в сочетании с, в сравнении с, в унисон с, в уровень с, не считаясь с, по аналогии с, вдогон за, вдогонку за, обок с, вкупе с, вместе с, вплотную с, вразрез с, вровень с, врозь с, вслед за, наедине с, наравне с, наряду с, невысоко над, несообразно с, по сравнению, рядком с, рядом с, следом за, совместно с, совокупно с, согласно с, созвучно с, сообразно с, сообща с, соответственно с, соразмерно с, сравнительно с, кончая, над, надо, начиная, перед, передо, по-за, по-над, по-под, пред, предо.

Правило снятия омонимии, использующее список предлогов: слово является существительным в случае, если оно имеет творительный падеж, а перед ним находится предлог из списка. Иначе слово – наречие.

Для проведения численных исследований точности разработанного метода в НКРЯ в подкорпусе с вручную снятой омонимией было отобрано 20 случайных фрагментов текста, содержащих кандидата на омонимию наречия и существительного. В таблице А.6 (Приложение А) приведены результаты

разрешения омонимии с помощью разработанного метода, а также проставленный морфологический параметр в НКРЯ, соответствующий части речи. Исследуемое слово выделено в таблице заглавными буквами. Как видно из таблицы, авторский метод на тестовых предложениях ни разу не ошибся, что говорит о полноте используемых словарей и списка предлогов.

3.6 Разработка метода автоматического снятия омонимии русских деепричастий

Деепричастие считается глагольной формой, однако, по мнению некоторых ученых, его следует считать самостоятельной (неизменяемой) частью речи, обозначающей добавочное действие при основном.

Следует отметить, что проблема омонимии деепричастий в русском языке стоит особенно остро. Даже в НКРЯ в ряде случаев деепричастия, имеющие омонимы с другими частями речи, размечены неверно.

В данном подразделе рассматривается вопрос автоматического снятия омонимии русских деепричастий. В связи с чем проведено разделение омонимии деепричастий в русском языке на пять видов, для каждого из них сформирован перечень омонимов из разработанного МС. С учетом построенного перечня омонимов для каждого вида разработаны правила для снятия частеречной и морфологической омонимии, которые реализованы в едином методе снятия омонимии русских деепричастий.

Таким образом, для нахождения всех омонимов деепричастий нужно для каждого из деепричастий, входящих в словарь, произвести поиск совпадающих по написанию словоформ по алгоритму: если список индексов для данной словоформы содержит более одного индекса (т.е. таких словоформ в словаре несколько), значит деепричастие имеет омонимы.

Ниже приведены некоторые примеры найденных в МС омонимов деепричастий:

бая | сущ одуш ед муж род

бая | сущ одуш ед муж вин

бая | деепр несов пер/не наст

благодаря | деепр несов перех наст

благодаря | предл дат

богатея | сущ одуш ед муж род

богатея | сущ одуш ед муж вин

богатея | деепр несов непер наст

буря | деепр несов перех наст

буря | сущ неод ед жен им

…

Анализируя полный полученный перечень омонимов деепричастий, выделено пять видов омонимии:

1) деепричастий и существительных;
2) деепричастий и предлогов;
3) деепричастий и прилагательных;
4) деепричастий и причастий;
5) деепричастий переходного и непереходного глагола [77, с. 39].

Схема работы предложенного метода разрешения омонимии деепричастий приведена на рисунке 3.3.

По аналогии с рисунком 3.1, переменная hom_type обозначает вид омонимии деепричастий из списка, приведенного выше. На вход поступает результат грамматической разметки – массив *T*, каждый элемент которого является массивом записей с грамматической информацией, где фигурирует часть речи – деепричастие. Выходными данными является массив *T**, содержащий список токенов со снятой омонимией, вариант выбранной разметки помечен «!».

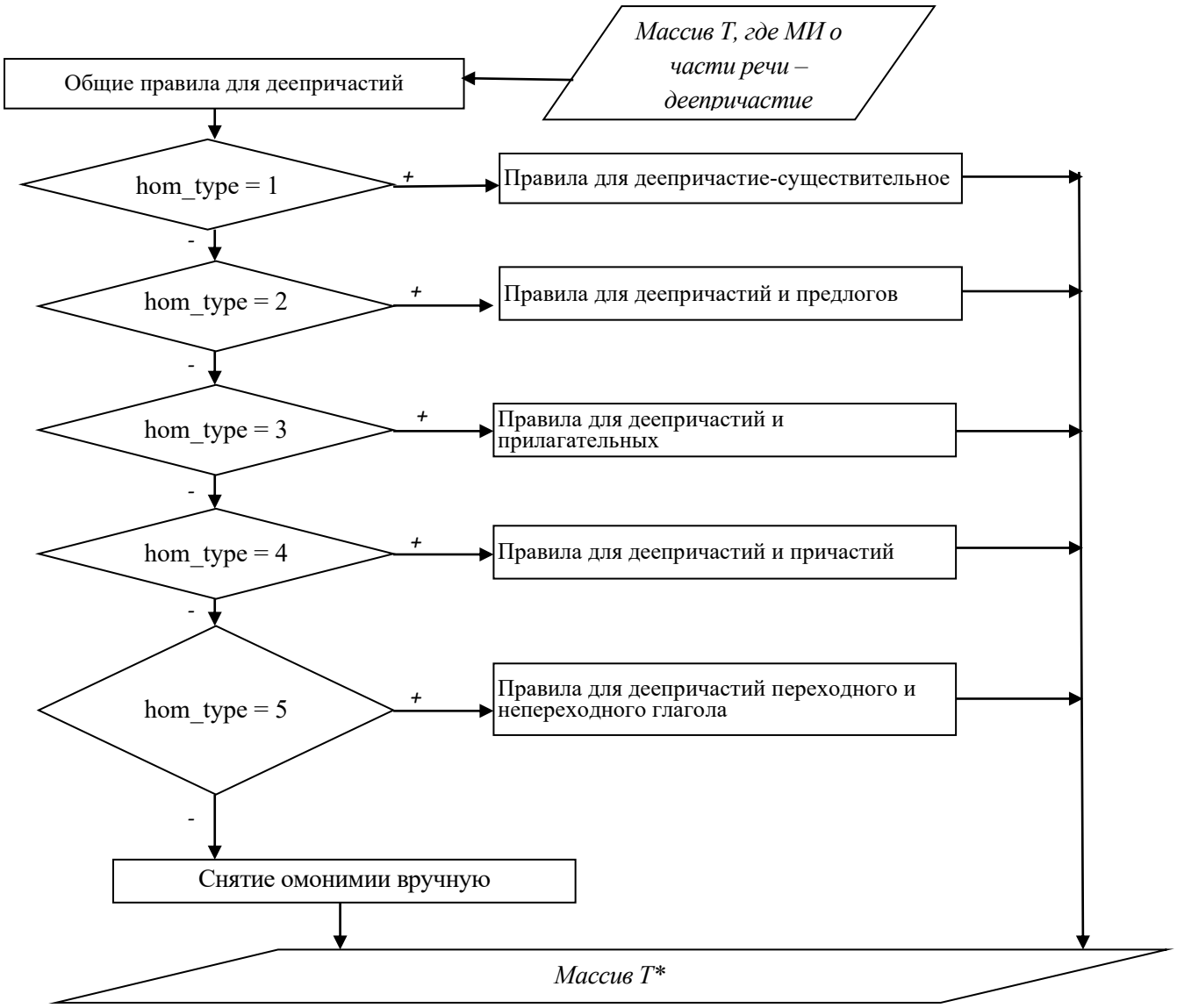

Рисунок 3.3 – Общая схема работы алгоритма снятия омонимии деепричастий

Сформированные перечни омонимов с помощью разработанного МС для каждого из видов приведены в таблице 3.3.

Таблица 3.3 – Перечни омонимов русских деепричастий, полученные из МС

№ вида омонимии	Перечень омонимов из МС
1)	бая, богатея, буря, воя, гвоздя, горя, гостя, доля, душа, ежа, заезжая, залив, заплыв, запоров, застав, клея, клича, корча, кроя, лая, лишая, мая, меча, моля, моря, моча, мытаря, нагоняя, надоев, нажив, налив, напев, нарыв, неволя, обогрев, отлив, отрыв, отсев, отстоя, пав, паря, переборов, перелив, перерыв, пища, плача, плюща, поборов, подлив, подогрев, пожив, полив, поля, пошив, приезжая, приколов, прилив, пристав, проезжая, проколов, пролив, прорыв, разлив, размыв, разогрев, разрыв, расколов, рея, ржа, родня, роя, руля, свища, сев, селя, сеча, сколов, слив, споров, сторожа, строя, суша, туша, уколов, устав, хворая, хромая, чая
2)	благодаря, включая, для, исключая

Продолжение таблицы 3.3– Перечни омонимов русских деепричастий, полученные из МС

3)	бухая, вещая, витая, вшив, горяча, заезжая, приезжая, проезжая, синя, скупая, строгая, хворая, хромая
4)	витая, обитая, питая
5)	вкалывая, досадив, досадивши, досыпая, запахнув, запахнувши, засыпая, метя, мешая, наказывая, находя, недосыпая, отсыпая, отточив, отточивши, парируя, перетрусив, перетрусивши, пикировав, пикируя, планируя, повалив, поваливши, поводя, помешав, помешавши, поправив, поправивши, провалив, проваливая, проваливши, саботируя, свалив, сваливая, сваливши, снуя, тая, точа, узрев, узревши, целя

Рассмотрим перечисленные выше случаи омонимии и сформулируем базу правил для ее автоматического разрешения.

Для автоматического снятия омонимии деепричастия в первую очередь необходимо руководствоваться двумя общими правилами.

Правило 1: если предложение содержит омоним, который может быть деепричастием, то необходимо, чтобы оно содержалось в отрезке предложения, выделенном знаком препинания в начале или в конце предложения или с двух сторон в середине предложения [97].

Правило 2: если отрезок содержит предикатив, личную форму глагола, краткое прилагательное или краткое причастие, то он не может содержать деепричастие. [77, с. 40] Исключение – слово «будучи».

Для разрешения омонимии деепричастий и существительных используются общие Правила 1 и 2.

Для автоматического разрешения омонимии деепричастий и предлогов используется Правило 3, разработанное для каждого слова из соответствующего в таблице 3.5 перечня.

Правило 3: если после слова есть существительное, которое согласуется по падежу с предлогом, то выбирается предлог. В противном случае – деепричастие. В частности:

3.1. Омоним *благодаря* определяется как предлог, если далее в пределах отрезка найдется существительное (местоимение-существительное) в дательном падеже. В противном случае это деепричастие.

3.2. Омоним *для* определяется как предлог, если далее в пределах отрезка найдется существительное (местоимение-существительное) в родительном падеже. В противном случае это деепричастие.

3.3. Омонимы *включая, исключая* как в случае предлога, так и в случае деепричастия употребляются совместно с существительным в родительном падеже, поэтому требуют дополнительных исследований.

Для автоматического разрешения омонимии деепричастий и прилагательных используется Правило 4.

Правило 4: если на отрезке, содержащем омоним деепричастие-прилагательное есть существительное, согласованное с предполагаемым прилагательным, то это действительно прилагательное.

Для разрешения омонимии деепричастий и причастий используется Правило 5 автоматического снятия омонимии для каждого слова из соответствующего списка.

Правило 5: если за словом

5.1. *обитая* следует предлог или слова *там*, *тут*, *здесь*, то это деепричастие. В остальных случаях – причастие. Примеры:.

5.2. *питая* следует существительное в винительном падеже, то это деепричастие. В остальных случаях – причастие. Пример:.

5.3. *витая* следует предлог *из*, или слова *с помощью*, *с использованием*, то это причастие. В остальных случаях – деепричастие.

Для автоматического разрешения омонимии деепричастий переходного и непереходного глагола разработано Правило 6.

Правило 6: деепричастие переходного глагола выбирается, если за ним следует существительное в винительном падеже, в противном случае выбирается деепричастие непереходного. [77, с. 42]

Примеры использования перечисленных Правил 1-6 приведены в таблице 3.4.

Таблица 3.4 – Использование правил автоматического разрешения омонимии деепричастий русского языка

№	Входное предложение	Метка о части речи	№ пра–вила
1	Свиная ТУША у мясника стоит тысячу рублей	сущ	1
2	Задохнулся, ТУША пожар в своей мастерской	деепр	1
3	Понесся дальше, сбивая снежинки хвостом и ЛАЯ от счастья.	деепр	2
4	От громкого ЛАЯ собак звенело в ушах.	сущ	2
5	СТРЕМЯСЬ помочь, он потянул дверь	деепр	2
6	Мальчик, БУДУЧИ определен в кадетский корпус, с раннего детства жил вне семьи	прич	2 исключ
7	Однако за два года, БЛАГОДАРЯ своей тяге к знаниям, выучила неплохо английский, да и другое.	предл	3.1
8	Будем молиться, БЛАГОДАРЯ Бога за то, что эта буря пронеслась	деепр	3.1
9	каждый день прилетал ворон, ДЛЯ его агонию вечно	деепр	3.2
10	отвели ДЛЯ этого специальные места	предл	3.2
11	Уехали все, ИСКЛЮЧАЯ древних стариков	предл	3.3
12	Он действовал жестко, ИСКЛЮЧАЯ нерадивых учеников	деепр	3.3
13	СКУПАЯ старуха выжила в самые сложные времена	прил	4
14	Он выживал в сложные времена, СКУПАЯ ценности	деепр	4
15	Он жил, ОБИТАЯ здесь уже год	деепр	5.1
16	Там дверь, ОБИТАЯ железом	прич	5.1
17	В стакане вода, ПИТАЯ вчера	прич	5.2
18	Дельта разливается, обильно ПИТАЯ окрестности водой	деепр	5.2
19	Веревка, ВИТАЯ из рогожи	прич	5.3
20	Он жил, ВИТАЯ в облаках	деепр	5.3
21	Он работал с утра, КОСЯ траву	перех	6
22	Она сидела, КОСЯ в его сторону	неперех	6

Существуют также деепричастия-омонимы с чисто семантическим отличием (лексическая омонимия). Пример: *«взрывая»* (одно значение от «рыть», другое – от «устраивать взрыв»). Здесь классификация определяется контекстом и на сегодняшний день не может быть выполнена автоматически с помощью морфологической и синтаксической информации без использования языкового моделирования.

Для проведения численных исследований точности разработанного метода в НКРЯ в подкорпусе с вручную снятой омонимией было отобрано 55 случайных фрагментов текста, содержащих кандидата на омонимию деепричастий разных видов из перечня 3.5. В таблице А.7 (Приложение А) приведены результаты разрешения омонимии с помощью разработанного метода в случае единственного кандидата на предикатив, а также проставленный морфологический параметр в НКРЯ, соответствующий части речи. Исследуемое слово выделено в таблице заглавными буквами. Обобщенные результаты и результаты по каждому виду омонимии деепричастий сведены в таблицу 3.5.

Таблица 3.5 – Точность метода снятия омонимии русских деепричастий

№ вида омонимии	Количество тестовых предложений	Количество неверных разметок авторским методом	Точность авторского метода, %
1)	15	0	100
2)	6	0	100
3)	21	1 (предложение 40)	95,2
4)	7	1 (предложение 47)	85,7
5)	7	1(предложение 56)	85,7
Всего	**55**	**3**	**94,6**

Как видно из таблицы 3.5, авторский метод, предложенный для снятия омонимии деепричастий, показывает высокие результаты даже в трудных случаях, что говорит о полноте используемого МС, из которого формируются перечни омонимов, а также базы разработанных правил автоматического разрешения омонимии для деепричастий, охватывающих тонкости русского языка.

3.7 Сравнение эффективности предложенного метода снятия омонимии с существующими решениями

Для проведения сравнительного анализа эффективности предложенных алгоритмов и правил снятия омонимии с известными аналогами выбраны два

морфоанализатора русского языка, являющихся бесспорными лидерами в области морфологической обработки русскоязычных текстов в NLP-задачах:

– MyStem – морфологический анализатор русского языка от компании Яндекс [98]. Его исходный код закрыт и принадлежит ООО Яндекс, но для возможности использования MyStem различных проектах разработаны программные библиотеки, предоставляет программный интерфейс к анализатору. В частности в рамках данного диссертационного исследования использовалась библиотека pymystem3 для языка программирования Python;

– Pymorphy2 – морфологический анализатор, разработанный на языке программирования Python [99], поддерживающий русский и украинский языки. Выполняет лемматизацию и морфологический анализ и синтез слов, работает со словарём OpenCorpora, а для незнакомых слов строит гипотезы.

Для тестирования этих анализаторов использовались те же тексты, на которых проверялась точность работы авторского метода, они приведены в Приложении А (таблицы А.1-А7) для каждого вида омонимии, описанного в этом разделе. Результаты численных исследований сравнения качества работы методов автоматического разрешения омонимии для различных групп омонимов сведены в таблицу 3.6. В таблице введена следующая нумерация групп омонимов:

1. Предикатив-наречие-краткое прилагательное в случае единственного кандидата на предикатив.
2. Предикативные словосочетания – неделимые словосочетания.
3. Предикативные словосочетания – словосочетания с отрицанием.
4. Предикативные словосочетания – разделимые словосочетания, которые не являются предложными группами.
5. Предикативные словосочетания – предложная группа.
6. Предикатив-существительное.
7. Наречие-существительное.
8. Деепричастия.

Обозначения морфоанализаторов, используемые в таблице: MySt для MyStem; pyM2 для Pymorphy2; work – авторский метод.

Таблица 3.6 – Точность методов снятия омонимии для различных групп омонимов

№	Кол-во тестовых предложений	Размечены правильно			Точность разметки, %		
		MySt	pyM2	work	MyStem	pymorphy2	work
1	40	21	19	38	52,5	47,5	95
2	20	0	0	19	0	0	95
3	20	2	2	18	10	10	90
4	20	3	3	20	15	15	100
5	8641	1901	2074	8590	22	24	99,5
6	20	7	7	18	39	39	90
7	20	9	9	20	45	45	100
8	55	27	28	53	49	51	96
Всего	**8837**	**1970**	**2142**	**8775**	**22**	**24**	**99,3%**

Как видно из таблицы 3.8, наилучшие показатели точности имеет авторский метод снятия омонимии. Морфоанализаторы Pymorphy2 и MyStem не распознают предикативы и предикативные словосочетания, а омонимию наречие-существительное и деепричастий снимают лишь частично.

Результаты морфологической обработки текста – последовательность лемм и соответствующей МИ каждого токена предложений текста со снятой омонимией, поступают на вход блока сематической обработки текста.

3.8 Выводы к разделу 3

В разделе 3 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 в некоторых случаях частеречную омонимию. Авторский метод разрешает частеречную омонимию предикативов и предикативных словосочетаний, деепричастий, групп наречие-существительное со средней точностью 99,3%. Такое высокое качество говорит о полноте используемых словарей, а также базы разработанных продукционных правил автоматического разрешения омонимии.

РАЗДЕЛ 4
РАЗРАБОТКА АЛГОРИТМОВ СЕМАНТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТА

В данном 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.

4.1 Общая схема работы системы обработки и анализа текстовой информации

Разработанная система состоит из двух блоков: блока морфологической обработки и семантической обработки текста. Алгоритмы первого блока описаны в предыдущем разделе, в результате лемматизации и снятия омонимии на вход блока семантической обработки поступает массив пар $\left\{\left\{\langle лемма_j^i, МИ_j^i\rangle\right\}_{j=1}^{N\_j}\right\}_{i=1}^{M}$ для каждого j-го токена i-го предложения текста, N_j – количество токенов в предложении j, M – количество предложений текста.

Общая схема работы системы обработки и анализа текстовой информации приведена на рисунке 4.1.

В текущем разделе описаны алгоритмы модулей блока семантической обработки системы, которых, как видно из рисунка 4.1, три:

– модуль лексической адаптации текста;

– модуль разбиения текста на семантически однородные фрагменты, который необходим для того, чтобы не терять смысл при упрощении текста, поскольку целесообразно работать с каждым из таких фрагментов в отдельности;

– модуль построения элемента плана текста, который извлекает основной смысл предложений текста в сжатом виде путем автоматического создания элементов плана за счет использования словаря отглагольных существительных.

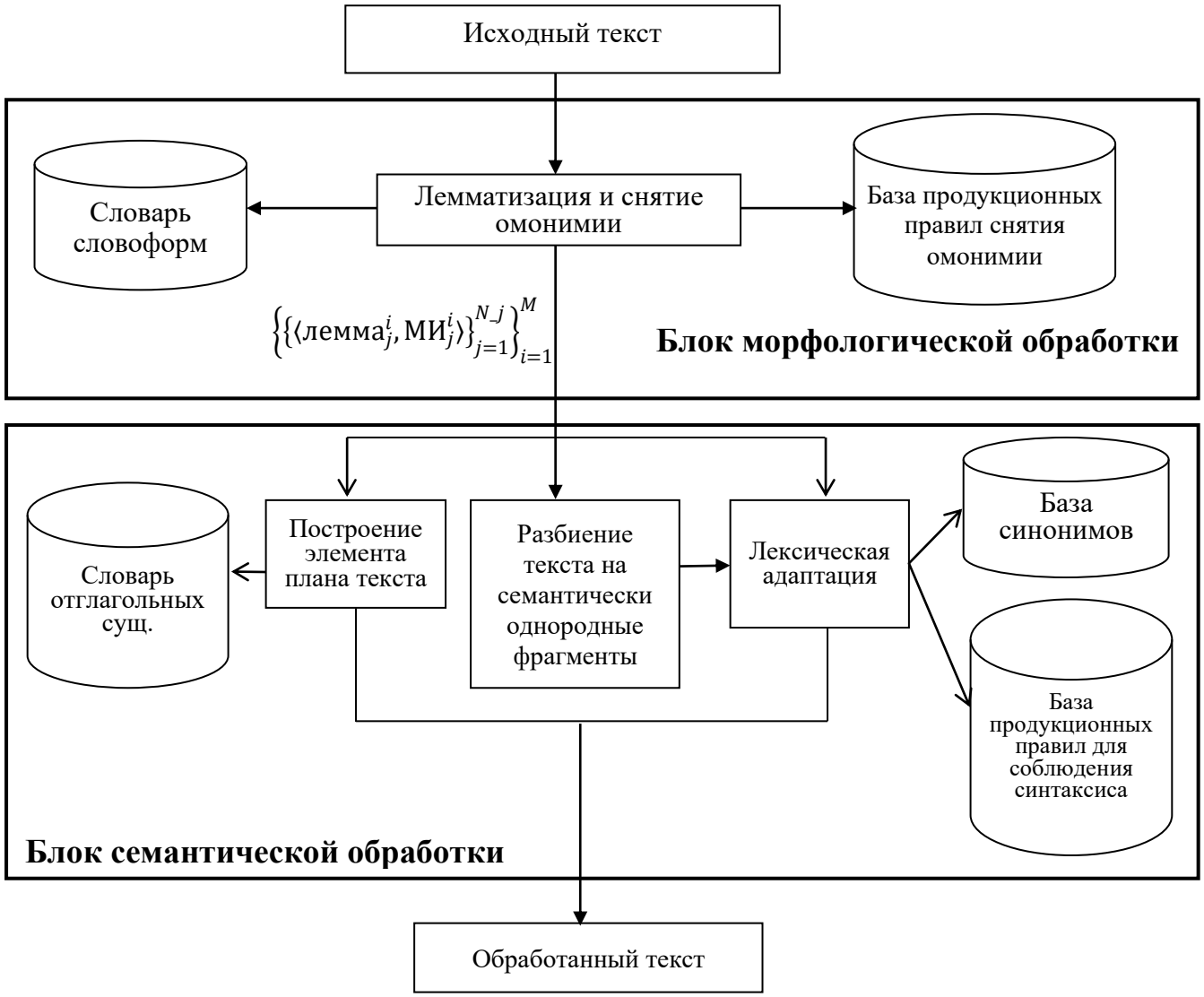

Рисунок 4.1 – Общая схема работы системы обработки и анализа текстовой информации

Алгоритмы модуля лексической адаптации наиболее сложные по сравнению с алгоритмами остальных модулей семантического блока. Адаптация текста может происходить на разных уровнях обработки текста: лексическая, грамматическая, синтаксическая [100, 101, 102].

Среди элементов адаптации, выполняемой человеком, могут быть выделены следующие:

1. Отказ от предложений или целых абзацев, которые могут быть опущены без существенных потерь для общего содержания текста.

2. Отказ от отдельных слов в предложениях на тех же условиях.

3. Замена предложений синтаксически более простыми.

4. Лексическое упрощение, а именно, замена отдельных слов и словосочетаний их более общими или более употребительными синонимами.

«Несмотря на активное изучение проблемы автоматизированной адаптации текста, решения пока найдены далеко не для всех исследовательских задач в этой области, а существующие методы применимы к сравнительно небольшому числу языков. Таким образом, очевидна необходимость дальнейшей разработки данной проблемы» [79, с.81]

Метод лексической адаптации использует словари, содержащие синонимические ряды. Идея предложенного метода синонимической замены кратко формулируется так: член синонимического ряда, встретившийся в тексте, должен быть заменен соответствующей доминантой.

В русском языке синонимы, как правило, обладают различными морфологическими параметрами, что создает трудности при автоматической замене, связанные с соблюдением правил синтаксиса в адаптированном тексте. В рамках данного исследования для решения этих проблем используется база продукционных правил для соблюдения синтаксиса, а также механизм обработки меток в словарях синонимов, разработанный автором.

Ниже приведено описание алгоритмов каждого из модулей семантического блока.

4.2 Разработка алгоритмов синонимических замен с целью упрощения (адаптации) русскоязычных текстов

4.2.1 Формирование базы синонимов на основе данных из открытых источников для системы синонимических замен

Лингвистической основой для разработки описанного метода синонимических замен могут служить словари синонимов [95, 103-112], а также частотные словари русского языка [113]. В данной работе используется «Словарь синонимов русского языка» З.Е. Александровой [95] и Электронная версия издания: О.Н. Ляшевская, С.А. Шаров «Частотный словарь современного русского языка (на материалах Национального корпуса русского языка)» [113].

В основе организации словаря З.Е. Александровой лежит понятие синонимического ряда. Как указано в предшествующем словарю описании, «Синонимический ряд начинается заглавным словом (доминантой). За ним следуют синонимы – члены ряда. Доминанта должна быть стилистически нейтральной, семантически прозрачной и емкой, по сравнению с другими членами ряда иметь большую употребительность и главное – наиболее широкую сочетаемость, общую со всеми членами ряда. За неимением нейтрального слова, отвечающего этим требованиям, доминантой может оказаться книжное, разговорное или даже устарелое слово. Иногда доминантой является словосочетание (например, ЖЕЛЕЗНАЯ ДОРОГА, СТАРАЯ ДЕВА)».

Отметим сразу, что при создании системы синонимических замен нельзя использовать доминанты и синонимические ряды словаря З.Е. Александровой совершенно формально, ибо этот словарь создан как справочник для людей, пишущих на русском языке. Так для доминанты АВТОМОБИЛЬ словарь предлагает следующий синонимический ряд:

автомобиль ! авто, автомашина, железный конь, железка, жестянка, кар, колеса, машина, мотор, тачка

Для человека, пользующегося словарем, это набор ценных подсказок, позволяющий найти и использовать наиболее выразительный и стилистически подходящий синоним. Но для компьютерной программы, которая всегда будет заменять член ряда доминантой, слова железка, жестянка, колеса, машина, мотор, тачка неприемлемы, ибо в каких-то ситуациях они будут обозначать совсем не автомобиль.

В связи с этим проведен анализ и сокращение предлагаемых синонимических рядов. Это большая работа филологического характера. Для того, чтобы замена не приводила к нелепому результату, смыслы, выражаемые недоминантным членом синонимического ряда, не должны выходить за пределы смыслов, выражаемых доминантой. В то же время доминанта может обладать и другими смыслами. Пример:

подтягивать !

подпевать

Здесь «подтягивать», помимо «подпевать», может использоваться иначе: «подтягивать гайку», «подтягивать брюки» и т.д. В частности, доминанта одного синонимического ряда может быть недоминантным членом другого синонимического ряда.

Кроме того, с помощью словаря О.Н. Ляшевская, С.А. Шаров «Частотный словарь современного русского языка (на материалах Национального корпуса русского языка)» для каждой группы проанализирована частотность доминанты и членов синонимического ряда, и при возможности в качестве доминанты выбран синоним с наибольшей частотностью.

Обратим внимание на то, что слово железка в том же словаре З.Е. Александровой может иметь другую доминанту ЖЕЛЕЗНАЯ ДОРОГА. В базе синонимов могут встречаться слова, которым соответствуют несколько доминант. Тогда система при синонимической замене выдаст несколько вариантов в скобках, так что нужное может быть выбрано одним щелчком мыши. Это относится к заменам как отдельных слов, так и словосочетаний.

На основе данных из открытых источников [95, 113] сформирована базы синонимов для системы синонимических замен. Состав базы синонимов отображен на рисунке 4.2. База синонимов состоит из 3 файлов:

- *База.txt* используется для синонимических замен отдельных слов. Содержит синонимические ряды в начальной форме, а также специальные метки для восстановления правильного синтаксиса. Полное содержание файла приведено в Приложении Г;

- *База-соч.txt* используется для синонимических замен словосочетаний. Содержит синонимические ряды, приведенные в начальную форму, т. е. словосочетания отлемматизированы, а также специальные метки для восстановления правильного синтаксиса. Полное содержание файла приведено в Приложении Д;

- *Форма.txt* используется для замены неизменяемых словосочетаний, которые используются только в той форме, как указано в базе, а также не требуют восстановления правильного синтаксиса. Содержит синонимические ряды не в начальной форме, т. е. без использования лемматизации. Полное содержание файла приведено в Приложении Е.

Рисунок 4.2 – Состав базы синонимов

Всего база синонимов насчитывает 24 935 слов, содержащихся в 7225 синонимических рядах, 13 100 словосочетаний, содержащихся в 3000 синонимических рядов, а также 3 850 неизменяемых словосочетаний в 750 синонимических рядах.

Рассмотрим более подробно формат данных фалов.

В соответствии с вышесказанным для системы замены отдельных слов создается файл *База.txt*, представляющий собой последовательность групп, каждая из которых начинается соответствующей доминантой (отмечается в файле восклицательным знаком), а затем содержит отрезок синонимического ряда (см. приложение Г). Пример такой группы:

бесконечность !

безграничность

безбрежность

безмерность

бескрайность

беспредельность

необозримость

необъятность

Для удобства группы разделены пробельными строками. Вместе с каждым словом в базу можно было бы включить все его словоформы. Это увеличило бы ее на порядки, но, представив ее в виде дерева, можно было бы обеспечить в ней, как и в словаре О.Н. Ляшевская, С.А. Шаров «Частотный словарь современного русского языка (на материалах Национального корпуса русского языка)», практически мгновенный поиск. Однако для того, чтобы база была легко обозримой с точки зрения семантики, в данной работе в файлы *База.txt* и *База-соч.txt* включены только начальные формы слов и используется лемматизация исходного текста [103].

При синонимических заменах словосочетаний используется файл *База-соч.txt* (см. приложение Д), аналогичный файлу База.txt и состоящий из групп вида

алкоголь !

спиртной напиток

крепкий напиток

горячительный напиток

зеленый змий

зеленый вино

дар вакх

Для удобства группы разделены пробельными строками. Доминанта отмечена знаком «!», а все члены синонимического ряда приведены в начальную форму (отлемматизированы). Доминантой может служить не только одно слово, но и словосочетание.

Не все члены синонимического ряда используются в тексте с одинаковыми морфологическими характеристиками. Например, в тексте «По весне настала распутица», происходит замена на доминанту синонимического ряда «бездорожье», которая имеет другие морфологические характеристики (другой род). Для корректной синонимической замены в результирующем тексте требуется восстановить правильный синтаксис: изменить род связанных с синонимом слов. Например, «По весне настало бездорожье» [114].

Зачастую для получения в замененном предложении правильного синтаксиса требуется изменить синтаксическое управление, т. е. восстановить правильный синтаксис. 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-[на] вин

означает, что, если существительное, стоит следом за заменяемым словом в творительном падеже с предлогом «с» то за заменяющим словом оно должно следовать в винительном падеже с предлогом «на». Отдельные элементы в этой записи могут отсутствовать. Например, 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» [115].

Рассмотрим более подробно пометки, использующиеся в базе синонимов, как при замене отдельных слов, так и словосочетаний:

1) Пометка «1» следом за доминантой означает, что при замене доминанта используется только в единственном числе. Пример:

поведение ! 1

поступок

«Он озадачил меня своими поступками» преобразуется в «Он озадачил меня своим поведением».

2) Пометка «2» следом за доминантой означает, что при замене доминанта используется только во множественном числе. Пример:

аплодисменты ! 2

овация

рукоплескание

«Концерт закончился овацией» преобразуется в «Концерт закончился аплодисментами».

3) Пометка «2» следом за членом синонимического ряда означает, что этот член заменяется доминантой только тогда, когда стоит во множественном числе. Пример:

артиллерия ! 1

орудие 2

пушка 2

«Орудия стреляют» преобразуется в «Артиллерия стреляет». «Орудие стреляет» программа оставляет без изменения.

4) Пометка /форма/ около члена синонимического ряда означает, что в процессе обработки слова оно не подвергается лемматизации. Пример:

способен !

горазд /форма/

«Он горазд плясать» преобразуется в «Он способен плясать».

5) Пометка /форма,/ (с запятой) около члена синонимического ряда означает, что в процессе обработки слова оно не подвергается лемматизации лишь в случае, когда после него стоит запятая или точка. Пример:

нет!

дудки /форма,/

«Дудки, ничего не выйдет!» преобразуется в «Нет, ничего не выйдет!»без предварительной лемматизации слова «дудки».

6) Пометка {в-во} означает, что при замене предлог «в» заменяется предлогом «во». Пример:

время ! 2 {в-во}

година

«В годину испытаний» преобразуется в «Во времена испытаний».

7) Пометка /сущ/ означает, что замена происходит только в том случае, когда член синонимического ряда определен в предложении как существительное. Пример:

бюллетень ! /сущ/

больничный

Если бы обсуждаемой пометки не было, то программа преобразовала бы предложение «Прошли больничный коридор» в «Прошли бюллетень коридор». [103]

8) Пометка +множ означает создание при замене вариантов единственного и множественного числа (в скобках), выбор между которыми осуществляется по щелчку. Пример: «Они живут в хоромах» преобразуется в «Они живут во дворце (дворцах) »

9) Еще одна специальная пометка связана с числительными:

два ! двойка^

три ! тройка^

четыре ! четверка^

пять ! пятерка^

шесть ! шестерка^

семь ! семерка^

восемь ! восьмерка^

девять ! девятка^

десять ! десятка^, десяток^

двадцать ! двадцатка^

тридцать ! тридцатка^

пятьдесят ! пятидесятка^

сто ! сотня^

пятьсот ! пятисотка^

Значок ^ у синонима означает, что замена производится только, если следом за ним стоит существительное или прилагательное. Пример: «Подошел десяток машин» поменяется на «Подошли десять машин». Но «Поехали на тройке» останется без изменения.

Для склонения сочетаний числительных с существительными типа «три лошади» используется отдельная подпрограмма, учитывающая род и одушевленность существительного.

10) # прош – при замене данного слова глаголом этот глагол ставится в прошедшем времени. Пример: «Он бряк на землю» преобразуется в «Он упал на землю» [116].

Для восстановления синтаксиса при замене словосочетаний прежде всего работают те же пометки, что и при замене отдельных слов. Но кроме этого, используются дополнительные:

!/<часть_речи>/ – среди омонимов слова-замены всегда выбирается указанная часть речи.

$ крат – если слово-замена – прилагательное, то оно ставится в краткой форме.

наст – при замене данного словосочетания, содержащего глагол прошедшего или будущего времени, перед словом-заменой не ставится словоформа глагола-связки «быть».

наст – при замене словосочетания, содержащего причастие прошедшего времени на глагол этот глагол ставится в настоящем времени

наст – эта метка применяется также, если участвующая в словосочетании форма глагола может интерпретироваться как будущее время.

Пример: Собрание выносит постановление.

буд – при замене данного словосочетания, содержащего глагол будущего времени, перед словом-заменой (прилагательным) ставится глагол-связка «будет».

наст,прош – при замене данного словосочетания, содержащего глагол прошедшего времени, заменяющий глагол ставится в настоящем и прошедшем времени (2 варианта).

прош – при замене данного словосочетания глаголом этот глагол ставится в прошедшем времени.

безл – при замене данного словосочетания, содержащего глагол, заменяющий глагол ставится в среднем роде 3-го лица (безличная форма предложения)

%1,2 – если слово-замена является глаголом, то при замене используются два варианта формы глагола – в единственном и множественном числе [79, с. 82].

№ N - указывает номер слова или слов в заменяемом словосочетании, с которым должна быть согласована замена. «№ 0» означает слово, которое стоит перед заменяемым словосочетанием. Используется также знак ~ . Он означает аналогичную функцию по отношению к слову, предшествующему заменяющему.

<N> - в заменяющем словосочетании из нескольких слов варьируется только форма слова с номером N (пометка только для доминанты).

Одновременное наличие метки № Ni,Nj у доминанты и метки № Nk,Nl у словосочетания означает, что форма слова доминанты с номером Ni определяется формой слова с номером Nk в заменяемом словосочетании. Форма слова доминанты с номером Nj определяется формой слова с номером Nl в заменяемом словосочетании.

>сущ – замена производится только в случае, если перед или после заменяемого словосочетания стоит существительное.

& – если доминанта – глагол, то он согласуется с подлежащим, за исключением случая будущего времени

Отметим также следующее: если База-соч содержит наречие, входящее в словарь, то его начальная форма совпадает с ним самим. В такой же форме оно включается в Базу-соч (без лемматизации).

Формат файла *Форма.txt* отличается от двух предыдущих. Он также содержит группы синонимов (как отдельных слов, так и словосочетаний). Однако члены синонимического ряда не приведены в начальную форму (см. Приложение Е). Пример элемента (синонимического ряда) из этого файла:

буквально !

в буквальном смысле слова

в прямом смысле слова

слово в слово

буква в букву

Данный файл используется для замены неизменяемых словосочетаний, которые используются только в той форме, как указано в базе, а также не требуют восстановления правильного синтаксиса.

4.2.2 Основные этапы работы системы синонимических замен

Основными этапами работы системы синонимической замены являются:

1. Лемматизация введенного для обработки текста.

2. Выделение в нем токена (слово или словосочетание), содержащегося в *Базе*.

3. Если оно не является доминантой, то программа заменяет его соответствующей доминантой.

4. Восстановление в полученном тексте правильного русского синтаксиса.

Блок-схема алгоритма реализации метода синонимических замен слов приведена на рисунке 4.3.

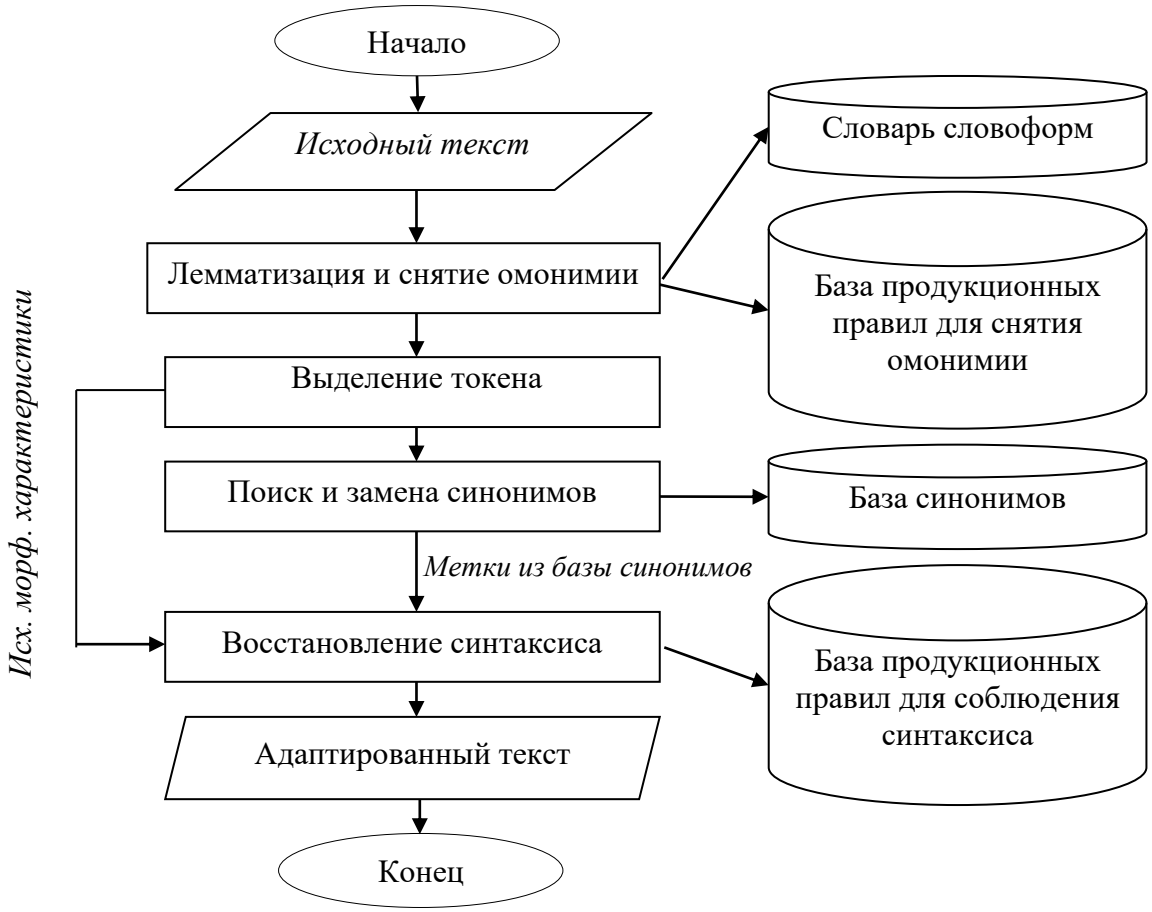

Рисунок 4.3 – Блок-схема алгоритма реализации метода синонимических замен слов

Метод синонимических замен опирается на базу продукционных правил, необходимых для соблюдения синтаксиса при заменах. Ниже приведены две базы продукций: синонимической замены отдельных слов и словосочетаний. В базе правил используются следующие обозначения: Wrd_Src – заменяемое слово;

Wrd_Repl – слово-замена; ЧР(w) – часть речи слова w; МИ(w) – переменная морфологическая информация слова w; Число(w), Пад(w), Род(w), Время(w) – число, падеж, род, время слова w соответственно.

4.2.3 Правила синонимической замены отдельных слов и неизменяемых словосочетаний

При синонимических заменах отдельных слов используется файл *База.txt*, формат которого описан выше. Система синонимических замен находит в тексте член синонимического ряда и заменяет его соответствующей доминантой, при этом восстанавливая правильный синтаксис. Более подробно правила порождения нужной формы для слова-замены при замене одиночных слов приведены в таблице 4.1. Wrd_Src и Wrd_Repl в данной таблице выделены жирным.

Таблица 4.1 – База продукций для синонимической замены отдельных слов

№ пра–вила	**Антецедент**	**Консеквент**	**Пример**
1	**(ЧР(Wrd_Repl) = (существительное ИЛИ местоимение-существительное)) И**		
1_1	ЧР(Wrd_Src) = (существительное ИЛИ местоимение-существительное)	МИ(Wrd_Repl) = МИ(Wrd_Src)	*мы столкнулись с* ***чинушей*** → *мы столкнулись с* ***бюрократом***
1_2	(Пад(Wrd_Src) = именительный) И (в предложении есть Word такое, что ЧР(Word) = (глагол ИЛИ краткое прилагательное ИЛИ причастие))	Род (Word) = Род(Wrd_Repl) И Число(Word) = Число(Wrd_Repl)	***капитал должен*** *работать на страну* → ***богатство должно*** *работать на страну*
1_3	ЧР(Word перед Wrd_Src) = прилагательное ИЛИ местоимение-прилагательное	МИ(Word) = МИ((Wrd_Repl)	*он купил новую* ***автомашину*** → *он купил новый* ***автомобиль***
2	(**ЧР(Wrd_Repl) = прилагательное ИЛИ местоимение-прилагательное ИЛИ причастие) И**		
2_1	ЧР(Wrd_Src) = (прилагательное ИЛИ местоимение-прилагательное ИЛИ причастие)	МИ(Wrd_Repl) = МИ(Wrd_Src)	*нам нужны* ***предприимчивые*** *люди* → *нам нужны* ***активные*** *люди*
2_2	ЧР(Wrd_Src) = НЕ (прилагательное ИЛИ местоимение-прилагательное ИЛИ причастие)	МИ(Wrd_Repl) = МИ(ближайшее существительное)	*пейзажи там – загляденье* → *пейзажи там красивые*
3	ЧР(Wrd_Src) = (глагол ИЛИ деепричастие)	МИ(Wrd_Repl) = МИ(Wrd_Src)	*он* ***лоботрясничал*** *весь вечер* → *он* ***бездельничал*** *весь вечер*

Продолжение таблицы 4.1

4	**(ЧР(Wrd_Repl) = числительное) И**		
4_1	ЧР(Wrd_Src) = (существительное ИЛИ местоимение-существительное)	МИ(Wrd_Repl) = МИ(Wrd_Src)	*я выбрал первые две фотографии из **дюжины** → я выбрал первые две фотографии из **двенадцати***
4_2	ЧР(Word после Wrd_Src) = существительное	МИ(Word) = МИ(Wrd_Repl)	***четверка** лошадей → **четыре** лошади*
4_3	ЧР(Word перед Wrd_Src) = (прилагательное ИЛИ местоимение-прилагательное)	(Число(Word) = множественное) И (Пад(Word 3) = Пад(Wrd_Repl))	*дадим сена нашей **четверке** лошадей → дадим сена нашим **четырем** лошадям*
4_4	(в предложении есть Word такое, что ЧР(Word) = глагол) И (Пад(Wrd_Src) = именительный)	Число(Word) = множественное	*едет **четверка** лошадей → едут **четыре** лошади*
5	(ЧР(Wrd_Src) = глагол) И (Время(Wrd_Src) = прошедшее) И (ЧР(Wrd_Repl) = наречие)	перед Wrd_Repl ставится глагол *быть* в форме *было*	*ему **нетерпелось** увидеть родных → ему **было невтерпеж** увидеть родных*

Правила первой группы (1_1-1_3) разработаны для слов-замен, которые являются существительными или местоимениями-существительными, правила второй группы (2_1-2_2) – для слов-замен, которые являются прилагательными, местоимениями-прилагательными, причастиями, для слов-замен, которые являются глаголами, разработано правило 3, правила группы 4 (4_1-4_4) – для слов-замен, которые являются числительными, правило 5 – для слов-замен наречий.

Определенная проблема при использовании правила 5 возникает, если группа, в которой доминанта является наречием, содержит слово, которое может быть как наречием, так и существительным. Пример:

отдельно !

изолированно

особняком

В предложении «*Он стоял особняком*» слово «*особняком*» выступает в функции наречия и должно заменяться словом «*отдельно*». В предложении же «*Он следит за особняком*» это слово выступает в функции существительного и указанной замены быть не должно.

Вопрос решается на основе разрешения омонимии наречие-существительное, которой выше был посвящен раздел 3.5.

Подчеркнем, что замена синонима осуществляется только тогда, когда он, не являясь доминантой, стоит в группе в начальной форме.

Например, имея в Базе синонимов группу:

десерт !

пирожное

сладкое

система синонимических замен преобразует предложение *Подадим сладкое* в *Подадим десерт*, а предложение *Подадим сладкое печенье* оставит без изменения. Причина в том, что в первом случае разработанный метод снятия омонимии (подробно описано в подразделе 3.2) идентифицирует *сладкое* как существительное, для которого в группе словаря *База.txt* есть лемма , а во втором – как словоформу прилагательного *сладкий*, для которого в этой группе леммы нет.

Отдельно реализован метод замены неизменяемых словосочетаний, которые чаще всего являются синонимическими эквивалентами наречий. В качестве словаря таких словосочетаний используется файл *Форма.txt*. Пример элемента (синонимического ряда) из этого файла:

буквально !

в буквальном смысле слова

в прямом смысле слова

слово в слово

буква в букву

Данный метод синонимических замен отличается тем, что в нем не применяется лемматизация. В базе синонимов, представленных в файле *Форма.txt*, все слова содержатся в виде необходимых словоформ, и для работы с ней исходный текст также не лемматизируется [117].

4.2.4 Алгоритм и правила синонимических замен словосочетаний

При синонимических заменах словосочетаний используется словарь, хранящийся в файле *База-соч.txt*, формат которого описан в подразделе 4.1. Обнаружив в тексте словосочетание, которое содержится в этом словаре, алгоритм заменяет его соответствующей доминантой.

Ввиду присущего русскому языку свободного порядка слов желательно, чтобы метод находил в базе словосочетание даже тогда, когда его слова встречаются в тексте в ином порядке (но подряд!). Разработанный метод синонимических замен анализирует одно за другим слова отлемматизированного исходного текста. Предложение после лемматизации и снятия омонимии представляет собой массив $P = \{lem[i]\}_{i=1}^{N}$. После обработки каждая лемма помечается, т.е. заносится в список помеченных лемм list. Алгоритм поиска заменяемого фрагмента предложения организован следующим образом.

Шаг 0. Инициализация: i = 1. Список помеченных лемм list пуст.

Шаг 1. lem[i] добавляется в list.

Шаг 2. Если существует строка $Str[j] = \{Str[i][k]\}_{k=1}^{N\_j}$ словаря *База-соч.txt*, что lem[i] – ее подстрока

то

Шаг 2.1. Если все леммы слов строки Str[j] входят в предложение P,

то

Шаг 2.1.1. Замена фрагмента предложения P от lem[i] до lem[i+N_j-1] доминантой строки Str[j] и пополнение list заменяемыми леммами

иначе

Шаг 2.1.2. Переход к следующей строке словаря j++, переход на шаг 2.

иначе

Шаг 2.2. Переход к следующей лемме i++ .

Шаг 3. Если list = P,

то

Шаг 3.1. Выход

иначе

Шаг 3.2. Переход на шаг 1.

Отметим, что алгоритм реализован таким образом, что он определяет для замены во введенном предложении возможно более длинное словосочетание.

При замене словосочетаний восстановление синтаксиса в полученном предложении намного сложнее, чем при замене отдельных слов. В этом случае необходимо определить, от какого слова в исходном тексте будет зависеть грамматическая форма слова-замены, то есть найти опорное слово. В случае, если замена состоит из одного слова, для нахождения опорного слова и восстановления правильного синтаксиса после замены прежде всего используются следующие правила [118].

Более подробно правила порождения нужной формы для слова-замены при замене словосочетаний приведены в таблице 4.2. Помимо обозначений, используемых в таблице 4.1, в таблице 4.2 введены дополнительные обозначения: Wrd_Repl – слово-замена; Phrase – заменяемое словосочетание; Wrd_Main – главное слово в словосочетании Phrase. Wrd_Repl и Phrase в таблице 4.2 выделены жирным.

Таблица 4.2 – База продукций для синонимической замены словосочетаний одним словом

№ пра–вила	**Антецедент**	**Консеквент**	**Пример**
1	**(ЧР(Wrd_Repl) = (существительное ИЛИ местоимение-существительное)) И**	**(Wrd_Main – первое по порядку существительное, местоимение-существительное или числительное в Phrase) И**	
1_1		(Число(Wrd_Repl) = Число(Wrd_Main)) И (Пад(Wrd_Repl) = Пад(Wrd_Main))	*у всех есть свои **слабые стороны**» → у всех есть свои **недостатки**,* *я нашел **десять копеек** → я нашел **гривенник***

Продолжение таблицы 4.2

1_2	(Род(Wrd_Repl) != Род(Wrd_Main)) И (Пад(Wrd_Main) = именительный) И (в предложении есть слово Word такое, что ЧР(Word) = (глагол ИЛИ краткое прилагательное ИЛИ причастие)	Род (Word) = Род(Wrd_Repl)	***нападающая сторона дезориентировала*** *противника* → ***агрессор дезориентировал*** *противника*
1_3	ЧР(Word перед Phrase) = прилагательное ИЛИ местоимение-прилагательное	(Род(Word) = Род(Wrd_Repl)) И (Пад(Wrd_Repl) = Пад(Word))	*мы гордимся* ***нашим воздушным флотом*** → *мы гордимся* ***нашей авиацией***
1_4	ЧР(Wrd_Main) = глагол И МИ_Время(Wrd_Main) = прошедшее	Замена = Word + Wrd_Repl, где лемма Word = быть И Время(Word) = прошедшее И Род(Word) = Род(Wrd_Repl)	*ему* ***птичьего молока не хватало*** → *у него* ***было изобилие***
2	**(ЧР(Wrd_Repl) = прилагательное ИЛИ местоимение-прилагательное ИЛИ причастие) И**	**(Word – главное слово для Wrd_Repl –это ближайшее существительное, местоимение-существительное или числительное не из Phrase) ИСКЛЮЧ. ИЛИ ((Wrd_Main – первое по порядку прилагательное ИЛИ причастие из Phrase) ИСКЛЮЧ. ИЛИ (Wrd_Main – первое по порядку существительное)) И**	
2_1		МИ(Wrd_Repl) = МИ(Word) ИСКЛЮЧ. ИЛИ МИ(Wrd_Repl) = МИ(Wrd_Main)	*старик* ***гол как сокол*** → *старик* ***бедный***
2_2	(ЧР(Wrd_Main) = глагол) И (Время(Wrd_Main) = (прошедшее ИЛИ будущее)) И (лемма Wrd_Main != быть)	Замена = Word + Wrd_Repl, где лемма Word = *быть* И Время(Word) = Время(Wrd_Main) И Род(Word) = Род (Wrd_Repl) И Число(Word) = Число (Wrd_Repl)	*Вопрос* ***стоял на повестке дня*** → *Вопрос* ***был актуален***
3	**(ЧР(Wrd_Repl) = глагол) И**		
3_1	(ЧР(Wrd_Main) = глагол) И (Wrd_Main ! = инфинитив)	МИ(Wrd_Repl) = МИ(Wrd_Main)	*преступника* ***посадят в тюрьму*** → *преступника* ***арестуют***, *он* ***прикован к постели*** → *он* ***болеет***
3_2	(ЧР(Wrd_Main) = глагол) И (Wrd_Main = инфинитив)	Wrd_Repl – инфинитив	*нужно* ***поднять настроение*** *публике* → *нужно* ***ободрить*** *публику*

Продолжение таблицы 4.2

3_3	(ЧР(Wrd_Main) ! = глагол) И (есть Word – существительное или местоимение-существительное в именительном падеже не из Phrase) И (нет форм вспомогательного глагола *быть*)	(Род(Wrd_Repl) = Род(Word)) И (Число(Wrd_Repl) = Число(Word)) И (Лицо(Wrd_Repl) = Лицо(Word)) И (Время(Wrd_Repl) = настоящее)	*они* ***в волнении*** → *они* ***волнуются***
3_4	(ЧР(Wrd_Main) ! = глагол) И (есть Word – существительное или местоимение-существительное в именительном падеже не из Phrase) И (есть форма вспомогательного глагола *быть* в прошедшем времени)	(Род(Wrd_Repl) = Род(Word)) И (Число(Wrd_Repl) = Число(Word)) И (Лицо(Wrd_Repl) = Лицо(Word)) И (Время(Wrd_Repl) = прошедшее) И (форма глагола *быть* удаляется)	*они были* ***в волнении*** → *они* ***волновались***
3_5	(ЧР(Wrd_Main) ! = глагол) И (есть Word – существительное или местоимение-существительное в именительном падеже не из Phrase) И (есть форма вспомогательного глагола *быть* в будущем времени)	Wrd_Repl = инфинитив	*они будут* ***в волнении*** → *они будут* ***волноваться***
4	**(ЧР(Wrd_Repl) = наречие) И**		
4_1	(ЧР(Wrd_Main) ! = глагол)	никаких преобразований не выполняется	*он выпекся* ***на славу*** → *он выпекся* ***хорошо***
4_2	(ЧР(Wrd_Main) = глагол) И (Время(Wrd_Main) = прошедшее)	Замена = *было* + Wrd_Repl	*у него* ***кровь стыла в жилах*** → *ему* ***было страшно***
4_3	(ЧР(Wrd_Main) = глагол) И (Время(Wrd_Main) = будущее)	Замена = *будет* + Wrd_Repl	(*у него* ***кровь застынет в жилах*** → *ему* ***будет страшно***
5	Лемма Wrd_Repl = сам И (есть слово Word, которое глагол не из Phrase)	(Род(Wrd_Repl) = Род(Word)) И (Число(Wrd_Repl) = Род(Word))	*они видели это* ***своими глазами»*** → *они видели это* ***сами***
6	(**ЧР(Wrd_Repl) = предикатив) И**		
6_1	(ЧР(Wrd_Main) ! = глагол)	никаких преобразований не выполняется	*это* ***проще пареной репы*** → *это* ***легко***

Продолжение таблицы 4.2

6_2	(ЧР(Wrd_Main) = глагол) И (Время(Wrd_Main)) = прошедшее) И (есть слово Noun, которое является существительным именительного падежа не из Phrase)	Замена = Word + Wrd_Repl, где лемма Word = *быть* И Время(Word) = Время(Wrd_Main) И Род(Word) = Род (Noun) И Число(Word) = Число (Noun)	*он **имел основания** сомневаться → он **был вправе** сомневаться*
6_3	(ЧР(Wrd_Main) = глагол) И (Время(Wrd_Main)) = прошедшее) И (есть слово Pronoun, которое является местоимением именительного падежа 1-го или 2-го лица не из Phrase)	Замена = Word + Wrd_Repl, где лемма Word = *быть* И Время(Word) = Время(Wrd_Main) И Род(Word) = Род(Pronoun) И Число(Word) = Число(Pronoun)	*я **имела основания** сомневаться → я **была вправе** сомневаться*

Наконец, опишем правила для замен, состоящих из нескольких слов (если нет пометки <N>). Эти правила представлены в таблице 4.3. Помимо обозначений, используемых в таблицах 4.1-4.2, в таблице 4.3 введены дополнительные обозначения: Phrase_Repl – словосочетание-замена; Verb_Repl, Noun_Repl, Adj_Repl – глагол, существительное, прилагательное в Phrase_Repl соответственно; Verb_Phrase, Noun_Phrase, Adj_Phrase – глагол, существительное, прилагательное в Phrase соответственно. Phrase_Repl и Phrase в таблице 4.3 выделены жирным.

Таблица 4.3 – База продукций для синонимической замены словосочетаний словосочетанием

№ пра–вила	**Антецедент**	**Консеквент**	**Пример**
1	**(Phrase_Repl содержит глагол (Verb_Repl) и существительное (Noun_Repl)) И**		
1_1	В Phrase есть глагол Verb_Phrase	МИ(Verb_Repl) = МИ(Verb_Phrase)	*отряд **покрыл расстояние** в шестьдесят километров → отряд **проделал путь** в шестьдесят километров*

Продолжение таблицы 4.3

1_2	В Phrase нет глагола, но есть краткое прилагательное или причастие Adj_Phrase	(Время(Verb_Repl) = настоящее) И (Число(Verb_Repl) = Число(Adj_Phrase))	*он всегда **верен своему слову** → он всегда выполняет обещание*
1_3	(В Phrase нет глагола, краткого прилагательного или причастия) И (перед Phrase стоит существительное Noun)	Лицо(Verb_Repl) = третье	*сестра **верная своему обещанию** → сестра выполняет обещание*
1_4	перед Phrase стоит местоимение Pronoun	Лицо(Verb_Repl) = Лицо(Pronoun)	*они **сыпали соль на рану** → они **бередили душу***
2	**(Phrase_Repl содержит глагол (Verb_Repl) и прилагательное/причастие (Adj_Repl)) И**		
2.1	(В Phrase есть глагол Verb_Phrase) И (В Phrase есть прилагательное Adj_Phrase)	(МИ(Verb_Repl) = МИ(Verb_Phrase)) И (МИ(Adj_Repl) = МИ(Adj_Phrase))	*он **вышел сухим из воды** → он **остался безнаказанным***
2.2	(Phrase содержит глагол Verb_Phrase) И (В Phrase нет прилагательного) И (перед Phrase стоит существительное Noun)	(МИ(Verb_Repl) = МИ(Verb_Phrase)) И (Пад(Adj_Repl) = творительный) И (Число(Adj_Repl) = Число (Noun))	*факты **получают огласку** → факты **становятся известными***
3	(Phrase_Repl содержит существительное (Noun_Repl) и прилагательное (Adj_Repl)) И (Phrase содержит существительное Noun_Phrase)	МИ(Adj_Repl) = МИ(Noun_Phrase)) И (МИ(Noun_Repl) = МИ(Noun_Phrase))	*он приглашает **артистку из погорелого театра** → он приглашает **плохую актрису***
4	(Phrase_Repl содержит глагол (Verb_Repl) с частицей «не») И (Phrase содержит глагол Verb_Phrase)	МИ(Verb_Repl) = МИ(Verb_Phrase)	*он **пустил на самотек** производство → он **не заботился** о производстве*
5	(Phrase_Repl содержит глагол (Verb_Repl) и другие части речи, которые не являются существительным, прилагательным/причастием) И (Phrase содержит глагол Verb_Phrase)	МИ(Verb_Repl) = МИ(Verb_Phrase)	*узников **вывели на свет божий** → узников **вывели наружу***

Все вышеописанные алгоритмы, опирающиеся на размеченные словари и базы продукций реализованы в едином методе синонимических замен, эффективность которого оценивалась на материалах из НКРЯ.

4.2.5 Тестирование работы системы синонимических замен на материалах Национального корпуса русского языка

Национальный корпус представляет современный русский язык во всём многообразии жанров, стилей, территориальных и социальных вариантов и т.п. Для проведения численных исследований эффективности разработанного метода упрощения текста из НКРЯ был отобран текстовый материал по 50 предложений для каждого из рассмотренных выше типов синонимической замены: отдельных слов, изменяемых и неизменяемых словосочетаний [119]. Пример работы системы лексической адаптации при замене отдельных слов представлен в таблице 4.4, а результаты работы системы при замене словосочетаний представлены в таблице 4.5.

Таблица 4.4. – Пример работы системы лексической адаптации (замена отдельных слов)

Материал из НКРЯ	**Результат работы программы после адаптации**
эта НЕРАЗБЕРИХА вокруг вашего вида спорта, наверное, мешает готовиться к стартам?	Этот БЕСПОРЯДОК вокруг вашего вида спорта, наверное, мешает готовиться к стартам?
И действительно, когда дело дошло до финала, в оркестре началась полная НЕРАЗБЕРИХА.	И действительно, когда дело дошло до КОНЦА, в оркестре начался полный БЕСПОРЯДОК.
А дома— полная НЕРАЗБЕРИХА.	А дома - полный БЕСПОРЯДОК .
РАСПУТИЦА ли помешала соседям съехаться по обыкновению к Марье Петровне, ненастье ли или другое что, а только она сидела этот раз почти без гостей.	БЕЗДОРОЖЬЕ ли помешало соседям съехаться по обыкновению к Марье Петровне , ненастье ли или другое что , а только она сидела этот раз почти без гостей.
Конечно, весенняя РАСПУТИЦА ПРЕДОПРЕДЕЛИЛА вопрос о качестве дорог.	Конечно, весеннее БЕЗДОРОЖЬЕ ПРЕДРЕШИЛО вопрос о качестве дорог.
На Руси считалось, что с этого дня начинается весенняя РАСПУТИЦА.	На Руси считалось, что с этого дня начинается весеннее БЕЗДОРОЖЬЕ.
Однажды к полковнику Тихомирову НАГРЯНУЛ дальний родственник - Сучков.	Однажды к полковнику Тихомирову ЯВИЛСЯ дальний родственник - Сучков.
По-настоящему Великий Новгород и несколько районов области вздрогнули, когда в наш регион внезапно НАГРЯНУЛ московский патруль.	По-настоящему Великий Новгород и несколько районов области вздрогнули, когда в наш регион ЯВИЛСЯ московский патруль.
- А я вас совсем другим представлял, - продолжал БАЛАГУРИТЬ Шепетуха, закуривая.	- А я вас совсем другим представлял , - продолжал ШУТИТЬ Шепетуха , закуривая .

Продолжение таблицы 4.4

Даже Станислав не нашел в себе ни сил, ни желания БАЛАГУРИТЬ.	Даже Станислав не нашел в себе ни сил, ни желания ШУТИТЬ.
Но, к сожалению, у нас нет необходимых служб, которые поддержали бы благородство и АЛЬТРУИЗМ людей, согласных отдать частицу себя ради спасения других.	Но, к сожалению, у нас нет необходимых служб, которые поддержали бы благородство и БЕСКОРЫСТИЕ людей, согласных отдать частицу себя ради спасения других.
Раз бросившись в ПРАЗДНОШАТАНИЕ, никак нельзя с ним скоро покончить.	Раз бросившись в БЕЗДЕЛЬЕ, никак нельзя с ним скоро покончить.

Таблица 4.5. – Пример работы системы лексической адаптации (замена словосочетаний)

Материал из НКРЯ	**Результат работы программы после адаптации**
Правительство Москвы приняло решение о том, чтобы не отмечать День города НА ШИРОКУЮ НОГУ.	Правительство Москвы приняло решение о том, чтобы не отмечать День города БОГАТО.
Он просто обожает работать бесплатно, ИЗ ЛЮБВИ К ИСКУССТВУ**,** и делать подарки своим друзьям.	Он просто обожает работать бесплатно, БЕСКОРЫСТНО, и делать подарки своим друзьям.
Впрочем, «подработка» – слово неправильное, денег она за это не получает, работает ИЗ ЛЮБВИ К ИСКУССТВУ, а точнее – для собственного удовольствия.	Впрочем, «подработка» – слово неправильное, денег она за это не получает, работает БЕСКОРЫСТНО, а точнее – для собственного удовольствия.
Он спустился на Петровку и привычными шагами, НЕ ОТДАВАЯ СЕБЕ ОТЧЕТА, зашел в маленькое артистическое кафе, кивнул знакомой барышне и спросил себе черного кофе с ватрушкой.	Он спустился на Петровку и привычными шагами БЕССОЗНАТЕЛЬНО зашел в маленькое артистическое кафе , кивнул знакомой барышне и спросил себе черного кофе с ватрушкой .
Эти кусты вымахали из маленьких прутиков, любовно вкопанных в землю лет двадцать назад новоселами, и с тех пор росли КАК БОГ НА ДУШУ ПОЛОЖИТ**,** забытые всеми..	Эти кусты вымахали из маленьких прутиков, любовно вкопанных в землю лет двадцать назад новоселами, и с тех пор росли БЕСПОРЯДОЧНО, забытые всеми.
В больницу она приехала В РАСТРЕПАННЫХ ЧУВСТВАХ и первым делом попросила вызвать дежурного врача, фамилия которой оказалась Соловьева.	В больницу она приехала РАССТРОЕННАЯ и первым делом попросила вызвать дежурного врача, фамилия которой оказалась Соловьева.
Как-то я В РАСТРЕПАННЫХ ЧУВСТВАХ даже разбил теннисную ракетку.	Как-то я РАССТРОЕННЫЙ даже разбил теннисную ракетку.
Как будто у тебя ДЕНЕГ КУРЫ НЕ КЛЮЮТ.	Как будто ты БОГАТЫЙ.
Но и таким людям он скажет: невелика радость будет им пережить бунт; пусть приготовятся ПОЛОЖИТЬ ЗУБЫ НА ПОЛКУ.	Но и таким людям он скажет: невелика радость будет им пережить бунт; пусть приготовятся ГОЛОДАТЬ.
Оптимистов много и сейчас; они продолжают на что-то надеяться, а у меня КОШКИ СКРЕБУТ НА СЕРДЦЕ.	Оптимистов много и сейчас; они продолжают на что-то надеяться, а я ТРЕВОЖУСЬ.

Продолжение таблицы 4.5

Ну конечно, куда этим перечницам Юдину и Кон¬стантинову?! Из них же ПЕСОК СЫПЕТСЯ. А ты, Алешка, наша молодая надежда.	Ну конечно, куда этим перечницам Юдину и Кон¬стантинову?! Они же СТАРЫЕ. А ты, Алешка, наша молодая надежда.
В общем-то, Роман – парень отнюдь не робкого десятка, но сейчас у него не то, чтобы ПОДЖИЛКИ ТРЯСЛИСЬ, нет, конечно, однако некоторая неуверенность наблюдалась.	В общем-то, Роман – парень отнюдь не робкого десятка, но сейчас он не то, чтобы ИСПУГАЛСЯ, нет, конечно, однако некоторая неуверенность наблюдалась.
Ты останешься не один, тебе будет весело. У меня КРОВЬ ЗАСТЫЛА В ЖИЛАХ. Бедная Надина!	Ты останешься не один, тебе будет весело. Я ИСПУГАЛСЯ. Бедная Надина!
Но я решил повезти его в Москву, немыслимо было его ОСТАВЛЯТЬ НА ПРОИЗВОЛ СУДЬБЫ.	Но я решил повезти его в Москву, немыслимо было о нем НЕ ЗАБОТИТЬСЯ.
И приходится через силу улыбаться и делать вид, что все это – ЧЕПУХА НА ПОСТНОМ МАСЛЕ.	И приходится через силу улыбаться и делать вид, что все это – ВЗДОР.

Для оценки эффективности разработанного метода упрощения текста полученные на выходе предложения анализировались по параметрам: сохранение семантики упрощенного текста; соблюдение в нем правильного синтаксиса упрощенного текста; упрощение – простота восприятия обработанного текста. Последний параметр оценивался субъективно [120]. Полученные оценки приведены в таблице 4.6.

Таблица 4.6 – Численные исследования эффективности метода лексической адаптации

Используемый словарь	**Кол-во предложений**	**Сохранение семантики**	**Соблюдение правильного синтаксиса**	**Упрощение**
Словарь отдельных слов (*База.txt*)	50	96%	96%	100%
Словарь словосочетаний (*База-соч.txt*)	50	98%	96%	100%
Словарь неизменяемых словосочетаний (*Форма.txt*)	50	98%	98%	100%
Всего	150	97%	97%	100%

Результаты тестирования позволяют утверждать, что разработанный метод лексической адаптации позволяет успешно осуществлять замену отдельных слов и словосочетаний в тексте с соблюдением правильного синтаксиса и сохранением семантики текста точностью выше 96%, что говорит о полноте базы

продукционных правил, позволяющих соблюдать правильный синтаксис упрощенного предложения, а также о хорошей наполненности размеченных словарей синонимов, сохраняющих семантику адаптированного текста.

4.3 Разбиение текста на семантически однородные фрагменты (абзацы)

Упрощение текста разумно выполнять, предварительно разбив его на тематически однородные фрагменты, оформленные в виде абзацев, и осуществлять упрощение в пределах каждого абзаца. В работе предложен алгоритм разбиения сплошного текста на абзацы, основанный на использовании часто встречающихся существительных (назовем их ключевыми словами) и местах их концентрации в тексте. Блок-схема алгоритма приведена на рисунке 4.4.

Формируется список лемм всех слов обрабатываемого фрагмента, которые являются существительными и встречаются в тексте не менее двух раз (условие L).

Количество вхождений словоформ, входящих в парадигму определенной леммы, будем называть частотой слова, которую обозначим через a. Помимо этого фиксируются номера первого и последнего предложения, где встречается слово: b и c соответственно. После чего вычисляется отношение

$$a/(c-b+1) \qquad (4.1)$$

Отношение (4.1) тем больше, чем больше частота слова и чем меньше отрезок, на котором это слово сосредоточено.

Алгоритм находит (первый) максимум отношения (4.1), определяет соответствующие номера b и c и выделяет в качестве абзаца отрезок текста от предложения с номером b до предложения с номером c включительно. Затем проводится подсчет количества предложений в образовавшихся предыдущих и последующих фрагментах текста (не входящих в уже выделенные абзацы) и к наибольшему из них при выполнении условия L применяется вышеописанная процедура. Если для него условие L не выполняется, то же делается для второго по размеру из образовавшихся фрагментов. Процесс продолжается до тех пор пока после выделения очередного абзаца для всех оставшихся фрагментов перестает выполняться условие L.

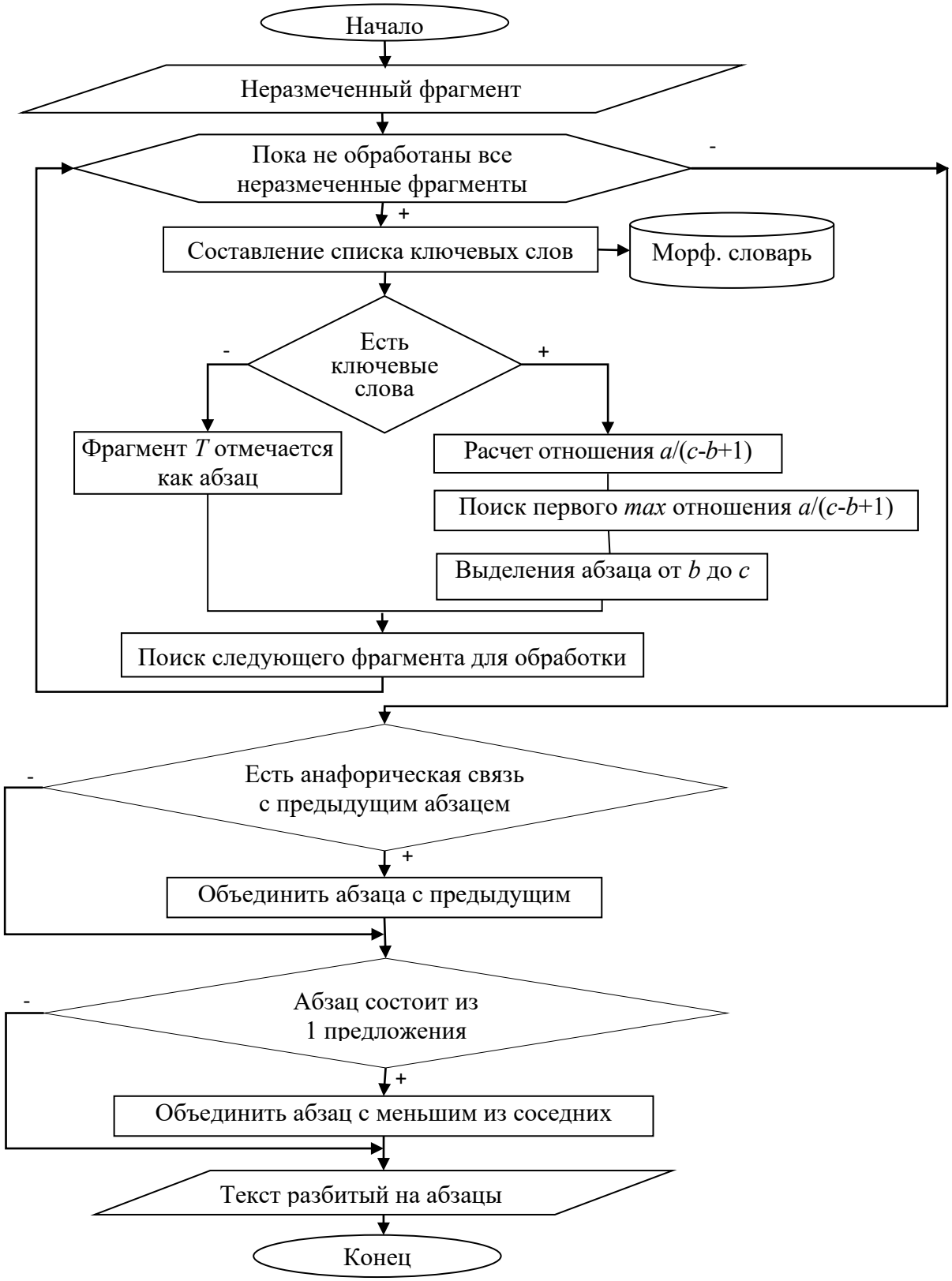

Рисунок 4.4. – Общая схема работы алгоритма разбиения текста на семантически однородные фрагменты

Следующий этап – анализ полученного разбиения на абзацы на предмет анафор и абзацев, состоящих их одного предложения.

Алгоритм анализирует те абзацы, которые в первом своем предложении содержат личные местоимения и указательные слова: *он* (*-а*; *-о*); э*тот*; *тот*; *это*, *там*; *туда*; *оттуда*; *столько*, поскольку именно они имеют анафорическую функцию. Если такое предложение найдено и в нем нет предшествующего существительного в том же роде и числе, то абзац присоединяется к предыдущему. Объединяются также соседние абзацы, на стыке которых оказался общий фрагмент прямой речи. При этом, если начало прямой речи выделено с помощью тире, то она, естественно, оформляется отдельным абзацем. К сожалению, конец этого абзаца пока приходится отмечать вручную. Наконец, абзац, состоящий из одного предложения, присоединяется к меньшему из соседних.

Проиллюстрируем этапы работы предложенного алгоритма на примере 15-ой главы романа Даниеля Дефо «Робинзон Крузо» в пересказе К.И. Чуковского. Вот фрагмент этой главы, представленный в виде сплошного текста:

«Конечно, было бы хорошо иметь лодку на этой стороне острова, поближе к моему дому, но как привести ее оттуда, где я оставил ее? Обогнуть мой остров с востока – от одной мысли об этом у меня сжималось сердце и холодела кровь. Как обстоит дело на другой стороне острова, я не имел никакого понятия. Что, если течение по ту сторону такое же быстрое, как и по эту? Разве не может оно швырнуть меня на прибрежные скалы с той же силой, с какой другое течение уносило меня в открытое море. Словом, хотя постройка этой лодки и спуск ее на воду стоили мне большого труда, я решил, что все же лучше остаться без лодки, чем рисковать из-за нее головой. Нужно сказать, что теперь я стал гораздо искуснее во всех ручных работах, каких требовали условия моей жизни. Когда я очутился на острове, я совершенно не умел обращаться с топором, а теперь я мог бы при случае сойти за хорошего плотника, особенно если принять в расчет, как мало было у меня инструментов. Я и в гончарном деле (совсем неожиданно!) сделал большой шаг вперед: устроил станок с вертящимся кругом, отчего моя работа стала и быстрее и лучше; теперь вместо корявых изделий, на которые было противно смотреть, у меня выходила очень неплохая посуда довольно правильной формы. Но никогда я, кажется, так не радовался и не гордился своей изобретательностью, как в тот день, когда мне удалось сделать трубку. Конечно, моя трубка была первобытного вида – из простой обожженной глины, как и все мои гончарные изделия, и вышла она не очень красивой. Но она была достаточно крепка и хорошо пропускала дым, а главное – это была все-таки трубка, о которой я столько мечтал, так как привык курить с очень давнего времени. На нашем корабле были трубки, но, когда я перевозил оттуда вещи, я не знал, что на острове растет табак, и решил, что не стоит их брать. К этому времени я обнаружил, что мои запасы пороха начинают заметно убывать. Это чрезвычайно встревожило и огорчило меня, так как нового пороха достать было неоткуда. Что же я буду

делать, когда у меня выйдет весь порох? Как я буду тогда охотиться на коз и птиц? Неужели я до конца моих дней останусь без мясной пищи?» [121]

В этом фрагменте текста содержится 18 предложений. Ниже приведена таблица 4.7, в которой приведены результаты вычислений. Во втором столбце таблицы после леммы указана частота (параметр *a*), затем в скобках приведены номера первого и последнего предложения, где встречается слово (параметры *b* и *c*). В предпоследнем столбце выведены значения отношения (4.1), в последнем – первый максимум этой величины, отмеченный тремя звездочками.

Таблица 4.7 – Результаты вычислений

Лемма	***a***	**(b, c)**	**a/(c-b+1)**	**Max a/(c-b+1)**
1 шаг – фрагмент [1;18]				
остров	5	(1, 13)	0.38	
трубка	4	(10, 13)	1.00	***
лодка	3	(1, 6)	0.50	
порох	3	(14, 16)	1.00	
время	2	(12, 14)	0.67	
дело	2	(3, 9)	0.29	
день	2	(10, 18)	0.22	
изделие	2	(9, 11)	0.67	
работа	2	(7, 9)	0.67	
сторона	2	(1, 3)	0.67	
течение	2	(4, 5)	1.00	
2 шаг – фрагмент [1;9])				
остров	4	(1, 8)	0.50	
лодка	3	(1, 6)	0.50	
дело	2	(3, 9)	0.29	
работа	2	(7, 9)	0.67	
сторона	2	(1, 3)	0.67	
течение	2	(4, 5)	1.00	***
3 шаг – фрагмент [6;9])				
лодка	2	(6, 6)	2	***
работа	2	(7, 9)	0.67	
4 шаг – фрагмент [7;9]				
работа	2	(7, 9)	0.67	***
5 шаг – фрагмент [14;18]				
порох	3	(14, 16)	1.00	***
6 шаг – фрагмент [1;3]				
остров	3	(1, 3)	1,5	***
сторона	2	(1, 3)	0.67	
7шаг – фрагмент [17;18]				
-	-	-	-	-

В результате алгоритм для приведенного фрагмента выдал следующий результат:

«Конечно, было бы хорошо иметь лодку на этой стороне острова, поближе к моему дому, но как привести ее оттуда, где я оставил ее? Обогнуть мой остров с востока – от одной мысли об этом у меня сжималось сердце и холодела кровь. Как обстоит дело на другой стороне острова, я не имел никакого понятия.

Что, если ТЕЧЕНИЕ по ту сторону такое же быстрое, как и по эту? Разве не может оно швырнуть меня на прибрежные скалы с той же силой, с какой другое ТЕЧЕНИЕ уносило меня в открытое море.

Словом, хотя постройка этой ЛОДКИ и спуск ее на воду стоили мне большого труда, я решил, что все же лучше остаться без ЛОДКИ, чем рисковать из-за нее головой.

Нужно сказать, что теперь я стал гораздо искуснее во всех ручных работах, каких требовали условия моей жизни. Когда я очутился на острове, я совершенно не умел обращаться с топором, а теперь я мог бы при случае сойти за хорошего плотника, особенно если принять в расчет, как мало было у меня инструментов. Я и в гончарном деле (совсем неожиданно!) сделал большой шаг вперед: устроил станок с вертящимся кругом, отчего моя работа стала и быстрее и лучше; теперь вместо корявых изделий, на которые было противно смотреть, у меня выходила очень неплохая посуда довольно правильной формы.

Но никогда я, кажется, так не радовался и не гордился своей изобретательностью, как в тот день, когда мне удалось сделать ТРУБКУ. Конечно, моя ТРУБКА была первобытного вида – из простой обожженной глины, как и все мои гончарные изделия, и вышла она не очень красивой. Но она была достаточно крепка и хорошо пропускала дым, а главное – это была все-таки ТРУБКА, о которой я столько мечтал, так как привык курить с очень давнего времени. На нашем корабле были ТРУБКИ, но, когда я перевозил оттуда вещи, я не знал, что на острове растет табак, и решил, что не стоит их брать.

К этому времени я обнаружил, что мои запасы ПОРОХА начинают заметно убывать. Это чрезвычайно встревожило и огорчило меня, так как нового ПОРОХА достать было неоткуда. Что же я буду делать, когда у меня выйдет весь ПОРОХ?

Как я буду тогда охотиться на коз и птиц? Неужели я до конца моих дней останусь без мясной пищи?» [121]

4.4 Автоматическое создание элемента плана текста. Использование отглагольных существительных

Опишем предлагаемый метод программного выявления смысла с помощью 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. При этом игнорируются характеристики глагола, такие как время, лицо, число. Остается только обозначение самого действия. Разработанный

метод заменяет глагол соответствующим отглагольным существительным и винительный падеж существительного (прямое дополнение) родительным.

Например, результатом работы метода с предложением: «По телевидению передают важное сообщение» является словосочетание «Передача сообщения». Оно и является носителем основной информации. Например, для предложения «Газета передала любопытное сообщение из города Воронежа» результат будет тем же самым.

Если упомянутое существительное стоит в именительном падеже, то допускаются соответствующие возвратные глаголы, в нашем примере глагола «передаваться».

Программа использует файл *Гл-сущ.txt.* Он состоит из групп, каждая из которых содержит две строки. Первая включает набор глаголов, а вторая – соответствующее отглагольное существительное. Группы разделены пробельными строками. Для демонстрации приведем фрагмент этого словаря:

```
абстрагировать,абстрагироваться
абстрагирование

авансировать,заавансировать,авансироваться,проавансировать
авансирование

автоматизировать,автоматизироваться
автоматизация

авторизовать,авторизоваться
авторизация

агитировать,заагитировать,поагитировать,разагитировать,сагитировать
агитация

агукать
агуканье

адаптировать,адаптироваться
адаптация

акклиматизировать,акклиматизироваться
акклиматизация

и т. д.
```

Получившийся текстовый корпус – словарь глаголов и отглагольных существительных имеет объем более 10 000 групп и представляет собой некоторый самостоятельный лингвистический продукт.

Пример работы программы автоматического создания плана текста приведен на рисунке 4.5.

Остальные подразделы посвящены описанию алгоритмов, словарей и баз продукций модуля лексической адаптации, работа которого невозможна без размеченной базы синонимов и набора правил соблюдения синтаксиса, лежащих в основе метода синонимических замен, который осуществляет лексическую замену слов и словосочетаний с соблюдением правильного синтаксиса и сохранением семантики текста.

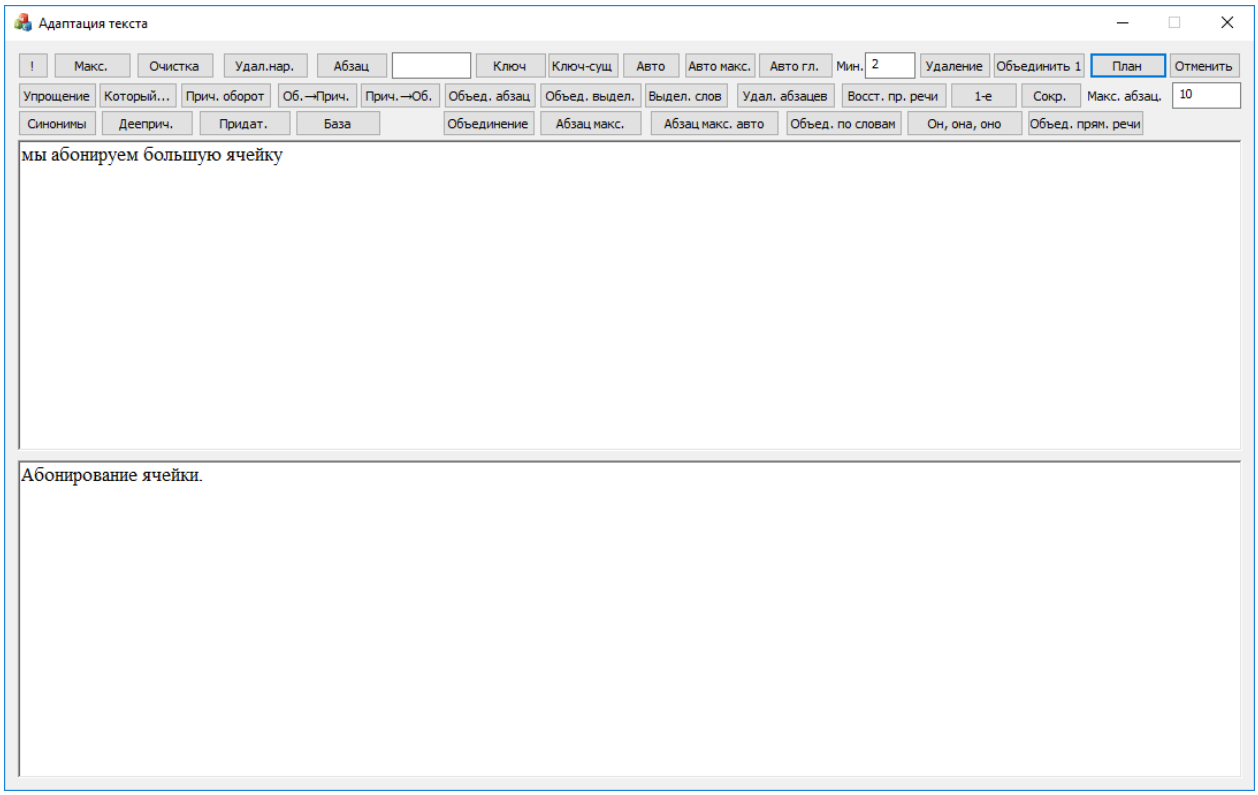

Рисунок 4.5 – Пример работы программы автоматического создания плана текста

4.5 Выводы к разделу 4

В рамках диссертационной работы для лексической адаптации текста на основе данных из открытых источников, представляющих собой словари синонимов и частотные словари русского языка, сформирован текстовый корпус – размеченная база синонимов. База состоит из трех частей, использующихся для синонимических замен отдельных слов, изменяемых и неизменяемых словосочетаний. Объем созданной размеченной базы синонимов составляет около 32 тыс. слов, содержащихся в 11 тыс. синонимических рядов.

При создании корпуса:

– проведен анализ и сокращение синонимических рядов из используемых словарей для сохранения семантики;

– для каждой группы синонимов проанализирована частотность доминанты и членов синонимического ряда, и при возможности в качестве доминанты выбран синоним с наибольшей частотностью;

– проведена разметка записей в базе синонимов, предложенный механизм обработки меток позволяет соблюдать правила синтаксиса в упрощенном тексте [79, с. 81].

Разработана база продукционных правил для сохранения правильного синтаксиса после лексической адаптации текста, состоящая из:

– базы продукций для синонимической замены отдельных слов;

– базы продукций для синонимической замены словосочетаний одним словом;

– базы продукций для синонимической замены словосочетаний словосочетанием.

Проведенные численные исследования на материалах НКРЯ 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%.

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.

Для построения элемента плана текста сформирован текстовый корпус – словарь отглагольных существительных, объемом более 10 000 групп, содержащих глаголы и соответствующие им существительные. Элемент плана текста получается заменой глагола в предложении соответствующим отглагольным существительным и винительного падежа существительного, являющегося прямым дополнением, родительным [79, с.82].

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 ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Русская грамматика. – Москва, 1980. – Т.1. – С. 8. – Текст : непосредственный.

2. Буриева, М. О симплификации языка в интернет-пространстве // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2021. № 5 (847). URL: https://cyberleninka.ru/article/n/o-simplifikatsii-yazyka-v-internet-prostranstve (дата обращения: 06.03.2025). – Текст : электронный.

3. Козинец, С. Б. Системные отношения в русском языке: Учебно-методическое пособие / С.Б. Козинец. – Саратов: ГАУ ДПО «СОИРО», 2019. – 64 с. – Текст : непосредственный.

4. Sikka, P. A Survey on Text Simplification / P. Sikka, V.K. Mago. 2020. URL: https://arxiv.org/abs/2008.08612 – DOI 10.48550/arXiv.2008.08612 (дата обращения: 06.03.2025). – Текст : электронный.

5. Способы упрощения текстов: плюсы, минусы, альтернативы // habr.com URL: https://habr.com/ru/articles/581526/ (дата обращения: 06.03.2025). – Текст : электронный.

6. Дмитриева, А. А. Количественное исследование стратегий упрощения в адаптированных текстах для изучающих русский язык на уровне L2 / А. А. Дмитриева, А. Н. Лапошина, М. Ю. Лебедева // Компьютерная лингвистика и интеллектуальные технологии : По материалам ежегодной международной конференции «Диалог» (2021), Москва, 16–19 июня 2021 года. Выпуск 20. – Москва: Российский государственный гуманитарный университет, 2021. – С. 191-203. – DOI 10.28995/2075-7182-2021-20-191-203. – EDN XVDVYJ. – Текст : непосредственный.

7. Inui K. [et al.] Text simplification for reading assistance: a project note // Proceedings of the second international workshop on Paraphrasing-Volume 16. Association for Computational Linguistics, 2003. С. 9-16.

8. Chandrasekar R., Srinivas B. Automatic induction of rules for text simplification // Knowledge-Based Systems. 1997. Т. 10. №3. С. 183-190.

9. Siddharthan A. A survey of research on text simplification // ITL-International Journal of Applied Linguistics. 2014. Т. 165. № 2. С. 259-298.

10. Использование метрики BLEU для оценки естественности текста лингвистических стегосистем / К. А. Ахрамеева, Е. Ю. Герлинг, Д. Ю. Мицковский, С. В. Прудников // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. – 2020. – № 2. – С. 73-80. – DOI 10.25586/RNU.V9187.20.02.P.073. – EDN EMCSGR. – Текст : непосредственный.

11. Petersen S. E., Ostendorf M. Text simplification for language learners: a corpus analysis // SLaTE. 2007. С. 69-72.

12. Coster, W., & Kauchak, D. Simple English Wikipedia: A New Text Simplification Task. Annual Meeting of the Association for Computational Linguistics. 2011.

13. Woodsend, K., & Lapata, M. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. Conference on Empirical Methods in Natural Language Processing. 2011.

14. De Belder J., Deschacht K., Moens M.F. Lexical simplification // Proceedings of ITEC2010: 1st international conference on interdisciplinary research on technology, education and communication. 2010.

15. De Belder J., Moens M.F. Text simplification for children // Prroceedings of the SIGIR workshop on accessible search systems. ACM, 2010. С. 19-26.

16. Kim, Y., Hullman, J.R., & Adar, E. DeScipher: A Text Simplification Tool for Science Journalism. 2015.

17. Lu, J., Li, J., Wallace, B.C., He, Y., & Pergola, G. NapSS: Paragraph-level Medical Text Simplification via Narrative Prompting and Sentence-matching Summarization. Findings. 2023.

18. Cripwell, L., Legrand, J., & Gardent, C. Document-Level Planning for Text Simplification. Conference of the European Chapter of the Association for Computational Linguistics. 2023.

19. Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and

Biljana Drndarevic. 2015. Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. ACM Trans. Access. Comput. 6, 4, Article 14 (June 2015), 36 pages. URL: https://doi.org/10.1145/2738046 (дата обращения: 26.06.2023).

20. Bott S., Saggion H., Mille S. Text Simplification Tools for Spanish // LREC. 2012. С. 1665-1671.

21. Сибирцева, В. Г. Автоматическая адаптация текстов для электронных учебников / Сибирцева В. Г., Карпов Н. В. // Новая русистика. – 2014. – №7. – С.19-33. – Текст : непосредственный.

22. Сибирцева, В. Г. Национальный корпус русского языка как основа новаторских электронных учебников / Сибирцева В. Г., Хоменко А. Ю., Баранова Ю. Н. – Образовательные технологии и общество. – Т . 16, № 3. – 2013. – С. 508-520. – Текст : непосредственный.

23. Burstein J. [et al.] The automated text adaptation tool // Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. – Association for Computational Linguistics, 2007. С. 3-4.

24. Martin, L., Fan, A., Villemonte de la Clergerie, E., Bordes, A., & Sagot, B. 2020. Multilingual Unsupervised Sentence Simplification. ArXiv, abs/2005.00352.

25. Liu, K., & Qiang, J. 2023. Sentence Simplification Using Paraphrase Corpus for Initialization. ArXiv, abs/2305.19754.

26. Lu, X., Qiang, J., Li, Y., Yuan, Y., & Zhu, Y. An Unsupervised Method for Building Sentence Simplification Corpora in Multiple Languages. Conference on Empirical Methods in Natural Language Processing. 2021.

27. Qiang, J., & Wu, X. Unsupervised Statistical Text Simplification. IEEE Transactions on Knowledge and Data Engineering, 33, 1802-1806. 2019.

28. Qiang, J., Li, Y., Zhu, Y., Yuan, Y., & Wu, X. Lexical Simplification with Pretrained Encoders. AAAI Conference on Artificial Intelligence. 2019.

29. Qiang, J., Liu, K., Li, Y., Yuan, Y., & Zhu, Y. 2023. ParaLS: Lexical Substitution via Pretrained Paraphraser. ArXiv, abs/2305.08146.

30. Katsuta, A., & Yamamoto, K. Improving text simplification by corpus expansion with unsupervised learning. 2019 International Conference on Asian Language Processing (IALP), 216-221.

31. Aprosio, A. P., Tonelli, S., Turchi, M., Negri, M., & Gangi, M. A. Neural Text Simplification in Low-Resource Conditions Using Weak Supervision. Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation. 2019.

32. Štajner, S. Automatic Text Simplification for Social Good: Progress and Challenges. Findings. 2021.

33. Agrawal, S., Xu, W., & Carpuat, M. A Non-Autoregressive Edit-Based Approach to Controllable Text Simplification. Findings. 2021.

34. Omelianchuk, K., Raheja, V., & Skurzhanskyi, O. Text Simplification by Tagging. Workshop on Innovative Use of NLP for Building Educational Applications. 2021.

35. Qiang, J., Li, Y., Zhu, Y., Yuan, Y., Shi, Y., & Wu, X. LSBert: Lexical Simplification Based on BERT. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3064-3076. 2021.

36. Zhang, X., & Lapata, M. 2017. Sentence Simplification with Deep Reinforcement Learning. ArXiv, abs/1703.10931.

37. Jordan Clive, Kris Cao, and Marek Rei. 2022. Control Prefixes for Parameter-Efficient Text Generation. In Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 363–382, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

38. Jiang, C., Maddela, M., Lan, W., Zhong, Y., & Xu, W. 2020. Neural CRF Model for Sentence Alignment in Text Simplification. ArXiv, abs/2005.02324.

39. Vu, T., Hu, B., Munkhdalai, T., & Yu, H. Sentence Simplification with Memory-Augmented Neural Networks. North American Chapter of the Association for Computational Linguistics. (2018).

40. Jurafsky, D., Martin, J. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

Third Edition draft. URL: http://www.web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf (дата обращения: 06.03.2025).

41. Пруцков А. В. Модели, методы и программы автоматической обработки форм слов в естественно-языковых интерфейсах : дис. ... доктора технических наук : 05.13.11 / Пруцков Александр Викторович; Место защиты: Рязан. гос. радиотехн. ун-т. – Рязань, 2015. – 279 с. – Текст : непосредственный.

42. Солодуб, Ю. П. Современный русский язык. Лексика и фразеология : учебник для студентов филологических факультетов и факультетов иностранных языков / Ю. П. Солодуб, Ф. Б. Альбрехт. – Москва, 2002. – С. 102. – Текст : непосредственный..

43. Гатауллин, Р. Р. Аналитический обзор методов разрешения морфологической многозначности / Р. Р. Гатауллин // Электронные библиотеки. – 2016. – Т. 19, № 2. – С. 98-114. – EDN YGKILF. – Текст : непосредственный.

44. Сокирко, А. В. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка / А. В. Сокирко, С. Ю. Толдова. URL: http://aot.ru/docs/RusCorporaHMM.htm (дата обращения 06.03.2025). – Текст : электронный..

45. Зеленков, Ю. Г. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов / Зеленков Ю. Г., Сегалович И. В., Титов В. А. // Сборник трудов Международной конференции «Диалог-2005». – Москва : Наука, 2005. – С. 616-638. – Текст : непосредственный.

46. Лесько, О. Н. Использование онтологии предметной области для снятия омонимии в естественно-языковых текстах / О. Н. Лесько, Ю. В. Рогушина // Проблеми програмування. – 2017. – № 2. – С. 61-71. – Текст : непосредственный.

47. Schmid, H. Probabilistic part-of-speech tagging using decision trees. In.: Proceedings of the international conference on new methods in language processing. 1994. P. 44-49.

48. Итеративное применение алгоритмов снятия частеречной омонимии в русском тексте / Епифанов М. Е., Антонова А. Ю., Баталина А. М. [и др.] // Компьютерная лингвистика и интеллектуальные технологии – труды Международной конференции «Диалог-2010». – Т. 9(16). – С. 119-123. – Текст : непосредственный.

49. Автоматическая обработка текстов на естественном языке и анализ данных : учеб. пособие / Большакова Е. И., Воронцов К. В., Ефремова Н. Э., Клышинский Э. С., Лукашевич Н. В., Сапин А. С. – Москва : Изд-во НИУ ВШЭ, 2017. – 269 с. – Текст : непосредственный.

50. Зинькина, Ю. В. Разрешение функциональной омонимии в русском языке на основе контекстных правил / Зинькина Ю. В., Пяткин Н. В., Невзорова О. А. // Сборник трудов Международной конференции «Диалог-2005». – Москва : Наука, 2005. – С. 198-202. – Текст : непосредственный.

51. Анисимов, А. В. Создание управляющего пространства синтаксических структур естественного языка / Анисимов А. В., Марченко О. О. Нагорный В. А. // Вестник Киевского университета, серия «Физико-математические науки». – 2011. – Выпуск 1. – С. 159-169. – Текст : непосредственный.

52. Лукашевич, Н. В. Тезаурусы в задачах информационного поиска / Н. В. Лукашевич. – Москва, 2010. – 396 с. – Текст : непосредственный.

53. Марченко, А. А. Контекстный семантический анализ текста. Система текстового мониторинга и качественного оценивания фокусного объекта / Марченко А. А., Никоненко А. А. // Искусственный интеллект. – 2008. – № 3. – С. 808–813. – Текст : непосредственный.

54. Brill E. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging // Computational Linguistics. 2002. Vol. 21, N 4. P. 543–565.

55. Сокирко, А. В. Быстрословарь: предсказание морфологии русских слов с использованием больших лингвистических ресурсов / А. В. Сокирко // Труды международной конференции «Диалог-2010. Компьютерная лингвистика и

интеллектуальные технологии». – Москва : Издательский центр РГГУ, 2010. – С. 450. – Текст : непосредственный.

56. Зализняк, А. А. Грамматический словарь русского языка / А. А. Зализняк. – Москва, Русский язык, 1980. – Текст : непосредственный.

57. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, pp. 320-332. 2015.

58. Порохнин, А. А. Анализ статистических методов снятия омонимии в текстах на русском языке / А. А. Порохнин // Вестник АГТУ. Серия: Управление, вычислительная техника и информатика. 2013. № 2. URL: https://cyberleninka.ru/article/n/analiz-statisticheskih-metodov-snyatiya-omonimii-v-tekstah-na-russkom-yazyke (дата обращения: 06.03.2025). – Текст : электронный.

59. Ермоленко, Т. В. Разработка алгоритмов и языковых моделей для мультиязычной системы автоматического аннотирования текстов разных жанров / Т. В. Ермоленко, В. И. Бондаренко, Я. С. Пикалев // Вестник Донецкого национального университета. Серия Г: Технические науки. – 2023. – № 2. – С. 22-43. – EDN KRDDOO. – Текст : непосредственный.

60. Mikulov, T. Recurrent neural network based language model / T. Mikulov, M. Karafiát, L. Burget, C. Jan, S. Khudanpur // Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010. 2010.

61. Hochreiter, S. Long Short-Term Memory / S. Hochreiter, J. Schmidhuber // Neural Computation. 1997.

62. Lourentzou, I. Adapting sequence to sequence models for text normalization in social media / I. Lourentzou, K. Manghnani, C.X. Zhai // Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019. 2019.

63. Alammar, J. The Illustrated GPT-2 (Visualizing Transformer Language Models). – URL: http://jalammar.github.io/illustrated-gpt2/ (дата обращения: 06.03.2025). – Текст : электронный.

64. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proc. NAACL-HLT, 2019, pp. 4171–4186.

65. Васильев, Д. Д. Использование языковых моделей T5 для задачи упрощения текста / Васильев Д. Д., Пятаева А. В. // Программные продукты и системы. – 2023. – Т. 36, № 2. – С. 228–236. – doi: 10.15827/0236-235X.142.228-236. – Текст : непосредственный.

66. Monteiro, J. C. Using a Pre-trained SimpleT5 Model for Text Simplification in a Limited Corpus / J. C. Monteiro, M. M. A. Aguiar, S. Araújo // Conference and Labs of the Evaluation Forum. 2022. pp. 2826–2831.

67. Raffel C., Shazeer N., Roberts A., Lee K., Narang S. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J. of Machine Learning Research, 2019, vol. 21, pp. 5485–5551.

68. Liu Y., Gu J., Goyal N., Li X., Edunov S. [et al.] Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 2020, vol. 8, pp. 726–742. doi: 10.1162/tacl_a_00343.

69. Brown, T. B., Mann B., Ryder N. [et al.]. Language models are few-shot learners. Proc. NeurIPS, 2020, pp. 1877–1901.

70. RuSimpleSentEval-2021 Shared Task: Evaluating Sentence Simplification for Russian / A. Sakhovskiy, E. Tutubalina, V. Malykh [et al.] // Computational Linguistics and Intellectual Technologies : Papers from the Annual International Conference «Dialogue», Moscow, June 16–19, 2021. Iss. 20. – Moscow: Russian state university for the humanities, 2021. P. 607-617. DOI 10.28995/2075-7182-2021-20-607-617.

71. Shatilov, A. A. Sentence simplification with ruGPT3 / A. A. Shatilov, A. I. Rey // Computational Linguistics and Intellectual Technologies : Papers from the Annual International Conference “Dialogue” (2021), Moscow, June 16–19, 2021. Iss. 20. Moscow: Russian state university for the humanities, 2021. P. 618-625. DOI 10.28995/2075-7182-2021-20-618-625.

72. Komleva, E. P. Sentence Simplification for Russian using Transfer Learning / E. P. Komleva, D. G. Anastasyev // Computational Linguistics and Intellectual Technologies, June 16–19, 2021. Iss. 20. S. Russian state university for the humanities, 2021. P. 1075-1080. DOI 10.28995/2075-7182-2021-20-1075-1080.

73. Fenogenova, A. Text Simplification with Autoregressive Models / A. Fenogenova // Computational Linguistics and Intellectual Technologies : Papers from the Annual International Conference “Dialogue” (2021), Moscow, 16–19 июня 2021 года. Vol. Выпуск 20. – Moscow: Российский государственный гуманитарный университет, 2021. – P. 227-234. – DOI 10.28995/2075-7182-2021-20-227-234. – Текст : непосредственный.

74. Vasil'ev D. D., Pyataeva A.V. Ispol'zovanie yazykovyh modelej T5 dlya zadachi uproshcheniya teksta [Using T5 Language Models for Text Simplification Task] // Programmnye produkty i sistemy. 2023. Vol. 36. № 2. P. 228–236. DOI: 10.15827/0236-235X.142.228-236.

75. Ниценко, А. В. Автоматическая лексическая адаптация русскоязычных текстов / Ниценко А. В., Шелепов В. Ю., Большакова С. А. // Журнал «Искусственный интеллект и принятие решений». – 2025. – № 1. – С. 95–107. – ISSN 2413-7383. – Текст : непосредственный.

76. Хаген, М. А. Полная парадигма. Морфология. URL: http://www.speakrus.ru/dict/#morph-paradigm (дата обращения: 10.12.2021). – Текст : электронный.

77. 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. – ISSN 2413-7383. – Текст : непосредственный.

78. Ниценко, А. В. Разделение сплошного текста на слова / А. В. Ниценко, В. Ю. Шелепов, С. А. Большакова // Проблемы искусственного интеллекта. – 2018. – № 3 (10). – С. 94–103. – ISSN 2413-7383. – Текст : непосредственный.

79. Ниценко, А. В. О некоторых подходах к проблеме автоматической адаптации русскоязычных текстов / А. В. Ниценко, В. Ю. Шелепов //

Программная инженерия: методы и технологии разработки информационно-вычислительных систем (ПИИВС-2020) : сборник научных трудов III Международной научно-практической конференции, Донецк, 25–26 ноября 2020 года. Т. 1. – Донецк: Донецкий национальный технический университет, 2020. – С. 77-83. – EDN XRNZWU. – Текст : непосредственный.

80. Ниценко, А. В. Об автоматическом построении дерева синтаксического подчинения / Ниценко А. В., Шелепов В. Ю. // XII Мультиконференция по проблемам управления (МКПУ-2019): Материалы XII мультиконференции (г. Геленджик, 23-28 сентября 2019 г.). Т. 1. – Ростов н/Д. : ЮФУ, 2019. – С. 119–121. – Текст : непосредственный.

81. Ниценко, А. В. О подчинительном дереве для простого распространенного русского предложения / А. В. Ниценко, В. Ю. Шелепов, С. А. Большакова // Проблемы искусственного интеллекта. – 2019. – № 2 (13). – С. 63–73. – ISSN 2413-7383. – Текст : непосредственный.

82. Циммерлинг, А. В. Предикативы и качественные наречия: классы слов и направления деривации / А. В. Циммерлинг // Русистика на пороге XXI века: проблемы и перспективы: материалы международной конференции. – Москва, 2003. – С. 54–59. – Текст : непосредственный.

83. Ниценко, А. В. К вопросу об автоматическом снятии омонимии русских предикативов / А. В. Ниценко, В. Ю. Шелепов, С. А. Большакова. // Сборник трудов VIII Международной конференции «Знания-Онтологии-Теории» (г. Новосибирск, 8-12 ноября 2021г.). – С. 217-224. – Текст : непосредственный.

84. Большакова, С. А. К вопросу о снятии омонимии в некоторых группах омонимов, включающих предикатив / С. А. Большакова, А. В. Ниценко, В. Ю. Шелепов // Донецкий международный круглый стол «Искусственный интеллект: теоретические аспекты и практическое применение» (ИИ-2022). – ДНР, Донецк: ГУ «Институт проблем искусственного интеллекта» (ГУ «ИПИИ»). – 25.05.2022. – С. 152-158. – Текст : непосредственный.

85. Ниценко, А. В. О снятии омонимии предикатив-наречие-краткое прилагательное в случае единственного кандидата на предикатив / А. В. Ниценко, В.

Ю. Шелепов, С. А. Большакова // Донецкий международный круглый стол «Искусственный интеллект: теоретические аспекты и практическое применение» (ИИ-2022). – ДНР, Донецк : ГУ «Институт проблем искусственного интеллекта» (ГУ «ИПИИ»). – С. 158-163. – Текст : непосредственный.

86. Ниценко, А. В. Лексико-синтаксический метод снятия омонимии в русскоязычных текстах / А. В. Ниценко, В. Ю. Шелепов, С. А. Большакова // Речевые технологии. – 2023 г. – № 2. – Москва : ИД «Народное образование». – С. 40–48. – Текст : непосредственный.

87. 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. 53–63. – ISSN 2413-7383. – Текст : непосредственный.

88. Ниценко, А. В. К вопросу об автоматическом снятии омонимии предикативов / А. В. Ниценко, В. Ю. Шелепов, С. А. Большакова // Материалы международного научного круглого стола «Искусственный интеллект: теоретические аспекты, практическое применение» (г. Донецк, 27 мая 2021г.). – 2021. – С. 124-126. – Текст : непосредственный.

89. Большакова, С. А. К вопросу о снятии омонимии «предикатив – предложная группа» / С. А. Большакова // Донецкий международный круглый стол «Искусственный интеллект: теоретические аспекты и практическое применение» (ИИ-2023). – ДНР, Донецк : ГУ «Институт проблем искусственного интеллекта» (ГУ «ИПИИ»). – С. 25-28. – Текст : непосредственный.

90. 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 г. – № 2 (29). – С. 49–57. – ISSN 2413-7383. – Текст : непосредственный.

91. 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. – ISSN 2413-7383. – Текст : непосредственный.

92. 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. – ISSN 2413-7383. – Текст : непосредственный.

93. Ниценко, А. В. 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. – ISSN 2413-7383. – Текст : непосредственный.

94. Национальный корпус русского языка URL: http://ruscorpora.ru/new/index.html (дата обращения: 06.03.2025). – Текст : электронны).

95. Александрова, З. Е. Словарь синонимов русского языка: Практический справочник: Ок. 11 000 синоним. рядов. – 11 изд., перераб. и доп. – Москва : Рус. яз., 2001. – 568 с. – Текст : непосредственный.

96. Алиева, Т. С. Словарь синонимов русского языка: с грамматическими приложениями / Т. С. Алиева. – Москва, 2001. – Текст : непосредственный.

97. Большакова, С. А. К вопросу об автоматическом снятии омонимии русских деепричастий / С. А. Большакова // Материалы международного научного круглого стола «Искусственный интеллект: теоретические аспекты, практическое применение» (г. Донецк, 27 мая 2021г.). – 2021. – С. 120-123. – Текст : непосредственный.

98. Морфологический анализатор Mystem 3.0 URL: https://yandex.ru/dev/mystem/?ysclid=mfb7zszbne612289850 (дата обращения 06.03.2025). – Текст : электронный.

99. Морфологический анализатор Pymorphy2. URL: https://pymorphy2.readthedocs.io/en/latest/index.html (дата обращения 06.03.2025). – Текст : электронный.

100. Weischedel Ralph M. Coping with ambiguity and unknown words through probabilistic models // Computational Linguistics. Cambridge, MA, USA: MIT Press,1993. V. 19, Issue 2. P. 361–382.

101. Ratnaparkhi, A. Maximum entropy model for part-of-speech tagging // Proceedings of the Empirical Methods in Natural Language Processing. Philadelphia, PA, USA, 1996. P. 133–142.

102. Гатауллин Р. Р. Методы, модели и программный инструментарий разрешения многозначности в текстах : дисс. ... канд. тех. наук : 05.13.11 / Гатауллин Рамиль Раисович; Место защиты: Казан. (Приволж.) федер. ун-т. – Казань, 2019. – 173 с.

103. Бабенко, Л. Г. Большой толковый словарь русских существительных: Идеографическое описание. Синонимы. Антонимы / Л. Г. Бабенко. – Москва, 2008. – Текст : непосредственный.

104. Бабенко, Л. Г. Словарь-тезаурус синонимов русского языка / Л. Г. Бабенко. – Москва, 2017. – Текст : непосредственный.

105. Бирих, А. К. Словарь фразеологических синонимов русского языка: свыше 8 000 русских фразеологизмов, 950 синонимических рядов / Бирих А. К., Мокиенко В. М., Степанова Л. И. – Москва, 2009. – Текст : непосредственный.

106. Горбачевич, К. С. Русский синонимический словарь / К. С. Горбачевич. – СПб., 1996. – Текст : непосредственный.

107. Горбачевич, К. С. Словарь синонимов русского языка: более 4 000 синонимов / К. С. Горбачевич. – Москва, 2012. – Текст : непосредственный.

108. Жуков, В. П. Словарь фразеологических синонимов русского языка: около 730 синонимических рядов / Жуков В. П., Сидоренко М. И., Шкляров В. Т.; под ред. В. П. Жукова. – Москва, 1987. – Текст : непосредственный.

109. Зимин, В. И. Учебный словарь синонимов русского языка / Зимин В. И., Алекторова Л. П. – Москва, 1994. – Текст : непосредственный.

110. Новый объяснительный словарь синонимов русского языка / Под общ. рук. Ю.Д. Апресяна. Вып. 1. – Москва, 1997; Вып. 2. – Москва, 2000. – Текст : непосредственный.

111. Русский семантический словарь. Толковый словарь, систематизированный по классам слов и значений / Под ред. Н.Ю. Шведовой. Т. 1-4. – Москва, 1998-2007. – Текст : непосредственный.

112. Ахманова, О. С. Словарь омонимов русского языка / Ахманова О. С. – Москва, 1986. – Текст : непосредственный.

113. Ляшевская, О. Н. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка) / Ляшевская О. Н., Шаров С. А. – Москва : Азбуковник, 2009. – Текст : непосредственный.

114. Ниценко, А. В. 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. – ISSN 2413-7383. – Текст : непосредственный.

115. 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. – ISSN 2413-7383. – Текст : непосредственный.

116. Большакова, С. А. Об автоматизированных системах адаптации русскоязычных текстов / С. А. Большакова // Материалы Донецкого международного научного круглого стола «Искусственный интеллект: теоретические аспекты и практическое применение» ИИ-2020. – Донецк: ГУ ИПИИ, 2020. – С. 34–39. – Текст : непосредственный.

117. Ниценко, А. В. О некоторых подходах к автоматическому извлечению информации из текста / А. В. Ниценко, В. Ю. Шелепов, С. А. Большакова, К. С. Ивашко // Материалы Донецкого международного научного круглого стола «Искусственный интеллект: теоретические аспекты и практическое применение» ИИ-2020. – Донецк: ГУ ИПИИ, 2020. – С. 148–152. – Текст : непосредственный.

118. Ниценко, А. В. Исследование омонимии предикативных словосочетаний на основе национального корпуса русского языка / А. В. Ниценко, В. Ю. Шелепов, С. А. Большакова // Материалы VII Международной научно-технической конференции «Современные информационные технологии в

образовании и научных исследованиях (СИТОНИ-2021) Донецк, 23 ноября 2021 г.» / Под общей редакцией В. Н. Павлыша – Донецк: Донецкий национальный технический университет (Донецк), 23.11.2021 г. – С. 510-514. – Текст : непосредственный.

119. 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 : ФГБНУ «ИПИИ», 2024. – 328 с. – С. 11–15. – Текст : непосредственный.

120. 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.

121. Дефо, Д. Жизнь и удивительные приключения морехода Робинзона Крузо : для среднего школьного возраста / Д. Дефо, К. И. Чуковский ; Даниель Дефо ; [пересказ с англ. Корнея Ивановича Чуковского]. – Москва : Самовар, 2009. – (Школьная библиотека). – ISBN 978-5-9781-0162-1. – EDN QUFEFT.

ПРИЛОЖЕНИЕ А

Примеры работы программной реализации метода снятии омонимии

Таблица А.1 – Пример работы программной реализации метода снятии омонимии предикатив-наречие-краткое прилагательное в случае единственного кандидата на предикатив

№	Текст из НКРЯ	Метка о части речи		Точность разметки
		НКРЯ	Авторс-кий метод	
1	Все дальнейшее просто АНТИНАУЧНО.	прил	прил	1
2	Вызывают меня к начальству: «Клевера больше сеять не будешь. АНТИНАУЧНО».	предик	предик	1
3	И вдруг среди галок найдутся отдельные АНТИНАУЧНО настроенные элементы, которые начнут летать с места на место?	нар	нар	1
4	Бесцерковный Голубев не отрицал Бога ? нельзя отрицать то, что тебе недоступно, такое отрицание АНТИНАУЧНО.	прил	прил	1
5	Он НЕВЕРОЯТНО душевный и притягательный.	нар	нар	1
6	Стивен Норингтон слепил НЕВЕРОЯТНО качественную картину и дал жизнь целому франчайзу.	нар	нар	1
7	Точно так же и многое видимое для Димы — НЕВЕРОЯТНО.	предик	предик	1
8	Но было НЕВЕРОЯТНО не видеть, как изменилась Соня!	предик	предик	1
9	И по-прежнему играю в театре, и по-прежнему НЕВЕРОЯТНО хочу сниматься в кино.	нар	нар	1
10	Он спал ГЛУБОКО и спокойно, но ровно через 20 минут он проснется.	нар	нар	1
11	Но различие слишком уж глубоко.	прил	прил	1
12	Как бы ГЛУБОКО ни было падение человека или народа	прил	прил	1
13	Думаю это еще мягко сказано…	нар	нар	1
14	неродную дочь, которую, МЯГКО говоря, недолюбливает	нар	нар	1
15	чтобы депутату было тепло и МЯГКО	предик	нар	0
16	Тепло, МЯГКО, ощущение себя растворено в воздухе,	предик	предик	1
17	Много ли надо человеку? МЯГКО, удобно, тепло.	предик	предик	1
18	тело его МЯГКО и съедобно	прил	прил	1
19	Забравшись с ногами на стул и запахнув ЛЕГКОМЫСЛЕННО розовый ночной халат,	нар	нар	1

Продолжение таблицы А.1

№	Текст из НКРЯ	Метка о части речи		Точность разметки
		НКРЯ	Авторс-кий метод	
20	Не стоит относиться к ним снисходительно или ЛЕГКОМЫСЛЕННО	нар	нар	1
21	Только потом ножницами надрезал конверт и, ЛЕГКОМЫСЛЕННО посвистывая, извлек содержимое.	нар	нар	1
22	И учителя пользуются, и — что ИНТЕРЕСНО — даже многие ученики скачивают презентации.	прил	прил	1
23	В универе очень ИНТЕРЕСНО течет жизнь, постоянно новые знакомства…	нар	нар	1
24	Думаю это еще МЯГКО сказано…	нар	нар	1
25	И ИНТЕРЕСНО куда на самом деле ушел 1 трлн рублей	нар	нар	1
26	Natalie, а не могли вы подробнее рассказать о поездке. Очень ИНТЕРЕСНО. Правда.	предик	предик	1
27	Нравилось. ИНТЕРЕСНО.	предик	предик	1
28	Было очень ИНТЕРЕСНО, но трудно!	предик	предик	1
29	Это не демонстрация, разрешения у властей брать не надо. Очень ИНТЕРЕСНО.	предик	предик	1
30	Неважно, что тут нет никакого экшена, спецэффектов и прочего, что ИНТЕРЕСНО зрителю сейчас.	прил	предик	0
31	не выглядя при этом комично	нар	нар	1
32	Веничка КОМИЧНО стукнул кулаком по пеньку	нар	нар	1
33	Нержин прижмурился и КОМИЧНО потряс головой	нар	нар	1
34	В Москве, в отличие от Тбилиси, зимой ХОЛОДНО	предик	предик	1
35	4 апреля, в день рождения Андрея Тарковского, здесь было еще ХОЛОДНО, накануне ночью выпал снег, отчего и Волга стала совсем белой.	предик	предик	1
36	где была принята критикой довольно ХОЛОДНО.	нар	нар	1
37	Окружные суды обычно располагались в старых, запущенных зданиях, в которых зимой было ХОЛОДНО, чадили печи, а летом — жарко и душно.	предик	предик	1
38	Во тьме белело ее прекрасное лицо, но сейчас оно было ХОЛОДНО.	прил	прил	1
39	его обращение стало ХОЛОДНО	прил	прил	1
40	В лесу было так хорошо, так тихо и СПОКОЙНО, что преступления казались выдумкой досужих сочинителей с воспаленным воображением.	предик	предик	1
Всего				38/40 95%

Таблица А.2 – Пример работы программной реализации метода снятия омонимии предикативных неделимых словосочетаний

№	Текст из НКРЯ	Метка о части речи		Точ-ность разметки
		ЭР	Авторский метод	
1	Премьер купается в Байкале, а потом — в Черном море, у него все В АЖУРЕ.	предик	нар	0
2	все у него В АЖУРЕ	предик	предик	1
3	Эти люди активно защищают своего руководителя на любых партийных собраниях, чувствуя себя В ДОЛГУ.	нар	нар	1
4	Мне не хотелось чувствовать себя у нее В ДОЛГУ, но надо было быть благодарным	нар	нар	1
5	И наконец, красивее быть В ДОЛГУ перед государством, чем одалживать у знакомых.	предик	предик	1
6	Давай зарабатывай, мы с Коляшей это место В ОХОТКУ уступим и свою копейку за труды сдадим, свети и охраняй.	нар	нар	1
7	проведав детей и привезя продуктов, поделав В ОХОТКУ или для видимости что-нибудь в огороде и поев ягод с кустов, они уезжали до следующих выходных,	нар	нар	1
8	Корытин В ОХОТКУ похрумкивал жареными карасиками, девок хвалил:	предик	предик	1
9	Скажешь! ? сказал Ринат, ударяя ее совсем уже В ОХОТКУ, с увлечением.	нар	нар	1
10	Обычно НА СЛУХУ другие органы ООН	предик	предик	1
11	Это те имена, которые НА СЛУХУ.	предик	предик	1
12	Проблема только в том, что на шпану похожи меньше всего ? женихи НА ВЫДАНЬЕ, да и только.	предик	предик	1
13	сыновей Господь не дал, однако дочери НА ВЫДАНЬЕ.	предик	предик	1
14	Наоборот, начинаешь понимать, что с Богом сражаться НЕ ПОД СИЛУ никому, ни хладнокровным убийцам, ни бизнесменам, ни даже всему государству Швейцария.	предик	предик	1
15	Создание сайтов из подобных модулей ПОД СИЛУ даже не очень опытным преподавателям и иногда используется в качестве практических работ для подготовленных студентов.	предик	предик	1
16	сохраняя при этом облик солидных и серьезных вещей, которым НЕ ДО ШУТОК.	предик	предик	1
17	Но немец пер, было не до шуток.	предик	предик	1
18	А если расклад таков, что плохо не ребенку, а наоборот родителю, а ребенку В САМЫЙ РАЗ?	предик	предик	1
19	У него об этом счастье сказано В САМЫЙ РАЗ.	нар	нар	1
20	Вот это будет хорошо, это мне В САМЫЙ РАЗ.	предик	предик	1
Всего				19/20 95%

Таблица А.3 – Пример работы программной реализации метода снятия омонимии предикативных словосочетаний с отрицанием

№	Текст из НКРЯ	Метка о части речи		Точ-ность разметки
		ЭР	Авторский метод	
1	Но, во-первых, НЕ ГРЕХ накануне профессионального праздника порадоваться тому, что газета по-прежнему востребована большой частью жителей республики,	предик	предик	1
2	Все в трактире заговорили громче, задвигали над столом кружками ? за такое важное дело НЕ ГРЕХ было хорошо выпить.	предик	предик	1
3	Это не порок наш и НЕ ГРЕХ - это замысел природы.	отриц	предик	0
4	А в нашем деле… НЕ ГРЕХ ошибиться великому политику.	предик	предик	1
5	вообще это НЕ ДЕЛО, гораздо лучше по-другому	предик	предик	1
6	Стремится к жизни, чтобы, вероятная логика, сделаться жизнью, как бы ни говорили (и справедливо), что, мол, это НЕ ДЕЛО поэзии.	отриц	предик	0
7	Но вообще это НЕ ДЕЛО, когда суды выступают с законодательной инициативой.	предик	предик	1
8	Да, НЕ ДЕЛО, но стремление стремлению рознь.	предик	предик	1
9	И НЕ БЕДА, что нет шедевров ? они опять же в избытке в постоянной экспозиции новой французской живописи.	предик	предик	1
10	И НЕ БЕДА, что на улице уже далеко за полночь.	предик	предик	1
11	видно, пошел НЕ ВПРОК,	нар	нар	1
12	Кашу ест ртом бесчувственным, она ему НЕ ВПРОК.	нар	нар	1
13	Затраты и труды надо почитать напрасными и никому НЕ ВПРОК.	нар	нар	1
14	карьерист, НЕ ПРОЧЬ загрести жар чужими руками.	предик	предик	1
15	Перед выборами политики НЕ ПРОЧЬ приникнуть к "живительным " капиталам пивоваров	предик	предик	1
16	Должники России, судя по всему, НЕ ЧЕТА ее кредиторам?	предик	предик	1
17	Все эти басни я рассказываю к тому, что серебряные, золотые и прочие свадьбы у нас были НЕ В ХОДУ.	предик	предик	1
18	Да-да, НЕ К МАСТИ козырь, как говорит обо мне твоя мама.	предик	предик	1
19	Поэтому в принципе им НЕ РЕЗОН заваливать меня.	предик	предик	1
20	Но мне НЕ СМЕШНО ни капельки.	предик	предик	1
Всего				18/20 90%

Таблица А.4 – Пример работы программной реализации метода снятия омонимии предикативных словосочетаний и разделимых словосочетаний, которые не являются предложными группами

№	Текст из НКРЯ	Метка о части речи		Точ-ность разметки
		ЭР	Авторс-кий метод	
1	в прокуратуру или КУДА ТАМ еще бежать	част	част	1
2	Балансировать экономику мыльного пузыря ? значит делать согражданам очень больно, КУДА ТАМ твои антинародные реформы Гайдара.	част	част	1
3	ВЕРОЯТНЕЕ ВСЕГО, антитела к одной подгруппе будут достаточно хорошо взаимодействовать со штаммами другой подгруппы.	ввод	ввод	1
4	Эта прошумевшая ветка, полная цветов и листьев, ВЕРОЯТНЕЕ ВСЕГО ветка белой акации	предик	предик	1
5	Но ВЕРОЯТНЕЕ ВСЕГО - город был именно там.	предик	предик	1
6	ВЕРОЯТНЕЕ ВСЕГО то, что красавица эта не была списана с натуры	предик	предик	1
7	Наверное, где-то на форуме есть уже такой рецепт, но ВСЕ РАВНО напишу, нравится мне этот салат, потому что быстро и вкусно.	част	част	1
8	– Не бойся, ВСЕ РАВНО не понадобится.	част	част	1
9	Фильм о человеческих мечтах и о том, что даже в такой дыре они ВСЕ РАВНО сбываются.	част	част	1
10	Понял напрасно, потому что это ничего не меняло, и я ВСЕ РАВНО был вынужден брать на время велик у своего одноклассника.	предик	предик	1
11	И, наконец, главное – одному ездить ВСЕ РАВНО интереснее.	предик	предик	1
12	У партизан перловки найдется. ? Это еще КАК СКАЗАТЬ!	предик	предик	1
13	А Марку ничего не оставалось, КАК СКАЗАТЬ в ответ спасибо.	соч: част+инф	соч: част+инф	1
14	отвергая саму мысль, что семья МОЖЕТ БЫТЬ такой, какая она описана в книге.	соч: глаг+глаг	соч: глаг+глаг	1
15	По ходу фильма думаешь, а почему эта история не МОЖЕТ БЫТЬ правдой?	соч: глаг+глаг	соч: глаг+глаг	1
16	то МОЖЕТ БЫТЬ даже еще не поздно.	предик	предик	1
17	Объем пробки ПРОЩЕ ВСЕГО найти с помощью мензурки.	нар	нар	1
18	Из упомянутых г-ном Бершадским сырьевых ресурсов ПРОЩЕ ВСЕГО решается проблема с бумагой	нар	нар	1
19	Казалось бы, проще всего было зайти к ней и спросить, где Сонька…	предик	предик	1
20	Для животных, не умеющих поддерживать постоянную температуру тела, ПРОЩЕ ВСЕГО подыскать для себя теплое убежище.	предик	предик	1
Всего				20/20 100%

Таблица А.5 – Пример работы программной реализации метода снятия омонимии предикатив–существительное

№	Текст из НКРЯ	Метка о части речи		Точ-ность разметки
		ЭР	Авторс-кий метод	
1	Опять бардак в доме ? КАРАУЛ!	предик	предик	1
2	В тех редких случаях, когда замученному первогодку, заступившему в КАРАУЛ, выдают оружие.	сущ	сущ	1
3	Гости пришли, а он прыг на потолок и МОЛЧОК!	сущ	сущ	1
4	Но мать об уходе МОЛЧОК - как не было ничего.	предик	предик	1
5	Когда хрупкое, когда ранимое, когда унылое. Просто БЕДА с ним.	предик	сущ	0
6	С ценами совсем БЕДА: свечи - 2 руб, матрацы - 170 руб, кружки - 48 коп…	предик	сущ	0
7	На обложке изображены манхэттеновские высотки… в момент взрыва. ЖУТЬ.	предик	предик	1
8	Замешалась звенящая ЖУТЬ.	сущ	сущ	1
9	ее охватывал жгучий СТЫД	сущ	сущ	1
10	Доложите мне лично! СТЫД! Позор!	предик	предик	1
11	нас принят мораторий на смертную казнь и возиться с этим сложным делом командованию НЕДОСУГ.	предик	предик	1
12	Мне решительно НЕДОСУГ подумать, кто и зачем прибрал к рукам мою!	предик	предик	1
13	У одного из них КРЫШКА часто подпрыгивает	сущ	сущ	1
14	Если новым президентом станет мент, нам КРЫШКА.	предик	предик	1
15	Каждая погода – БЛАГОДАТЬ.	предик	предик	1
16	Есть еще в России удивительные уголки, где на человека может и впрямь снизойти БЛАГОДАТЬ Божья	сущ	сущ	1
17	ли он и замечал петербургское небо, то рассеянно, никогда не ощущая вот этого "московского " движения души: какая БЛАГОДАТЬ…	предик	предик	1
18	Если вам летать ОХОТА…	предик	предик	1
19	ОХОТА язык ломать?	предик	предик	1
20	На въезде - щит с надписью "ОХОТА запрещена".	сущ	сущ	1
Всего				18/20 90%

Таблица А.6 – Пример работы программной реализации метода снятия омонимии наречия и существительного

№	Текст из НКРЯ	Метка о части речи		Точ-ность разметки
		НКРЯ	Авторский метод	
1	Ну я прямо и говорю: поставьте мне авансом	нар	нар	1
2	А теперь идите за авансом, Верочка.	сущ	сущ	1
3	нормы обязывают налогоплательщиков увеличивать налоговую базу налога на добавленную стоимость только на те денежные суммы, которые ими получены АВАНСОМ	нар	нар	1
4	деньги, которые будут даны мне, являются АВАНСОМ	сущ	сущ	1
5	уроки сделали и вечером БЕГОМ в клуб к Ботвиннику	нар	нар	1
6	совершала кросс в парк и там делала всевозможные спортивные упражнения, а потом БЕГОМ мы возвращались назад.	нар	нар	1
7	Я подумал, что, может быть, он занимается БЕГОМ, и только что прибежал со стадиона.	сущ	сущ	1
8	И с ужасом понимаешь, что и он тоже заражен временем, а главное - еще в утробе отравлен паническим страхом перед его БЕГОМ	сущ	сущ	1
9	разом закричали они, пытаясь унять прерывистое, сбитое шибким БЕГОМ дыхание.	сущ	сущ	1
10	ВРЕМЕНАМИ я его смотрю, потом смотрю опять и снова.	нар	нар	1
11	На нем ВРЕМЕНАМИ бушуют опасные волны, которые называют «плесками», от них, вероятно, и появилось название этого озера.	нар	нар	1
12	зрителя ВРЕМЕНАМИ настигает ощущение, будто он смотрит документальный фильм.	нар	нар	1
13	Учитель русской словесности был еще хуже: среднего роста, довольно худощавый брюнет, смотревший на всех ВОЛКОМ	нар	нар	1
14	Так оно вроде, знаешь, все одно и то же, хоть ВОЛКОМ вой.	нар	нар	1
15	Автору этот дядя кажется похожим на героя романа Саши Соколова "Между собакой и ВОЛКОМ "	сущ	сущ	1
16	Я был ВОЛКОМ, а ты меня лупил кочергой…	сущ	сущ	1
17	что заставляло Крылова ДОБРОМ вспоминать одинаковых Ритку и Светку	нар	нар	1
18	тяжелый там народец, не сразу удалось с ними ДОБРОМ	нар	нар	1
19	Как же, знаете, приятно писать про любимые фильмы, фильмы, которые насквозь пропитаны светом и ДОБРОМ.	сущ	сущ	1
20	Казаков моих ДОБРОМ наделила	сущ	сущ	1
Всего				20/20 100%

Таблица А.7 – Пример работы программной реализации метода снятия омонимии деепричастий

№	Текст из НКРЯ	Метка о части речи		Точ-ность разметки
		НКРЯ	Авторс-кий метод	
Существительное-деепричастие				
1	Пиранья более опасна. Она без рыка и без ВОЯ съедает в речке все живое.	сущ	сущ	1
2	лишь бы не слышать страшного ВОЯ немецких снарядов	сущ	сущ	1
3	Вылетал из другого, свистя и ВОЯ по дороге.	деепр	деепр	1
4	что какая-то несчастная вдова, ВОЯ, стояла перед ним на коленях под дождем	деепр	деепр	1
5	прохрипел он в ухо, раскачивая его и почти ДУША.	деепр	деепр	1
6	Все режем как ДУША подскажет, кто мелко, кто крупно.	сущ	сущ	1
7	но и совершить настоящий ПРОРЫВ в области технологий банковского обслуживания.	сущ	сущ	1
8	ПРОРЫВ 34-й армии наткнулся на встречный удар дивизии	сущ	сущ	1
9	ПРОРЫВ канал между Европой и Африкой	деепр	деепр	1
10	ПЛАЧА от боли, горя, обид и смеясь от счастья	деепр	деепр	1
11	Та, ПЛАЧА и смеясь, заговорила о том, как бесчестно поступил он, когда сталкивал ее в овраг.	деепр	деепр	1
12	Но дед не слышал их тихого ПЛАЧА, он как бы оглох.	сущ	сущ	1
13	Он не хотел ехать среди пенья и ПЛАЧА.	сущ	сущ	1
14	Проверить можно, ПРОКОЛОВ вилкой через фольгу.	деепр	деепр	1
15	Первый опыт проведения такого крупного турнира оказался вполне удачным - организационных ПРОКОЛОВ не случилось	сущ	сущ	1
Предлог-деепричастие				
16	БЛАГОДАРЯ таким стоматологам, я остался без единого зуба и шансов поставить протезы.	предл	предл	1
17	Однако за два года, БЛАГОДАРЯ своей тяге к знаниям, выучила неплохо английский, да и другое.	предл	предл	1
18	Будем молиться, БЛАГОДАРЯ Бога за то, что эта буря пронеслась	деепр	деепр	1
19	Старая женщина, обрядившись в теплый платок да ватник, вышла провожать их, прощаясь и БЛАГОДАРЯ	деепр	деепр	1
20	каждый день прилетал ворон, ДЛЯ его агонию вечно	деепр	деепр	1
21	отвели ДЛЯ этого специальные места	предл	предл	1
Прилагательное-деепричастие				
22	Как вы представляете себе науку: Сивка-Бурка, ВЕЩАЯ каурка?	прил	прил	1
23	Ах, ворон, ворон, ВЕЩАЯ птица!	прил	прил	1
24	Запугивал зрителей, ВЕЩАЯ с экранов телевизора о новом всемирном потопе	деепр	деепр	1
25	Страх одолел его, ГОРЯЧА кровь	деепр	деепр	1
26	Уж очень она ГОРЯЧА, товарищ подполковник	прил	прил	1
27	я это видел, ЗАЕЗЖАЯ иногда к ней и днем	деепр	деепр	1

№	Текст из НКРЯ	Метка о части речи		Точ-
28	И сразу же, не ЗАЕЗЖАЯ домой, подъехал в Большой дом.	деепр	деепр	1
29	Заезжая аристократия в джинсах и местная публика куда более серого вида.	прил	прил	1
30	И за стеклами очков в золотой оправе СКУПАЯ мужская слеза.	прил	прил	1
31	Дед очень симпатичный, он тебя пальцем не тронет, хотя ты злая и СКУПАЯ.	прил	прил	1
32	А в России, если предприятие встает на ноги (как наше, например), его тут же пытаются обанкротить, применяя всевозможные финансовые махинации и СКУПАЯ его потом за бесценок.	деепр	деепр	1
33	в порыве ревности и обиды поранил ее лицо ножом, который всегда имел при себе, постоянно что-то СТРОГАЯ.	деепр	деепр	1
34	Приходилось прогуливать все уроки и сильно рисковать, учеба давалась мне с таким трудом, что СТРОГАЯ дама завуч неоднократно предлагала маме сводить меня к детскому психиатру.	прил	прил	1
35	И он утверждает, что это ? "СТРОГАЯ математика "?!	прил	прил	1
36	Еще есть ХРОМАЯ кассирша, перемещающаяся по кинотеатру с непонятной целью, и киномеханик, которого мы почти не видим	прил	прил	1
37	ХРОМАЯ девушка, заметная и на киносеансах, попала в больницу, в женское отделение..	прил	прил	1
38	Еще не совсем оправившись от болезни, ХРОМАЯ, он начал свою карьеру, по сути дела, заново.	деепр	деепр	1
39	Я бежала - буквально бежала! - в кухню, ХРОМАЯ при каждом шаге.	деепр	деепр	1
40	Эта хрупкая и на вид чрезвычайно слабая, ХВОРАЯ Юлия Михайловна с тонкими ручками, пергаментно-белым лицом была, надо сказать, необыкновенно упряма.	прил	деепр	0
41	Откуда-то с холщовых вершин спускалась тонкая витая веревка	прил	прил	1
42	Витая пара: категории, обжим, советы по работе	прил	прил	1
	Причастия- деепричастия			
43	Каких постановлений тыщи, в ветвях ВИТАЯ, стучит твой пальчик, неостывший после свиданья?	деепр	деепр	1
44	успокаивала она Ягуна, ВИТАЯ рядом с ним прозрачной тенью	деепр	деепр	1
45	При всей изощренности моей фантазии, даже ВИТАЯ в облаках, я не нахожу ответа ни на один вопрос.	деепр	деепр	1
46	Всю жизнь он проохотился, ОБИТАЯ зимой в рваной палатке, и хорошего дома не имел.	деепр	прич	0
47	Тугая, ОБИТАЯ на зиму войлоком, она тяжело подается, сквозь щель влетает колкое морозное облако.	прич	прич	1
48	ОБИТАЯ в высоких широтах Арктики, гренландский кит способен пробивать лед толщиной 20 - 30 сантиметров.	деепр	деепр	1

№	Текст из НКРЯ	Метка о части речи		Точ-
Деепричастия переходного и непереходного глагола				
49	Она усадила Лену, а сама стояла против нее, то ВКАЛЫВАЯ швейную иглу в отворот блузки, то вынимая опять.	перех	перех	1
50	Бритый врач не совсем верной рукой сдавил в щипок остатки мяса, ВКАЛЫВАЯ в руку Турбину иглу маленького шприца.	перех	перех	1
51	Имею я, ВКАЛЫВАЯ на четырех работах, право на отдых?	неперх	неперх	1
52	остыдил себя в глазах двора явлением непристойным, ДОСАДИВ вместе и ляхам и россиянам.	неперх	неперх	1
53	Различие темпераментов в работе, и вот: я нарычал на Ольгу, она огрызнулась, швырнула резиновые перчатки и уехала, не ДОСАДИВ грядку с капустой.	перех	перех	1
54	Работая как проклятый, НЕДОСЫПАЯ, он сумел обратить на себя внимание престижных изданий	неперх	неперх	1
55	И тут замахнулся паж разведенными ножницами, ЦЕЛЯ острый конец себе в горло.	перех	неперх	0
Всего				52/55 94,6%

ПРИЛОЖЕНИЕ Б

Содержание файла «Список дициплин»

агробиология, агрономия, агрофизика, агрохимия, агроэкология, аквакультура, акупунктура, акустика, акушерство, алгебра, альгология, анатомия, андрология, анестизиология, антропогенез, антропология, архитектура, астробиология, астрономия, астрофизика, аэробиология, аэрофотосъемка, библиология, биогеография, биоинформатика, биология, биотехнология, биофизика, биохимия, ботаника, ветеринария, виноградарство, вирусология, гематология, геммология, генетика, география, геодезия, геология, геометрия, геоморфология, геофизика, геохимия, гериатрия, герменевтика, герпетология, гидрогеология, гидрография, гидрология, гинекология, гистология, гляциология, граждановедение, диабетология, диетология, дизайн, драматургия, египтология, живопись, журналистика, зоология, зубопротезирование, иммунология, интерлингвистика, информатика, искусствоведение, история, ихтиология, каллиграфия, кардиология, кардиохирургия, картография, кибернетика, киноведение, климатология, кораблестроение, космонавтика, криобиология, криогеника, криптозоология, ксенобиология, культурология, лесоводство, лимнология, лингвистика, литература, литературоведение, литология, логика, логистика, маркетинг, математика, материаловедение, медицина, менеджмент, метеорология, механика, микология, микробиология, минералогия, мифология, морфология, музыка, музыковедение, неврология, нейробиология, нейропсихология, нейрохиругия, нейрохирургия, нефрология, нутрициология, океанография, океанология, онкология, онтология, оология, оптика, орнитология, ортодонтия, ортопедия, отоларингология, офтальмология, палеоантропология, палеобиология, палеогеография, палеоклиматология, палеонтология, паразитология, парапсихология, патология, педагогика, педиатрия, периодонтология, планетология, политология, помология, поциолингвистика, почвоведение, поэтика, право, прагматика, приматология, психиатрия, психоанализ, психология, психометрия, психопатология, психотерапия, психофизика, пчеловодство, радиология, ревматология, религиеведение, рисование, риторика, садоводство, семантика, семиотика, серология, синтаксис, систематика, системотехника, скульптура, социолингвистика, социология, спелеология, статистика, стоматология, строительство, сценография, таксономия, теория, терапия, тератология, термодинамика, технология, травматология, трансфузиология, урология, фармакология, фармация, физика, физиология, физиотерапия, филология, философия, фольклор, фонетика, фонология, фотография, хемоинформатика, химия, хирургия, хореография, хронобиология, цитология, экология, экономика, электромагнетизм, электротехника, эмбриология, эндокринология, энология, энтомология, эпидемиология, эпистемология, эстетика, этика, этноботаника, этнография, этноистория, этнолингвистика, этнология, этология.

ПРИЛОЖЕНИЕ В

Словарь для снятия омонимии наречия и существительного

авансом # (нар) брать, взыскать, взыскивать, взять, внести, вносить, вознаградить, вознаграждать, выдавать, выдать, выплатить, выполнить, делать, заплатить, засчитывать, начислять, оплатить, платить, получать, получить, поощрить, поощрять, премировать, сделать, уплатить

авансом # (сущ) воспользоваться, довольствоваться, заручиться, интересоваться, обеспечить, ограничиваться, ограничиться, пользоваться, пренебрегать, пренебречь, распорядиться, распоряжаться

артелью # (сущ) воспользоваться, восхититься, восхищаться, заинтересоваться, залюбоваться, заниматься, заняться, интересоваться, любоваться, пользоваться, полюбоваться, править, руководить, управлять

артелью # (нар) выскакивать, выскочить, действовать, держаться, набрасываться, навалиться, прятаться, работать, сбегать, сбежать, скрываться, скрыться, собираться, спрятаться, сражаться, убегать, убежать, удирать, удрать, юркнуть

бегом # (сущ) воодушевиться, воодушевляться, заниматься, заняться, увлекаться, увлечься

бегом # (нар) бежать, взбежать, возвращаться, доставить, записывать, писать, мчаться, направиться, нести, принести, донести, занести, отнести, нестись, отправлять, покидать, пуститься, рвануть, ринуться, устремиться

визави # (нар) находится, сидеть, смотреть, рассматривать, разглядывать, танцевать, уставиться

вихрем # (сущ) быть, стать

вихрем # (нар) вертеться, завертеться, взвиться, взвиваться, врываться, ворваться, кружиться, закружиться, крутиться, закрутиться, лететь, налететь, залететь, долететь, метаться, взметнуться, мчаться, помчаться, промчаться, примчаться, носиться, нестись, пронестись, уносить, поднимать, поднять, разбросать, распространять, ринуться, сметать, становиться

волей # (сущ) вершить, владеть, завладеть, наделить, обладать, овладеть, подавлять, помыкать, пренебрегать, распоряжаться, руководить, руководствоваться, склониться становиться, сломиться, совладать, управлять

волей # (нар) идти, пойти, ехать, поехать, отправится, бросить, развязаться

временами # (нар) вспоминать, забывать, забываться, запаздывать, захаживать, напоминать, опаздывать, помнить

временами # (сущ) восхищаться, заниматься, интересоваться

волком # (нар) броситься, взвыть, выть, глядеть, жить, завыть, зыркать, наброситься, озираться, смотреть

волком # (сущ) быть, обернуться, оказаться, остаться, стать

гольём # (нар) взять, гнать, прогонять, пить

голяком # (нар) выпускать, выскочить, выходить, гнать, идти, лежать, носится, оставаться, остаться, погнать, погнаться, посидеть, поспать, пройтись, пускать, сидеть, скакать, спать, станцевать, стоять, танцевать, топтаться, ходить,

гуськом # (нар) бежать, брести, вести, войти, входить, выбегать, выбираться, выбраться, вывести, выдвигаться, выйти, выползать, выползти, выстраиваться, выстроить, выстроиться, вытянуться, выходить, двигаться, заходить, ездить, ехать, запрягать, запрячь, идти, передвигаться, плестись, плыть, подниматься, подходить, ползти, поплестись, пробираться, проезжать, пройти, проплывать, проплыть, проходить, растягиваться, растянуться, следовать, спускать, спускаться, танцевать, тащиться, шагать

добром # (сущ) владеть, воспользоваться, делиться, дорожить, завладеть, заинтересоваться, интересоваться, набивать, набить, наделить, наполнить, обернуться, отвечать, поделиться, поживиться, пользоваться, разбрасываться, распорядиться, распоряжаться, светиться, являться

добром # (нар) благодарить, воздать, закончиться, искупить, кончиться (не) назвать, называть, отблагодарить, откликаться, отвечать, ответить, отплатить, отозваться, платить, подкупать, подкупить

дугой # (нар) висеть, выгнуть, выгнуться, вытягиваться, вытянуться, загнуть, загнуться, изгибаться, изогнуться, согнуть, согнуться

дугой # (сущ) заниматься, заняться, любоваться, залюбоваться, полюбоваться, восхищаться, восхититься, пользоваться, воспользоваться

залпом # (нар) выпалить, выпивать, выпить, выстрелить, допить, осушить, пить, прочитать, читать

залпом # (сущ) накрыть, накрывать, ответить, отвечать

зараз # (нар) разг., брать, взять, вывозить, выдать, выкладывать, выпить, выпечь, выплатить, вытащить, дать, давать, делать, забирать, забрать, заряжать, набрать, осушить, охватить, передать, подавать, получать, привезти, съесть, укрепить

капельку # (нар) взять, выдать, говорить, давать, делать, думать, ждать, забыть, заниматься, заказать, запомнить, изменить, исправить, отдохнуть, показать, помнить, понимать, поправить, разбираться, сидеть, скучать, спать, терпеть, успокоиться, чувствовать

крестом # (нар) завязать, завязывать, зачеркнуть, зачеркивать, перечеркнуть, разрезать, раскинуть, складывать

крестом # (сущ) благословить, испугать, махать, махнуть, наградить, отметить, поманить, проткнуть, ударить, украшать

крохотку # (нар) отдохнуть

крошечку # (нар) думать, завидовать, играть, остепениться, отдохнуть, повременить, потесниться, приютиться, робеть

крошку # (нар) думать

мигом # (нар) вернуться, вспомнить, добывать, догадаться, зарядить, избавиться, испортить, опомниться, откликаться, отправить, оформить, очнуться, очутиться, понять, попадать, появиться, приструнить, проснуться, разобраться, ремонтировать, скрыться, успокоиться, хватать

мигом # (сущ) дорожить, казаться, насладиться, наслаждаться

миром # (нар) завершить, решить, уладить

миром # (сущ) владеть, восхищаться, интересоваться, любоваться, наслаждаться, овладеть, править

молчком # (нар) бежать, волочиться, давать, двигаться, делать, ехать, идти, отдавать, отдать, ползти, работать, садиться, сидеть, смотреть, собираться, стоять, тащить, терпеть, уходить, ходить, шагать

моментом # (сущ) воспользоваться, жить, казаться, насладиться, обладать, пользоваться, распоряжаться, считать

моментом # (нар) забывать, запомнить, зашить, определиться, освоить, снимать

навалом #, грузить, сваливать, складывать, скидывать, ссыпать

наскоком # (нар) делать, действовать, думать, работать

нахрапом # (нар) брать, взять, действовать, лезть

невидимкой # (сущ) быть, казаться, оставаться, притвориться, стать, становиться, украсить

невидимкой # (нар) взбежать, возвращаться, подходить, выходить, приостановиться, скрываться

неволей # (сущ) довольствоваться, грозиться, угрожать

неволей # (нар) женить, мучить, наказывать, отправить, содействовать

неглиже # (нар) гулять, ходить, бежать, быть

особняком # (сущ) владеть, восхищаться, заведовать, интересоваться, обзаводиться, пользоваться, управлять, хвалиться, хвастаться

особняком # (нар) гулять, действовать, держаться, ехать, жить, поселиться, путешествовать, сидеть, стоять, шагать, идти

охотой # (сущ) забавляться, заведовать, заниматься, заняться, кормиться, промышлять, развлекаться, развлечься, отвлекаться, отвлечься, увлекаться, увлечься

охотой # (нар) предлог с, говорить, рассказывать, сообщать

ошибкой # (сущ) быть, воспользоваться, заинтересоваться, казаться, обернуться, оказаться, признавать, считать, упрекать, являться

порой # (сущ) назвать, называть

порой # (нар) бывать, видеть, видеться, встречаться, выручать, думать, забывать, замечать, знать, казаться, помогать, проявлять, случаться, слышать, чудится, чувствовать

пулей # (сущ) зарядить, пробить, пронзить, прострелить, раздробить, ранить, сбить, сразить, стрелять, убить

пулей # (нар) вскочить, выскочить, лететь, мчаться, нестись, отправить, прибежать, сбегать, слетать

самотеком # (нар) идти, действовать, пускать

силком # (сущ) ловить

силком # (нар) брать, взять, выводить, заставить, отправить, тащить, уводить

силом # (нар) забирать, тащить, удерживать

скопом # (нар) выскакивать, выскочить, действовать, держаться, набрасываться, наброситься, навалиться, нападать, прятаться, ругаться, сбегать, сбежать, скрываться, скрыться, собираться, сопротивляться, спрятаться, сражаться, убегать, убежать, удирать, удрать, юркнуть

скопом # (сущ) управлять

скороговоркой # (сущ) вдохновиться, воодушевиться, воспользоваться, заинтересоваться, интересоваться, озадачить, озадачиться, увлекаться, увлечься

скороговоркой # (нар) бормотать, бубнить, вещать, говорить, голосить, затараторить, зачитывать, изложить, лепетать, объяснить, объяснять, отвечать, пробубнить, произнести, произносить, протараторить, сообщать, сообщить, тараторить, твердить, шептать

стрелой # (сущ) вертеть, выбить, воспользоваться, дорожить, жертвовать, зарядить, покарать, пользоваться, поразить, пробить, пронзать, пронзить, ранить, стрелять, убивать, убить, ударить, ударять, управлять

стрелой # (нар) бежать, броситься, влетать, вывернуться, вылетать, вылететь, вытянуться, лететь, метнуться, мчаться, нестись, пронестись, проноситься, рвануть, просвистеть, ринуться, унестись, уноситься, уходить

украдкой # (нар) взглядывать, вздрагивать, вздыхать, всматриваться, выглядывать, высматривать, глядеть, действовать, наблюдать, оглядываться, оглянуться, рассматривать, следить, смотреть

утицей # (сущ) быть, вертеть, воспользоваться, кормить, пользоваться, покормить, прокормиться, хвастаться

утицей # (нар) барахтаться, величать, выкатиться, выкатываться, выходить, выплывать, закачаться, качаться, ковылять, крикнуть, назвать, называть, плыть, проковылять, проплывать, ходить

уточкой # (сущ) баловать, восхититься, впечатлять, гордиться, жертвовать, угощать

уточкой # (нар) ковылять, переваливаться, плыть, покачиваться, поплыть, проплывать, ходить

утречком # (нар) вставать, встать, выбегать, выйти, выходить, готовить, действовать, делать, доносить, допрашивать, думать, заходить, ехать, жалобить, жаловаться, проснуться, узнавать, узнать

утречком # (сущ) восхищаться, дорожить, пренебрегать, пренебречь

целиком # (нар) брать, взять, вмещаться, вытеснить, грузить, зависеть, заполнить, захватить, захлестнуть, нагружать, освоить, охватить, повториться, поглотить, погрузить, погрузиться, поддержать, поддерживать, предстать, принадлежать, проглотить, рассмотреть, раствориться, смотреть, согласится, тратить, уничтожить, читать

цепочкой # (сущ) восхититься, восхищаться, зазвенеть, звякнуть, любоваться, лязгнуть, обладать, украсить

цепочкой # (нар) бежать, выстраиваться, выстроиться, двигаться, ехать, идти, построиться, растянуться, спускаться, стать, тянуться, уводить, уходить

цепью # (сущ) бряцать, взмахнуть, греметь, громыхать, замахнуться, звенеть, крепить, приковать, размахивать, сковать

цепью # (нар) выбираться, выстроиться, вытянуться, двигаться, замыкаться, идти, наступать

цугом # (нар) выезжать, вытянуться, двигаться, ездить, ехать, заезжать, запрягать, отправиться, подъезжать, скакать, ставить, стоять

честью # (нар) искупить, назвать, называть

честью # (сущ) гордиться, дорожить, жертвовать, клясться, обладать, пожертвовать, пользоваться, поступиться, признавать, рисковать, руководствоваться, ручаться, считать

чередой # (нар) выходить, нисходить, проходить

чредой # (нар) выходить, нисходить, проходить

чуточку # (нар) жалеть, испугать, испугаться, обгореть, огорчить, повзрослеть, подождать, подумать, пожаловаться, поплакать, почитать, приврать, пристыдить, спешить, списывать, убедить, увидеть, фантазировать, халтурить, чувствовать, щадить,

ПРИЛОЖЕНИЕ Г

База для замены неизменяемых словосочетаний (фрагмент)

аппетитно !
со смаком

бдительно !
во все глаза
в оба глаза

бегло !
не затрудняясь
без запинки

безвинно !
без вины
ни за что ни про что
ни за хрен ни про хрен

беззаботно !
не зная забот
в свое удовольствие
как птица небесная
как птичка божья

вдоволь !
сколько влезет
в свое удовольствие
сколько душе угодно
до отвала
под завязку
от пуза
в обжор

безнадежно !
гиблое дело
дохлое дело

безнаказанно !
без последствий
без всяких последствий

безопасно !
не подвергаясь опасности

безответно !
без взаимности

безразлично !
все равно
не все ли равно
какая разница
нет никакой разницы
без разницы
все едино
что в лоб что по лбу
один черт
хрен редьки не слаще

бескорыстно !
без корысти
для души
из любви к искусству
за прекрасные глаза
ради прекрасных глаз
из чести

бесплатно !
на даровщину
на даровщинку
на дармовщину
на дармовщинку
на халяву
за так
за спасибо
за одно спасибо
за прекрасные глаза
за красивые глаза
ради прекрасных глаз

беспокойно !
как на вулкане

беспорядочно !
в беспорядке

беспрекословно !
без отговорок
без всяких отговорок
без возражений
без всяких возражений
без никаких
как миленький

беспрепятственно !
без помех
без задержки

беспричинно !
без причины
без всякой причины
без повода
без всякого повода
сам не зная почему
ни с того, ни с сего
с бухты-барахты
с кондачка
за здорово живешь
не говоря худого слова

бессвязно !
ни в склад, ни в лад

бесследно !
без следа

бессознательно !
сам того не сознавая
не отдавая себе отчета

бесцельно !
неизвестно зачем
не зная зачем
неведомо зачем
неизвестно зачем

близко !
по соседству
в нескольких шагах
в двух в трех шагах
перед глазами
под боком
под носом
под самым носом
двор об двор
рукой подать

богато !
не считаясь с затратами
не стесняясь расходами
с размахом
на широкую ногу
на большую ногу
на барскую ногу

больше !
в большей степени
в большей мере

боязливо !
со страхом
с опаской

брезгливо !
с омерзением
с отвращением

буквально !
в буквальном смысле слова
в прямом смысле слова
слово в слово
буква в букву

быстро !
полным ходом
как на нарах
стремительными темпами
гигантскими шагами
стремительными шагами
не по дням, а по часам
как на дрожжах
как грибы
как пулемет
со скоростью пулемета
быстрым шагом
быстрыми шагами
с ветерком
во всю прыть
сломя голову
что есть духу
во весь дух
во весь мах
во весь опор
с быстротой молнии
как на крыльях
как метеор
со всех ног
не чуя под собой ног
не слыша под собой ног
не чувствуя под собой ног
без оглядки
откуда прыть взялась
только пятки засверкали
одна нога здесь, а другая там
не теряя времени
не теряя времени даром
не тратя времени
не тратя времени даром
за короткое время
за короткий срок
в короткий срок
в одно мгновение
в мгновение ока
в момент
в один момент
в один миг
в два счета
одним духом
единым духом
одним пыхом
живой ногой
живой рукой
раз-два и готово
живым манером
как ошпаренный
как ужаленный
стриженая девка косы не заплетет

важно !
много значит
имеет важное значение
имеет большое значение
имеет принципиальное значение
вопрос жизни и смерти

вверх !
в высоту

вверх ногами !
вверх тормашками
вверх тормашки

вдвое !
в два раза

вдвоем !
в паре

вдобавок !
вместе с тем
к тому же
кроме того
кроме всего прочего
мало того
более того
сверх того
сверх всего
в довершение
в довершение всего
к довершению
к довершению всего
ко всему еще
ко всему еще и
в придачу
на придачу
а тут еще

вдохновенно !
с увлечением
с подъемом
с энтузиазмом
вкладывая душу
вкладывая всю душу
с душой
с огоньком

вдребезги !
на мелкие части
на мелкие куски

вдруг !
а ну как
паче чаяния
чего доброго
не ровен час
чем черт не шутит

вереницей !
друг за другом
один за другим
один за одним

верно !
верой и правдой

вероятно !
по всей видимости
по всей вероятности
судя по всему
скорее всего

вернее всего
может случиться
может быть
очень может быть
должно быть
надо думать
надо полагать
как видно
пожалуй что
надо быть
может статься
должно статься
должно полагать
не исключено
к тому идет
к тому дело идет

вроде бы !
в некотором роде
своего рода
словно бы
будто бы
как будто
можно подумать

как ни странно !
можете себе представить

вертикально !
в вертикальном положении

вечером !
в вечернее время
под вечер
к вечеру
на ночь глядя

взаймы !
в долг
с возвратом
в одолжение

взволнованно !
с замиранием сердца
с сердечным замиранием

вместе !
соединенными усилиями
всем скопом
всем миром
в сообществе
в одном строю
в одном ряду
плечо к плечу
плечом к плечу
бок о бок
рука в руке
рука в руку
рука об руку
под одной крышей
в общей сложности
в совокупности
вместе взятые

внешне !
по виду
по внешнему виду
на вид
с виду
на взгляд
из себя

внимательно !
затаив дыхание
стараясь не проронить ни слова
стараясь не пропустить ни слов
стараясь не пропустить ни слов

внутренне !
в глубине души
в глубине сердца

внутри !
в середине
в недрах
в глубине
в утробе
в середке

внутрь !
в середину
в середку

внятно !
с чувством, с толком, с расстановкой

воедино !
в одно целое

возможно !
насколько можно
как только можно
вообще !
не вдаваясь в подробности
не входя в подробности
в общих чертах
в основном
в целом
в общем
в общем и целом
по большому счету
вообще говоря
в общем-то
по идее
в принципе

впервые !
в первый раз

вперевалку !
с развальцем
с перевальцем
с перевальцей
переваливаясь с боку на бок

впереди !
во главе
в первых рядах
в авангарде
на первом месте

вплотную !
грудь в грудь
борт о борт

впору !
подходящего размера
в самый раз
в самую пору
в аккурат
тютелька в тютельку
по росту

впредь !
в дальнейшем
на будущее время

впрок !
на пользу
про запас

и т. д.

ПРИЛОЖЕНИЕ Д

База для синонимических замен отдельных слов (фрагмент)

бог !
боженька
вседержитель
господь
предвечный
всевышний

богородица !
богоматерь
пречистая

божий !
всевышний
господень
господний

коран !
алкоран

троица !
пятидесятница

рассчитаться !
в расчете

абстракция !
абстрагирование
абстракт
отвлечение
умозрение

автоматический !
самодействующий

автомобиль !
автомашина
кар

авторитет !
престиж

авторитетный !
признанный

авторучка !
стило

агрессор !
захватчик

ад !
аид
гадес
геенна
преисподняя
тартар
тартарары

азбука морзе !
морзянка

азиат !
азиатец

аккуратный !
порядливый

активист !
общественник

активно !
деятельно
инициативно

активность !
инициативность
предприимчивость

активный !
деятельный
инициативный
предприимчивый

актриса !
актерка
артистка
комедиантка

актерствовать !
лицедействовать

актуальность !
злободневность

актуальный !
злободневный
наболевший
назревший

акушерка !
повитуха

алкоголь !
спиртное
хмельное

аллюр !
побежка

алмаз !
адамант
бриллиант
диамант

алмазный !
адамантовый
бриллиантовый

алоэ !
столетник

алтарь !
жертвенник

алыча !
ткемали

альпийский !
высокогорный

альпинист !
восходитель
горовосходитель

американец !
янки

американский !
заатлантический
заокеанский

аморалист !
разложенец

анализ !
разбор

анархия !
безвластие
безначалие

английский !
англосаксонский
британский

англичанин !
англосакс
британец
бритт

анонимный письмо !
анонимка

анонимный !
безымянный
неподписанный

антивоенный !
антимилитаристический
антимилитаристский

антиквар !
антикварий
старинщик

антинаучность !
ненаучность

ненаучный !
антинаучный

антисемит !
жидомор
юдофоб

антисемитизм !
юдофобство

антисемитский !
юдофобский

аплодировать !
рукоплескать

аплодисменты ! 2
овация
рукоплескание

арена !
манеж

арест !
задержание

арестовать !
заарестовать
засадить

аристократия ! 1
барин 2

арифметика !
цифирь

аромат !
благовоние

артиллерийский !
орудийный

артиллерист !
пушкарь

артиллерия ! 1
орудие 2
пушка 2

архитектор !
зодчий

архитектура !
зодчество

астролог !
звездочет

атеизм !
безбожие
безверие
неверие

атеист !
безбожник
неверующий

аукцион ! 1
торг 2

афера !
шахер-махер
дирижабль
монгольфьер

бабочка !
мотылек

багрянец !
багрец
пурпур

бакенбарда !
бачки

балахон !
хламида

балерина !
танцовщица

балетмейстер !
хореограф

балетный !
хореографический

баня !
мыльня
сауна

барахтаться !
бултыхаться

барствовать !
сибаритничать
сибаритствовать

басистый !
басовитый

бахрома !
бахромка

бахча !
баштан

бдительно !
неослабно
неусыпно

бдительный !
всевидящий
недреманный
недремлющий
неослабный
неусыпный

бегать !
рыскать

беглость !
незатрудненность

беглый !
незатрудненный
беднеть !
нищать
оскудевать
скудеть
скуднеть

бедно !
мизерно
небогато
нищенски
по-нищенски
скудно
убого

бедность !
необеспеченность
нищенство
нищета
скудность
скудость
убогость
убожество

бедный !
бедняцкий
беспорточный
голопятый
голоштанный
малоимущий
небогатый
необеспеченный
нищенский
нуждающийся
обездоленный
скудный

бедняк !
беднота
беспорточник
голодранец
голоштанник
голытьба
голыш
голь
гольтепа
голяк
паупер

беда !
несчастье
бездолье
горесть
злоключение
злополучие
злосчастие
невзгода
трагедия

бедствовать !
бедовать
перебиваться
скудаться

безбрачие !
целибат

безветренный !
затишный
штилевой

безветрие !
штиль

безвкусие !
безвкусица

безвозвратный !
невозвратимый
невозвратный

безволие !
бесхарактерность
бесхребетность
мягкотелость
слабоволие
слабодушие
слабохарактерность
слюнтяйство
тряпичность

белокровие !
лейкемия

блондин !
белокурый+[сущ]
белобрысый+[сущ]
блондинистый+[сущ]

блондинка !
белокурая+[сущ]
белобрысая+[сущ]
блондинистая+[сущ]

и т. д.

ПРИЛОЖЕНИЕ Ж

База для синонимических замен словосочетаний (фрагмент)

актуальный ! $ крат
стоять на повестка день

предстоять !
на повестка день /(у) род-дат

алкоголь ! 1
спиртной напиток
крепкий напиток
горячительный напиток
зеленый змий
зеленый вино
дар вакх

пьянство !
беспробудный пьянство
беспросыпный пьянство

бедный !
бедный как церковный мышь
бедный как церковный крыса
сирый и убогий
голый как сокол
бедный как ир
бедный как иов
бедный как лазарь
ни грош /(у) род-им end
ни полушка за душа /(у) род-им
ни грош ни копье за душа /(у) род-им
в карман вошь на аркан /(у) род-им
в карман вошь на аркан да блоха на цепь /(у) род-им
ни кол ни двор /(у) род-им
ни кол, ни двор, ни куриный перо /(у) род-им
ни плошка ни ложка /(у) род-им
яко нагой
яко благой
яко нет ничего /(у) род-им
каждый копейка на счету /(у) род-им
считать копейка

безденежье ! 1
стесненный обстоятельство
пустой карман |(с) тв-[в] пр
пустой карман /(у) род-им |[в]
тощий карман |(с) тв-[в] пр
тощий карман /(у) род-им |[в]
карманный чахотка |(с) тв-[в] пр
карманный чахотка /(у) род-им |[в]
денежный чахотка |(с) тв-[в] пр
денежный чахотка /(у) род-им |[в]
ветер свистеть в карман /(у) род-им |[в]
финансы петь романс /(у) род-им |[в]

нуждаться !
испытывать нужда
испытывать потребность
иметь нужда
испытывать нехватка \род-[в] пр
считать копейка

голодать !
жить впроголодь
пухнуть с голод
сидеть голодный
щелкать зуб
класть зуб на полка
положить зуб на полка
лапа сосать
питаться манна небесный
питаться акрида и дикий мед
жить на пища святой антония
сидеть на пища святой антония
жить на антониевый пища
вкушать от пища святой антония
есть нечего /дат-им
кусать нечего /дат-им

испытывать !
подвергать испытание
подвергать проверка
испытывать на себя
узнавать на свой опыт
узнавать на собственный опыт
испытывать на свой шкура
испытывать на собственный шкура
чувствовать на свой шкура
чувствовать на собственный шкура

испытать !
подвергнуть испытание
подвергнуть проверка
испытать на себя
узнать на свой опыт
узнать на собственный опыт
испытать на свой шкура
испытать на собственный шкура

почувствовать на свой шкура
почувствовать на собственный шкура

испить чаша ! № 1,2
испить до дно чаша № 1,4
выпить горький чаша № 1,3
испить горький чаша № 1,3
выпить до дно чаша № 1,4
выпить до дно горький чаша № 1,5

пить чаша ! № 1,2
пить до дно чаша № 1,4

мучиться !
терпеть мучение
испытывать мучение
пить горький чаша
пить до дно горький чаша
принимать мука
принимать много мука
сердце кровь обливаться /(у) род-им
душа надрываться /(у) род-им
душа разрываться /(у) род-им
сердце надрываться /(у) род-им
сердце разрываться /(у) род-им

нет конца !
конца-краю нет
конца-края нет
конца-краю не видно
конца-края не видно
конца-краю не видать
конца-края не видать
конец не видно
не окинуть взгляд /вин-дат
не окинуть взгляд \вин-дат
ни конец ни край не видно
ни конец ни край нет

ходить !
мерить верста

безрассудный ! $ крат
шальной голова
шальной голова /(у) род-им

беспутный !
буйный голова
саврас без узда
забубенный голова

безумствовать !
вести себя безрассудно
поступать безрассудно
поступить безрассудно

тревожиться !
ощущать тревога
ощущать беспокойство
не знать покой
не находить себе место
принимать близко к сердце
сходить с ум от беспокойство
как на иголка
сидеть как на иголка
сидеть как на уголье
сердце замирать /(у) род-им
сердце ныть /(у) род-им
сердце щемить /(у) род-им
сердце не на место /(у) род-им
душа не на место /(у) род-им
кошка скрести на сердце /(у) род-им
кошка скрести на душа /(у) род-им
душа болеть /(у) род-им
сердце болеть /(у) род-им

встревожиться !
ощутить тревога
ощутить беспокойство
лишиться покой
выйти из равновесие
прийти в смятение
сердце сжаться /(у) род-им
сердце екнуть /(у) род-им

спешить !
пороть горячка
спешить на курьерский
спешить как на курьерский
не терпеться /дат-им

бередить душу !
сыпать соль на рана

разбередить душу !
посыпать соль на рана
насыпать соль на рана

беспорядок !
поэтический беспорядок
художественный беспорядок
первозданный хаос
дым коромысло

вавилонский столпотворение
как мамай воевать # наст
как мамай пройти # наст

сложный !
крепкий орешек

неслаженный !
ни склад ни лад

бессемейный !
без род и племя

один ! /мест прил/
в одиночество № 0
в единственный число |пр-им

бессмыслица ! 1
птичий язык
сапог всмятку
колокольня в уксус
сорок бочка арестант
тарабарский грамота

вздор !
завиральный идея
свинячий петрушка
ерунда на постный масло
чепуха на постный масло
бред собачий
чушь собачий
бред сивый кобыла
пустой звук

несомненный !
не вызывающий сомнение № 2
не подлежащий сомнение № 2
ярко выраженный № 2

выделяться !
лезть в глаза
бросаться в глаза

очевидный !
говорить сам за себя

благозвучный !
ласкать слух

нравиться !
радовать глаз
ласкать глаз
радовать взор
внушать симпатия
внушать к себя симпатия
вызывать симпатия
вызывать к себя симпатия
располагать к себя

богатый !
при деньги
деньги счет не знать
деньги кура не клевать /(у) род-им
деньги водиться /(у) род-им
деньги шевелиться /(у) род-им
капитал водиться /(у) род-им
капитал шевелиться /(у) род-им

обогащаться !
составлять состояние
сколачивать состояние
наживать деньги
наживать капитал
набивать карман
набивать кошелек
ковать деньги
загребать деньги лопата
грести деньги лопата
огребать деньги лопата
огребать золото лопата
зашибать деньга

обогатиться !
составить состояние
сколотить состояние
сколотить капитал
нажить деньги
нажить капитал
набить карман
набить кошелек
зашибить деньга
набить деньга
нагреть рука

разбогатеть !
деньги завестись /(у) род-им

изобилие !
птичий молоко не хватать /дат-[у] род

ловкач !
гусь лапчатый
продувной бестия
тонкий штука

палец в рот не класть /дат-им |вин-им

хитрый !
хитрый как лис
с хитринка
себе на ум
на мякина не провести /вин-им # наст
на мякина не провести /род-им # наст
на кривая не объехать /вин-им # наст
на вороная не объехать /вин-им # наст
на саврасый не объехать /вин-им # наст
на коза не объехать /вин-им # наст

болезненный !
подверженный заболевание
в что только душа держаться /(у) род-им
еле-еле душа в тело /(у) род-им

болезненность !
хрупкий здоровье
слабый здоровье
подверженность заболевание

болеть !
припадать здоровье
лежать пласт
лежать как пласт
лежать в лежка
прикованный к постель # наст
быть прикованный к постель
лежать на одер болезнь
тяжело болеть ! <2>
при смерть
быть при смерть

умирать !
отправляться на тот свет
сходить в могила
сходить в гроб
ложиться в могила
ложиться в гроб
давать дуб
отдавать конец
кончать жизнь
уходить из жизнь
расставаться с жизнь
кончать счет с жизнь
отдавать бог душа
уходить в иной мир
уходить в лучший мир
лежать при смерть
при последний издыхание
лежать на смертный одер
день сочтенный /род-им # наст
день быть сочтенный /род-им

и т. д.

ПРИЛОЖЕНИЕ И

Справки и свидетельство

Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное научное учреждение
«ИНСТИТУТ ПРОБЛЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА»
(ФГБНУ «ИПИИ»)

ул. Артёма, д. 118Б, г. Донецк, Донецкая Народная Республика, 283048,
тел.: (856) 311-34-24, e-mail: gu_ipii@mail.ru; http://guiaidn.ru
ОКПО 99649438; ОГРН 1229300155182; ИНН/КПП 9309021845/930901001

01.07.2025 № 173/1/01-01
на № _____ от _____

СПРАВКА
о внедрении результатов
диссертационной работы С.А. Большаковой
на тему: «Совершенствование методов компьютерной обработки текстовой информации в аспекте задач, связанных с омонимией и синонимией», представленной на соискание степени кандидата технических наук по специальности 2.3.1. Системный анализ, управление и обработка информации, статистика (технические науки) в научно-исследовательской деятельности

Результаты диссертационного исследования Большаковой С.А. нашли отражение в проведении научно-исследовательских работ, выполненных отделом распознавания речевых образов ФГБНУ «Институт проблем искусственного интеллекта» с участием автора:

1) Научно-исследовательская работа «Исследование и разработка методов семантического анализа и интерпретации потоков данных интеллектуальными системами» (№ Г/Р 0118D000003) (2018 - 2020 гг.);

2) Научно-исследовательская работа «Исследование и разработка методов снятия омонимии в естественно-языковых текстах внутри парадигмы русского слова» (№ Г/Р 0121D000017) (2021 -2023 гг.);

3) Научно-исследовательская работа «Исследование и разработка методов обработки данных и естественно-языковых текстов с применением онтологий» (№ гос. учета в ЕГИСУ НИОКТР 123092600030-4) (2023 -2025 гг.).

И.о. директора ФГБНУ «ИПИИ» С.Б. Иванова

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ НАУЧНОЕ УЧРЕЖДЕНИЕ
"РЕСПУБЛИКАНСКИЙ АКАДЕМИЧЕСКИЙ НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ И ПРОЕКТНО-КОНСТРУКТОРСКИЙ ИНСТИТУТ ГОРНОЙ ГЕОЛОГИИ, ГЕОМЕХАНИКИ, ГЕОФИЗИКИ И МАРКШЕЙДЕРСКОГО ДЕЛА"
(ФГБНУ "РАНИМИ")

Российская федерация
283001, Донецкая Народная Республика, городской округ Донецкий
город Донецк, ул. Челюскинцев, 291
Тел.: +7 (856) 300 27 91; Тел/факс: +7 (856) 300 27 92
E-mail: ranimi@ranimi@org

05.02.2025 № 04.02-07/34/1
на _____ от _____

СПРАВКА

о внедрении результатов диссертации С.А. Большаковой «Совершенствование методов компьютерной обработки текстовой информации в аспекте задач, связанных с омонимией и синонимией», представленной на соискание ученой степени кандидата технических наук по специальности 2.3.1. Системный анализ, управление и обработка информации, статистика (технические науки)

Методы обработки текстовой информации, разработанные в диссертации С.А. Большаковой «Совершенствование методов компьютерной обработки текстовой информации в аспекте задач, связанных с омонимией и синонимией», используются в отделе компьютерных технологий ФГБНУ «РАНИМИ» для организации формирования и хранения информационных массивов.

И.о. директора
ФГБНУ «РАНИМИ»,
д-р техн. наук В.А. Дрибан

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025611191

Экспериментальное программное обеспечение для морфологической разметки текста со снятием омонимии

Правообладатель: ***Федеральное государственное бюджетное научное учреждение "Институт проблем искусственного интеллекта" (RU)***

Авторы: ***Шелепов Владислав Юрьевич (RU), Ниценко Артем Владимирович (RU), Большакова Светлана Анатольевна (RU)***

Заявка № **2024693205**

Дата поступления **26 декабря 2024 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***16 января 2025 г.***

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0692e7c1a6300bf54f240f670bcc2026
Владелец **Зубов Юрий Сергеевич**
Действителен с 10.07.2024 по 03.10.2025

Ю.С. Зубов