На правах рукописи

Бутенко Юлия Ивановна

МОДЕЛИ И МЕТОДЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ В ПАРАЛЛЕЛЬНОМ КОРПУСЕ

Специальность 2.3.8 –
«Информатика и информационные процессы»

Автореферат
диссертации на соискание ученой степени
доктора технических наук

Москва – 2025

Работа выполнена в Федеральном государственном автономном образовательном учреждении высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)

Официальные оппоненты:

Барахнин Владимир Борисович,
доктор технических наук, доцент, ведущий научный сотрудник Федерального государственного бюджетного научного учреждения «Федеральный исследовательский центр информационных и вычислительных технологий»

Елизаров Александр Михайлович
доктор физико-математических наук, профессор, профессор Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета

Котельников Евгений Вячеславович,
доктор технических наук, доцент, профессор Школы вычислительных социальных наук Автономной некоммерческой образовательной организации высшего образования «Европейский университет в Санкт-Петербурге».

Ведущая организация:

Федеральное государственное бюджетное учреждение науки «Институт проблем управления имени В. А. Трапезникова» Российской академии наук

Защита состоится «____» _____ 2026 г. в _____ часов на заседании диссертационного совета 24.1.224.03 на базе ФИЦ ИУ РАН по адресу: 119333, Россия, Москва, ул. Вавилова, д. 42.

С диссертацией можно ознакомиться в библиотеке ФИЦ ИУ РАН по адресу: Москва, ул. Вавилова, д.42 и на официальном сайте https://www.frccsc.ru

Автореферат разослан «___» _____2025г.

Ученый секретарь
диссертационного совета 24.1.224.03,
к.т.н. И. А. Рейер

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В современном мире важнейшая роль отведена большим данным и методам их анализа, при этом само понятие «большие данные» подразумевает работу с огромными потоками информации, которая регулярно обновляется и поступает из разных источников, с целью увеличения эффективности её функционирования. Примером структурированного представления больших данных являются параллельные корпуса, представленные в виде множества текстов-оригиналов, написанных на каком-либо исходном языке, и текстов-переводов этих исходных текстов на один или несколько других языков.

В настоящее время существует значительное количество параллельных корпусов с разными языковыми парами, совершенствуется технология их формирования, разметки, выравнивания и вывода статистических данных. Однако параллельные корпуса научно-технических текстов, в которых представлены отдельные подкорпуса по узким предметным областям, имеют незначительный объемы размеченных данных, что, с одной стороны, является препятствием для фундаментального описания отдельных языков для специальных целей и как следствие не может отразить их особенности в базах данных и знаний. С другой стороны, подавляющее большинство параллельных корпусов, в которых одним из языков выступает русский, разрабатываются авторами вручную или с использованием ограниченного количества средств разметки текстов, что существенно влияет на их объем.

При этом, постоянное увеличение объема переводных научно-технических текстов свидетельствует о необходимости, с одной стороны, разработки систем автоматического и автоматизированного перевода, а с другой стороны, проведения работ по унификации и стандартизации национальной терминологии. Отсутствие упорядоченных коллекций научно-технических текстов и созданных на их основе терминологических баз данных и знаний существенно тормозит развитие и совершенствование средств искусственного интеллекта по автоматической обработке научно-технических текстов.

Анализ современных параллельных корпусов показал, что они создаются лингвистами вручную, что требует значительных временных затрат на выполнение рутинных процедур разметки и выравнивания параллельных текстов. Вместе с тем, широкий спектр применения параллельных корпусов при решении ряда теоретических и практических задач свидетельствует о необходимости создания таких информационных ресурсов. Эта отрасль является недостаточно формализованной, слабо автоматизированной, существующие методы работы не универсальны, а операции по разметке научно-технических текстов выполняют лингвисты собственноручно отдельно для каждого вида разметки. Следовательно, с целью автоматизировать и ускорить процедуру автоматической обработки научно-технических текстов при создании параллельных корпусов необходима инструментальная поддержка процедуры обработки параллельных научно-технических текстов.

Таким образом, разработка теоретических основ построения

информационных моделей и методов решения задач автоматической обработки научно-технических текстов при создании параллельного корпуса с применением методов структурного, терминологического и стилистического моделирования является актуальной проблемой и имеет существенное научное и хозяйственное значение.

Проведенные в диссертационной работе исследования находятся в русле приоритетного направления развития науки, технологий и техники РФ «Информационно-телекоммуникационные системы», а также соответствуют критической технологии РФ «Нано-, био-, информационные, когнитивные технологии».

Степень разработанности темы диссертационного исследования.

Современное состояние корпусной лингвистики, а также аспекты создания, использования и обработки параллельных корпусов текстов представлены в работах отечественных и зарубежных ученых, а именно В.П. Захарова, С. Ю. Богдановой, М.Г. Кружкова, М.В, Хохловой, Д.В. Сичинавы, С.О. Шереметьева, Ч. Сяохуэй, M. Barlow, O. Bojar, M. Scott, P. Rayson. Особенности структурной разметки текстовых документов описаны в работах О.А. Горбань, М.В. Косовой, О.С. Ринчиновой, М.Ю. Мухиной, И. Яна. Обработке терминологических единиц в текстах на естественном языке посвящены труды Н.В. Лукашевич, Э.С. Клышинского, Н.А. Кочетковой, И.О. Кузнецова, Н.А. Астраханцева, Terryn A., S. Janicke. Выявление машинно-переведенных и машинно-сгенерированных текстов представлено в работах Ю.В. Чеховича, М.Н. Черкасовой, В.В. Николаева, G. Jawahar, L. Dugan. Семантическая разметка текстов стала предметом изучения таких ученых C. Fillmore, C. Baker, M. Palmer, D.Gildea. Е.Б. Козеренко, Н.Ф. Хайрова, Л.Д. Бадмаева, T. Yousef, S. Janicke, G. Neubig занимались вопросами выравнивания текстов в параллельных корпусах.

Однако в работах указанных ученых не приведены пути автоматической обработки англо- и русскоязычных научно-технических текстов в аспекте создания параллельного корпуса. Вместе с тем, создание такого корпуса поспособствует развитию подходов к обработке естественного языка за счет формирования филологически корректной базы данных размеченных текстов на русском и английском языках. Более того, в указанных работах не используются способы разноуровневого анализа научно-технических текстов, что на сегодняшний день является наиболее перспективным направлением при обработке естественно-языковых текстов.

В данной работе применен комплексный подход к проблеме обработки научно-технических текстов на русском и английском языках. Теоретические положения и практические результаты, полученные в ходе выполнения данного исследования, основаны на идеях зарубежных и отечественных специалистов в области искусственного интеллекта и лингвистики, не противоречат сути языковых явлений, а также не накладывают ограничений на естественный язык, использованный в научно-технических текстах, что является отличительной особенностью и преимуществом данной работы.

Объектом исследования в работе является модели, методы и

программные средства обработки научно-технических текстов.

Предмет исследования: разработка моделей, методов и программных средств обработки композиционной структуры научно-технических текстов, автоматического извлечения многокомпонентных терминов и номенклатурных наименований, выявления машинно-сгенерированных и машинно-переведенных русскоязычных научно-технических текстов в аспекте создания параллельного корпуса.

Цель и задачи исследования. Целью исследования является повышение эффективности автоматической разметки и выравнивания лингвистических единиц разной формальной структуры в параллельном корпусе путем автоматизации процесса обработки научно-технических текстов.

Для достижения цели были поставлены и решены следующие задачи:

- представление и обработка иерархически-структурированных научно-технических текстов;
- представление структурного состава терминологических словосочетаний, а также разработка способов их разметки и выравнивания в параллельных научно-технических текстах;
- представление структурного состава номенклатурных наименований, а также разработка способов их разметки и выравнивания в параллельных научно-технических текстах;
- представление способов выявления машинно-сгенерированных и машинно-переведенных научно-технических текстов или их фрагментов в параллельном корпусе.
- разработка инструментальных средств и прикладной технологии обработки научно-технических текстов при создании параллельного корпуса научно-технических текстов;
- практическая реализация разработанных моделей, методов и инструментальных средств для решения прикладных задач специальной и учебной лексикографии, информационного поиска и обработки коллекций текстов на английском и русском языках.

Методы исследования. Для решения поставленных задач в диссертации используются: методология системного анализа, методы компьютерной лингвистики, машинного обучения, информационного поиска, математической статистики, программной инженерии.

Научная новизна. Научной новизной проведенного исследования являются теоретические основы построения моделей и создания методов обработки англо- и русскоязычных научно-технических текстов, направленные на проектирование параллельного корпуса, что имеет важное хозяйственное значение в области информатики, а именно:

1. Усовершенствованы модели иерархически-структурированных научно-технических текстов, за счет добавления межуровневых элементов и оценки значимости каждого структурного элемента при создании параллельного корпуса, что позволяет более эффективно обрабатывать научно-технические тексты на разных уровнях языковой системы.
2. Получили дальнейшее развитие модели и методы разметки и

выравнивания англо- и русскоязычных терминологических единиц из научно-технических текстов, отличающиеся от существующих возможностью извлечения терминов с правыми определениями, что позволяет использовать эти модели и методы при обработке текстов при создании параллельного корпуса.

3. Впервые разработаны модели и метод разметки номенклатурных наименований в научно-технических текстах на русском и английском языках, что позволяет повысить эффективность разметки научно-технических текстов за счет учета лексических единиц, в состав которых входят произвольные буквенно-числовые последовательности в том числе символы разных алфавитов.

4. Впервые предложены методы выявления машинно-сгенерированных и машинно-переведенных текстов на основе семантико-синтаксических особенностей русского языка.

5. Разработан прототип системы управления корпусными данными, который в отличие от существующих корпусных менеджеров позволяет управлять корпусными данными на разных этапах их обработки, а также формировать различные наборы данных для машинного обучения.

Теоретическая значимость работы. Полученная научная новизна вносит развитие в аппарат теоретической информатики в области решения важной научной проблемы автоматической обработки научно-технических текстов. Методические результаты работы могут быть использованы в системах автоматической обработки естественных языков для специальных целей и при разработке различных информационно-поисковых систем широкого назначения.

Практическая значимость работы. Практическая ценность работы заключается в создании программного средства, позволяющего использовать разработанные теоретические основы, модели и методы обработки англо- и русскоязычных научно-технических текстов для создания параллельного корпуса. Данное программное средство позволяет автоматизировать рутинный процесс обработки англо-и русскоязычных параллельных текстов и увеличить объемы филологически компетентных баз данных размеченных научно-технических текстов.

Положения, выносимые на защиту:

1. Концепция, базовые принципы и стратегия создания параллельного корпуса, отличающиеся новой научной идеей обработки языковых объектов как системы взаимосвязанных компонентов при обработке научно-технических текстов.

2. Модели композиционной структуры научно-технических текстов, использующихся как источники для наполнения параллельного корпуса научно-технических текстов.

3. Модели англо- и русскоязычных многокомпонентных терминологических единиц и методы их разметки и выравнивания в параллельном корпусе научно-технических текстов.

4. Модели англо- и русскоязычных номенклатурных наименований и метод их разметки в параллельном корпусе научно-технических текстов.

5. Методы выявления машинно-сгенерированных и машинно-переведенных текстов или их фрагментов в научно-технических текстах на

основе актуального членения предложения в русском языке.

6. Концепция и прототип системы управления корпусными данными параллельного корпуса англо- и русскоязычных научно-технических текстов.

Степень достоверности результатов. Достоверность научных результатов работы подтверждается непротиворечивостью и согласованностью с известными фактами и исследованиями в рассматриваемой области, высокой степенью сходимости теоретических результатов с данными экспериментов и определяется применением теоретических и методологических основ разработок ведущих ученых в области обработки естественного языка, корректным и обоснованным использованием математического аппарата, экспериментальными исследованиями разработанных моделей и методов

Соответствие диссертации паспорту специальности. Тема и основные результаты диссертации соответствуют следующим областям исследований паспорта специальности 2.3.8 – Информатика и информационные процессы.

2 Техническое обеспечение информационных систем и процессов, в том числе новые технические средства сбора, хранения, передачи представления информации. Комплексы технических средств, обеспечивающих функционирование информационных систем и процессов, накопления и оптимального использования информационных ресурсов.

5 Лингвистическое обеспечение информационных систем и процессов. Методы и средства проектирования словарей данных, словарей индексирования и поиска информации, тезаурусов и иных лексических комплексов. Методы семантического, синтаксического и прагматического анализа текстовой информации для представления в базах данных и организации интерфейсов информационных систем с пользователями.

11 Разработка принципов организации и технологий реализации систем управления базами данных и знаний, создание специализированных информационных систем управления текстовыми, графическими и мультимедийными базами данных. Создание языков описания данных, языков манипулирования данными, языков запросов.

Апробация результатов диссертации. Основные результаты работы докладывались и обсуждались на X Международной научно-практической конференции студентов, аспирантов и молодых ученых «Информационные технологии в науке, бизнесе и образовании» (Москва, 2018), Всероссийской научной конференции «Нейрокомпьютеры и их применение» (Москва, 2018, 2019, 2020, 2022), II Всероссийской национальной научной конференции студентов, аспирантов и молодых ученых «Молодежь и наука: актуальные проблемы фундаментальных и прикладных исследований» (Комсомольск-на-Амуре, 2019, 2020), Международной научно-практической конференции «Современное технологическое образование» (Москва, 2019), II Всероссийской научно-практической конференции «Системы управления полным жизненным циклом высокотехнологичной продукции в машиностроении: новые источники роста» (Москва, 2019), Международном форуме «Цифровые технологии в инженерном образовании: новые тренды и опыт внедрения» (Москва, 2019), XII Всероссийской конференции молодых ученых и специалистов (с

международным участием) «Будущее машиностроения России» (Москва, 2019), Международной научной конференции «Фундаментальные и прикладные задачи механики», посвященная 100-летию со дня рождения Академика Константина Сергеевича Колесникова (Москва, 2019), Международном молодежном научном форуме «ЛОМОНОСОВ-2020» (Москва, 2020), Академических чтениях по космонавтике «Королевские чтения» (Москва, 2019, 2021, 2022), Международной конференции «Моделирование в инженерном деле» (Москва, 2019, 2020, 2022), II Всероссийской молодёжной научно-практической конференции с международным участием «LinguaNet» (Севастополь, 2020), Межвузовской заочной конференции аспирантов, соискателей и молодых ученых «Наука, технологии и бизнес» (Москва, 2020, 2022, 2024), VI Международном форуме «Instrumentation Engineering, Electronics and Telecommunications – 2020» (Ижевск, 2020), II Международной научно-практической конференции «Лингвистические и культурологические аспекты современного инженерного образования» (Томск, 2021), International Conference «Aviation Engineering and Transportation» (AviaEnT) (Иркутск, 2020), XXI Международной научно-технической конференции «Развитие информатизации и государственной системы научно-технической информации» (РИНТИ-2022) (Минск, 2022), Международной ИТ-конференции «Ключевые тренды развития искусственного интеллекта: наука и технологии» (Москва, 2023).

Публикации. По теме диссертации опубликовано 60 научных работ, из которых 27 статей в научно-технических журналах, входящих в перечень ВАК, 20 – в изданиях, входящих в международные наукометрические базы Scopus и Web of Science. В трудах российских и международных конференций опубликовано 29 работ.

Личный вклад соискателя. Все выносимые на защиту результаты и положения, составляющие основное содержание диссертационного исследования, разработаны и получены лично автором или при его непосредственном участии. В работах, опубликованных в соавторстве, соискателю принадлежит определяющая роль при решении задач развития теоретических основ создания информационных моделей и методов обработки научно-технических текстов. В работах [9, 11, 12, 16, 29, 38, 39, 50] соискателю лично принадлежит общий подход к извлечению англо- и русскоязычных многокомпонентных терминов на основе синтаксических шаблонов, подкрепленных морфологической информацией о каждой словоформе. В работах [6, 43] соискателем предложен подход к созданию учебных и специальных словарей на основе параллельного корпуса научно-технических текстов, в работах [8, 17, 44] соискателем предложен общий подход к установлению семантических ролей в научно-технических текстах, принципы семантико-синтаксического анализа научно-технических текстов. В работах [25, 27, 60] соискателю лично принадлежит принципиальная постановка задачи анализа композиционной структуры научно-технических текстов и проработка основных подходов к их анализу. В работах [5, 13, 21, 23, 30, 32, 36, 45, 47, 52, 59] соискателем проработаны общие принципы информационного поиска в сложно-структурированных научно-технических текстах на английском и

русском языках. В работах [37, 40, 46, 48, 49, 51, 53, 55, 58] соискателем предложены различные подходы к использованию параллельного корпуса в лингвистике и лингводидактике. В работах [18, 22, 26, 31, 33, 34, 54, 56, 57] автор принимал участие при создании баз данных интеллектуальных систем в аспектах анализа научно-технических текстов.

Структура и объем работы. Диссертация состоит из введения, 6 разделов, заключения, списка использованных источников, содержащего 298 наименований. Основная часть работы содержит 304 страницы, включая 103 рисунка и 41 таблицу.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность темы диссертации, сформулирована цель исследования, указаны основные задачи и научные положения, выносимые на защиту, охарактеризованы научная новизна, теоретическая и практическая значимость, указаны используемые методы исследования.

Первая глава посвящена обзору и анализу современного состояния исследований в области обработки научно-технических текстов при создании и наполнении корпусов текстов, а также применимости существующих моделей, методов, программных средств обработки научно-технических текстов в аспекте создания параллельного корпуса.

В **разделах 1.1-1.3** представлен обзор одно-и многоязычных параллельных и специальных корпусов, а также рассмотрены особенности и основные этапы создания параллельных корпусов. Анализ степени автоматической разметки и выравнивая текстов в рассмотренных корпусах приведен в таблице 1.

Таблица 1. Анализ степени автоматизации обработки текстов в параллельных и специальных корпусах

	Корпуса	**Кол-во**	**Выравнивание** (всего/автоматически)	**Разметка** (всего/автоматически)
1	Параллельные корпуса с русским языком	10	По предложениям (10/6)	Морфологическая (4/4), метатекстовая (2/0), семантическая (1), структурно-дискурсная (1/0), библиограф. (1/0)
2	Параллельные многоязычные корпуса	36	По предложениям (27/4), по абзацам (5/0), по текстам (4)	Не размечены (8), токенизация (11/11), морфологич. (7/7), семантич. (2/0), синтаксич. (2/0), фонетич. (1/0),
3	Параллельные двуязычные корпуса	54	По предложениям (45/14), по абзацам (2/2), по текстам (7)	Не размечены (14), токенизация (21/21), морфологич. (14/14), синтаксич.(3/1), семантич. (1/0), терминологич. (1/0)
4	Специальные корпуса	40	-	Не размечены (10), лемматизация (10/10), морфологич.(23/23), терминологич (3/1), структурн.(6/0), синтаксич. (10/6), семантич.(6/0), метаразметка (11).

На основе проведенного анализа следует, что жанровое разнообразие текстов в рассмотренных корпусах имеет широкий разброс от

общеупотребительной и художественной тематикой текстов до отдельных узких отраслей для разных языковых пар. При этом, параллельные корпуса, с одной стороны, широко используются при проведении различных исследований, а с другой стороны, выявлено, что они имеют значительные ограничения в объеме, тематике, спектре языков, а также средствах автоматической обработки текстов для наполнения корпусов.

Кроме того, основываясь на результатах проведенного анализа, не было выявлено ни одного программно-реализованного репрезентативного параллельного корпуса научно-технических текстов, где бы одним из одним из языков выступал русский.

В **разделе 1.4** показано, что существующие средства автоматической обработки текстов имеют ряд ограничений, которые на практике оказывают влияние на обработку научно-технических текстов при создании параллельного корпуса, так как не учитывают их отличительные особенности. Так, научно-технические тексты имеют ярко выраженную композиционную структуру, которая не может быть отображена стандартными средствами структурной разметки, учитывающей только деление текста на абзацы, приложения и слова. На лексическом уровне ключевой особенностью научно-технических текстов является использование специальной терминологии. При этом, несмотря на значительные успехи в аспекте автоматического извлечения терминов из специальных текстов, существующие подходы и программные средства не рассчитаны на извлечение терминов из текстов на нескольких языках одновременно, а также не учитывают некоторые виды специальной лексики такие как номенклатурные наименования, которые в своем составе содержат буквенно-числовые последовательности (БЧП) с использованием разных алфавитов.

Далее в разделе проанализированы различные виды разметок, обоснована целесообразность их реализации в параллельном корпусе научно-технических текстов. Обоснована необходимость разметки машинно-сгенерированных и машинно-переведенных текстов или их фрагментов, так как такие виды текстов могут на практике привести к искажению полученных результатов при решении целого ряда практических задач, основанных на корпусных данных.

Обзор исследований, связанных с автоматизированной обработкой научно-технических текстов при создании параллельного корпуса, и апробированных программных средств подтверждает актуальность и новизну разработанных в диссертационном исследовании моделей, методов и программных средств. На основании проведенного анализа в выводах к первой главе диссертации поставлены задачи диссертационного исследования

Вторая глава посвящена созданию моделей научно-технических текстов. Источниками для наполнения параллельного корпуса отобраны следующие виды текстов: научно-технические статьи, учебники, нормативная документация. Отличительной особенностью перечисленных текстов является наличие строгой иерархической структуры и ограниченного перечня элементов, входящих в тексты, а также закрепление определённого места в тексте за каждым из элементов.

В **разделе 2.1.** проанализированы языковые особенности изложения текста научно-технической статьи, его композиционная структура, а на их основе предложена модель текста статьи. В нотациях Бекуса-Наура композиционную структуру текстов научно-технических статей можно задать следующим образом:

$$St_i ::= \langle X^1, X^2, X^3 \rangle$$

где X^1 – реферативный раздел научно-технической статьи, X^2 – корпус научно-технической статьи, X^3 – информативный раздел научно-технической статьи.

X^1 – реферативный раздел научно-технической статьи, состоящий из следующих элементов:

$$X^1 ::= \langle x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17} \rangle | \langle x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17} \rangle$$

где x_{11} – код УДК, x_{12} – название статьи, x_{13} – информация об авторах, x_{14} – место работы авторов, x_{15} - контактная информация авторов, x_{16} – аннотация, x_{17} – ключевые слова.

X^2 – корпус научно-технической статьи можно представить в виде набора из следующих элементов:

$$X^2 ::= \langle x_{21}, x_{22}, x_{23}, x_{24}, x_{25} \rangle | \langle x_{21}, x_{22}, x_{23}, x_{25} \rangle | \langle x_{21}, x_{23}, x_{25}, \rangle | \langle x_{21}, x_{23}, x_{24}, x_{25}, \rangle$$

где x_{21} – введение, x_{22} – материал и методы, x_{23} – результаты, x_{24} – обсуждение результатов, x_{25} - заключение.

X^3 – информативный раздел научно-технической статьи, для которого справедливо

$$X^3 ::= \langle x_{31}, x_{32} \rangle | \langle x_{32} \rangle$$

где x_{31} – примечания, x_{32} – ссылки на источники.

На Рис.1 представлена полученная структурная схема элементов текста научно-технической статьи.

Рис. 2 – Структурные элементы текста научно-технической статьи

В **разделе 2.2** рассмотрены учебные тексты как источники для наполнения параллельного корпуса. Композиционная структура учебно-научного текста в первом приближении состоит из реферативного раздела, корпуса научно-технической статьи и информативного раздела и ее можно задать следующим образом:

$$St_i ::= \langle X^1, X^2, X^3 \rangle$$

где X^1 – реферативный раздел учебно-научного текста, X^2 – корпус учебно-научного текста, X^3 – информативный раздел научно-учебного текста.

X^1 – реферативный раздел учебно-научного текста, состоящий из следующих элементов:

$$X^1 ::= \langle x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, \rangle | \langle x_{11} x_{12}, x_{13}, x_{14} \rangle$$

где x_{11} – название, x_{12} – автор(ы), x_{13} – оглавление, x_{14} – введение, x_{15} - предисловие.

X^2 – корпус учебно-научного текста, который можно представить в виде набора из следующих элементов:

$$X^2 ::= \langle x_{21}, x_{22}, x_{23} \rangle | \langle x_{21}, x_{22} \rangle | \langle x_{21}, x_{23} \rangle | \langle x_{21}, \rangle$$

где x_{21} – основной текст, x_{22} – вопросы, x_{23} – задания и упражнения.

X^3 – информативный раздел учебно-научного текста, для которого справедливо

$$X^3 ::= \langle x_{31}, x_{32}, x_{33} \rangle | \langle x_{31}, x_{33} \rangle | \langle x_{32}, x_{33} \rangle | \langle x_{33} \rangle$$

где x_{31} – примечания, x_{32} – приложения, x_{33} - ссылки на источники.

На Рис.2 представлена полученная структурная схема элементов учебно-научного текста.

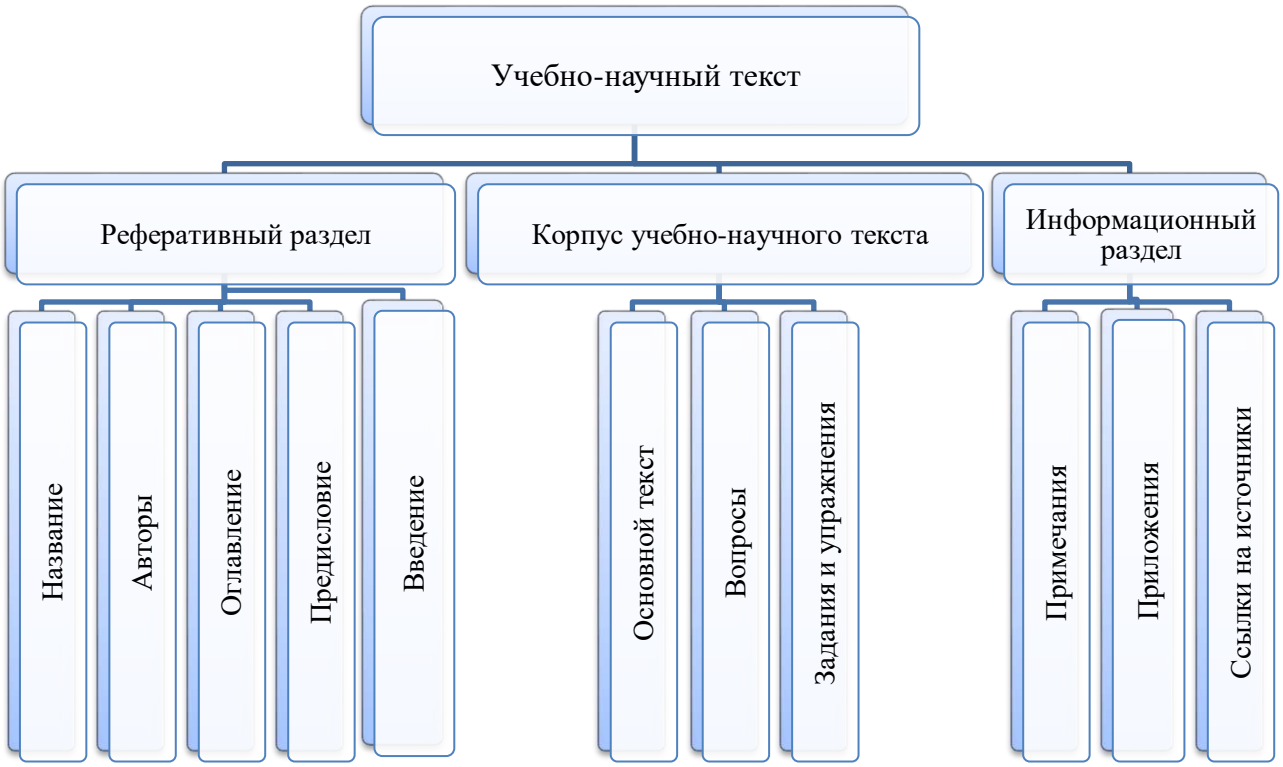

Рис. 2 – Структурные элементы учебно-научного текста

В **разделе 2.3** проанализированы тексты нормативной базы,

композиционную структуру которых можно задать следующим образом:

$$St_i ::= \langle X^1, X^2, X^3 \rangle | \langle X^1, X^2 \rangle,$$

где X^1 – предварительная часть стандарта, X^2 – часть требований и рекомендаций, X^3 – информативная часть.

Предварительная часть стандарта X^1 состоит из следующих элементов:

$$X^1 ::= \langle x_1, x_2, x_3, x_4 x_5, x_6, x_7, x_8, x_9 \rangle | \langle x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 \rangle | \langle x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_9 \rangle |$$
$$| \langle x_1, x_2, x_3, x_4, x_5, x_7, x_8, x_9 \rangle | \langle x_1, x_2, x_3, x_5, x_6, x_7, x_8, x_9 \rangle | \langle x_1, x_2, x_4, x_5, x_6, x_7, x_8, x_9 \rangle |$$
$$| \langle x_1, x_2, x_5, x_6, x_7, x_8, x_9 \rangle | \langle x_1, x_2, x_4, x_5, x_7, x_8, x_9 \rangle | \langle x_1, x_2, x_4, x_5, x_6, x_8, x_9 \rangle | \langle x_1, x_2, x_4, x_5, x_6, x_7, x_9 \rangle |$$
$$| \langle x_1, x_2, x_4, x_5, x_6, x_7, x_8 \rangle | \langle x_1, x_2, x_5, x_7, x_8, x_9 \rangle | \langle x_1, x_2, x_5, x_6, x_8, x_9 \rangle | \langle x_1, x_2, x_5, x_6, x_7, x_9 \rangle |$$
$$| \langle x_1, x_2, x_5, x_6, x_7, x_8 \rangle | \langle x_1, x_2, x_5, x_7, x_8, x_9 \rangle | \langle x_1, x_2, x_5, x_8, x_9 \rangle | \langle x_1, x_2, x_5, x_7, x_9 \rangle | \langle x_1, x_2, x_5, x_7, x_8 \rangle |$$
$$\langle x_1, x_2, x_5 \rangle,$$

где x_1 – титульный лист, x_2 - предисловие, x_3 – содержание, x_4 – введение, x_5 - название, x_6 –область применения, x_7 – нормативные ссылки, x_8 – термины и определения понятий, x_9 –обозначения и сокращения.

X^3 – информативная часть, для которой справедливо

$$X^3 ::= \langle x_{11} \rangle | \langle x_{12} \rangle | \langle x_{11}, x_{12} \rangle$$

где x_{11} – приложение, x_{12} – библиографические данные. На Рис.3 представлена полученная структурная схема элементов стандарта.

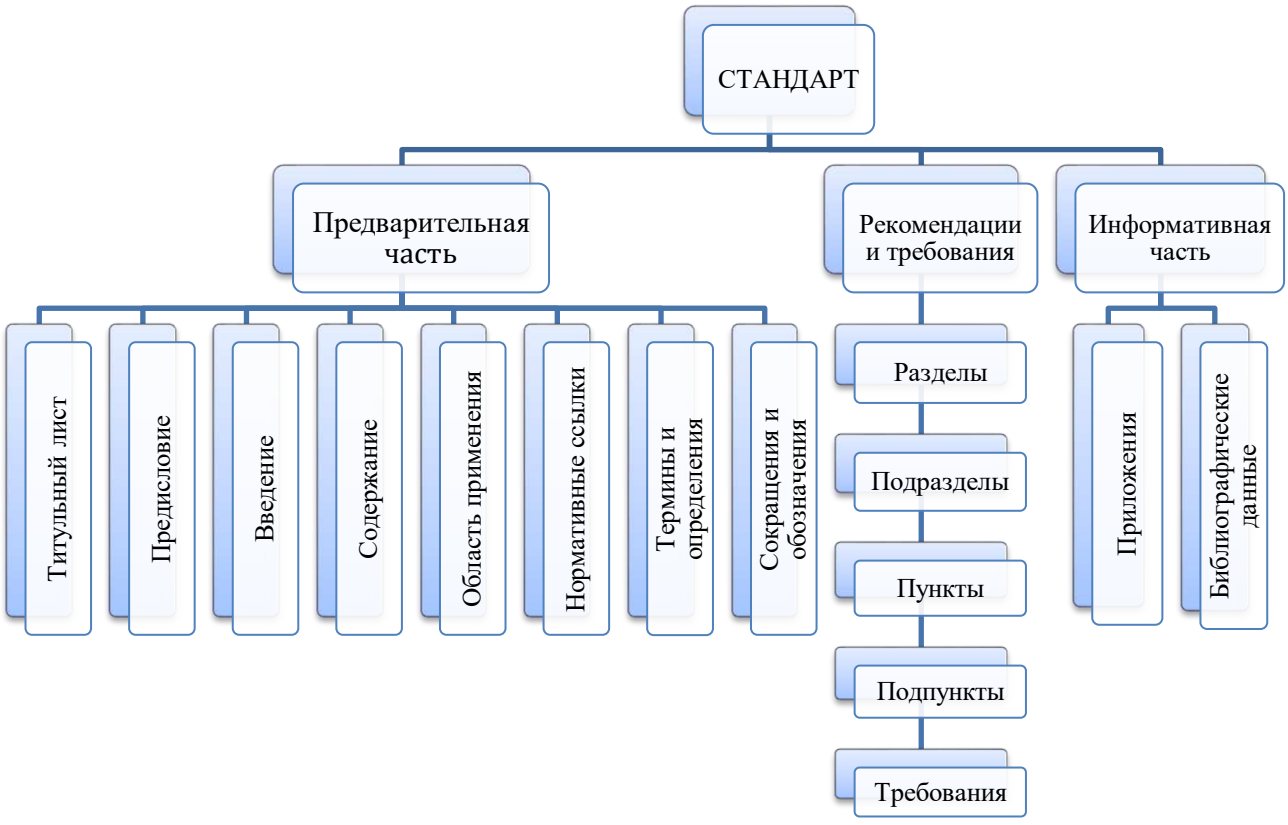

Рис. 3 – Структурные элементы текста стандарта

На основе проведенного анализа композиционных структур научно-технический текст St целесообразно представить в виде:

$$St = \langle E^L, R \rangle$$

где E – структурный элемент, R – отношения между структурными элементами, L – уровень структурного элемента. При этом $L = \{l_1, \dots l_8\}$, де l1 – документ, l2 – основная часть, l3 – раздел, l4 – подраздел, l5 – пункт, l6 – подпункт, l7 – абзац, l8 – предложение.

В **разделе 2.4.** показано, что представление научно-технического текста в виде упорядоченного набора структурных элементов дает возможность анализировать с помощью математических методов как вероятностную, так и логическую структуру всего текста.

Наличие структурной разметки научно-технических текстов при создании параллельного корпуса значительно расширяет исследовательский потенциал корпуса, что в свою очередь позволяет при разработке систем обработки естественного языка учитывать композиционные особенности научно-технических текстов, в целом, и их отдельных структурных компонентов, в частности. Так использование структурной разметки при обработке текстов научно-технических статей позволяет сократить более чем на 25% объем терминов-кандидатов за счет обработки только значимой части текста статьи и отсечения таких структурных элементов как данные об авторах и их аффилиации, списка использованных источников, иллюстративных примеров и т.д. (Табл. 2).

Таблица 2. Анализ эффективности разметки терминов с предварительной структурной разметкой

		терминов-кандидатов			
			основная часть	из них не обраб.	
	Язык НТТ	всего		терминов	%
1	Русский язык	653	480	173	**26,6**
2	Английский язык	742	515	227	**30,3**

Третья глава посвящена разработке моделей и методов разметки и выравнивания многокомпонентных русско- и англоязычных терминов с правыми определениями и номенклатурных наименований в параллельном корпусе научно-технических текстов.

В **разделе 3.1** представлена обобщенная структура многокомпонентного термина, состоящая из ядерного элемента, левого и правых определений, где ядерный элемент – главный элемент в структуре многокомпонентного термина, который вступает в словоизменительную парадигму в предложении. Левое и правое определения уточняют значение ядерного элемента. При этом левое определение чаще всего выражено именами прилагательными, которые наследуют грамматические признаки рода, числа и падежа имени существительного, перед которым стоят, а правые определения выражены именными группами, которые стоят после ядерного элемента, и при этом их грамматические характеристики остаются неизменными (Рис 4).

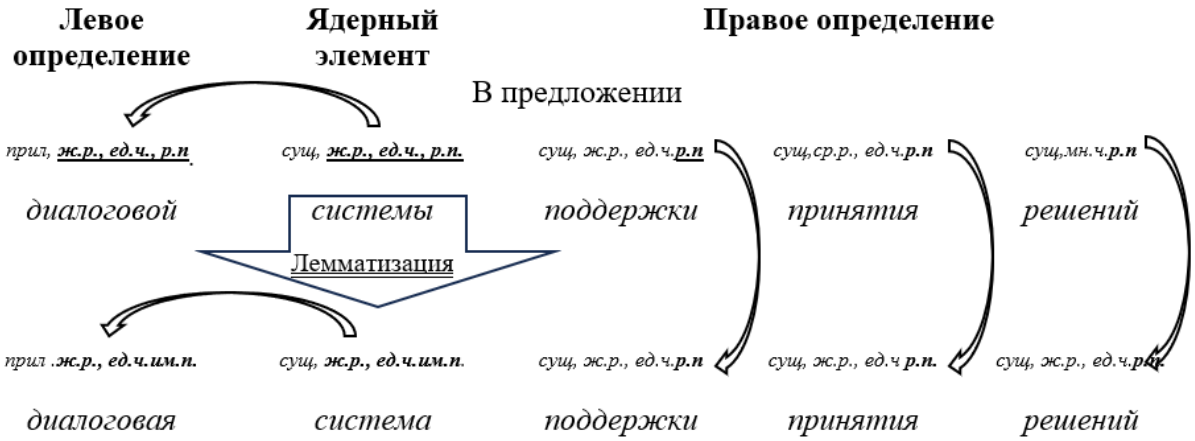

Рис. 4. Схема наследования грамматических признаков при нормализации многокомпонентного термина

Выделено 24 модели русскоязычных и 21 модель для англоязычных многокомпонентных терминов, в которых указаны грамматические характеристики компонентов термина, например, «имя прилагательное + имя существительное (ядерный элемент) + имя существительное в творительном падеже» (*холодная сварка давлением)*. При создании моделей англоязычных терминов целесообразно ориентироваться на порядок слов, а ядерным элементом всегда будет последнее существительное или последнее существительное перед предлогом, если он есть, например, термин *diffusion welding in vacuum* имеет модель «имя существительное + имя существительное (ядерный элемент) + предлог + имя существительное».

В **разделе 3.2** проанализированы номенклатурные наименования как особый класс специальной лексики. Номенклатурное наименование – это терминологическое обозначение частотного специального понятия какой-либо области знания, дисциплины или тематической области, которое состоит из двух лексико-синтаксических компонентов, синтаксически главный из которых является термином, словом или словосочетанием общего языка и обозначает специальное родовое понятие данной области, а синтаксически подчиненный – является условным, внешним знаком, номеном и служит для выделения из родового понятия именно данного частного понятия, фиксируемого в специальных описаниях, толкованиях и единицах, например, *разгонная ступень Блок ДМ-3, ракета-носитель Протон-М, РСЗО Град.* Англо- и русскоязычные номенклатурные наименования образованы по следующим моделям: аббревиатура + слово, аббревиатура + слово, аббревиатура + буквенно-числовая последовательность (БЧП), термин + слово, термин + слово, термин + БЧП, аббревиатура + слово + БЧП, аббревиатура + термин + БЧП, аббревиатура + слово + БЧП, аббревиатура + термин + слово, термин + слово + БЧП. В указанных моделях термин – это лексическая единица, входящая в специальный словарь корпуса, слово – любая лексическая единица английского или русского языков, БЧП – произвольный набор римских или арабских цифр, букв русского или английского алфавитов, других символов.

В **разделе 3.3.** предложен метод к автоматической разметке англо- и

русскоязычных многокомпонентных терминов на основе предложенных моделей терминологических словосочетаний, представленный на рис. 5.

Рис. 5. Метод разметки многокомпонентных терминов в научно-технических текстах

В качестве примера рассмотрим извлечение терминов для следующего фрагмента текста: *Как наука химия природных соединений возникла одновременно с органической химией.*

1. На первом этапе проводим морфологический анализ:

Как *союз* *наука* *СУЩ,неод,жр ед,им*, *химия* *неод,жр ед,им* *природных* *ПРИЛ,кач мн,рд* *соединений* *СУЩ,неод,ср мн,рд* *возникла* *ГЛ,сов,неперех жр,ед,прош,изъяв* *одновременно* *нар* *с* *предл* *органической* *ПРИЛ жр,ед,рд* *химией* *СУЩ,неод,жр ед,ТВ*.

2. Убираем части речи, которые не входят в состав терминов:

~~*Как* *союз*~~ *наука* *СУЩ,неод,жр ед,им*, *химия* *неод,жр ед,им* *природных* *ПРИЛ,кач мн,рд* *соединений* *СУЩ,неод,ср мн,рд* ~~*возникла* *ГЛ,сов,неперех жр,ед,прош,изъяв*~~ ~~*одновременно* *нар* *с* *предл*~~ *органической* *ПРИЛ жр,ед,рд* *химией* *СУЩ,неод,жр ед,ТВ*.

Таким образом, остаются следующие цепочки слов:

наука *СУЩ,неод,жр ед,им*

химия *неод,жр ед,им* *природных* *ПРИЛ,кач мн,рд* *соединений* *СУЩ,неод,ср мн,рд*

органической *ПРИЛ жр,ед,рд* *химией* *СУЩ,неод,жр ед,ТВ*.

3. Проверяем полученные термины-кандидаты на наличие в них «стоп-слов», указанных в специальной зоне словаря.

4. Полученные цепочки слов соотносим с шаблонами терминологических

словосочетаний, имеющихся в базе структурных моделей терминов.

наука *СУЩ,неод,жр ед,им* -принадлежит к моделям терминов

химия *неод,жр ед,им* *природных* *ПРИЛ,кач мн,рд* *соединений* *СУЩ,неод,ср мн,рд*

органической *ПРИЛ жр,ед,ТВ* *химией* *СУЩ,неод,жр ед,ТВ*

Полученные термины-кандидаты проверяем по словарю корпуса. Если термин-кандидат есть в словаре корпуса, то добавляем контекст его употребления в словарную статью. В противном случае на основе шаблона словарной статьи создаем новую для найденного термина и добавляем контекст его употребления.

В **разделе 3.4** описан метод разметки номенклатурных наименований, основные этапы реализации которого представлены на рис. 6.

Рис 6. Метод разметки номенклатурных наименований в научно-технических текстах

Рассмотрим разметку номенклатурных наименований в русскоязычных научно-технических текстах на примере следующего фрагмента текста:

Приведены результаты реконструкции вращательного движения малого спутника Аист-2Д по данным бортовых измерений векторов угловой скорости и напряженности магнитного поля Земли, полученным летом 2016 г.

На первом и втором этапах проводится морфологическая и терминологическая разметки текста. Ядерные элементы многокомпонентных терминов обозначены *.

Приведены КР_ПРИЧ,сов,прош,страд мн результаты СУЩ,неод,мр мн,им реконструкции СУЩ,неод,жр ед,рд ***[вращательного ПРИЛ ср,ед,рд движения СУЩ,неод,ср ед,рд] [малого ПРИЛ,ср ед,рд спутника* СУЩ,неод,мр ед,рд]*** *Аист СУЩ,од,мр ед,им -2Д НЕИЗВ по ПР данным СУЩ,неод,хр,pl мн,дт* ***[бортовых ПРИЛ мн,рд измерений* СУЩ,неод,ср мн,рд]*** *векторов СУЩ,неод,мр мн,рд* ***[угловой ,ПРИЛ,мр ед,вн скорости*СУЩ,неод,жр ед,рд]*** *и СОЮЗ напряженности СУЩ,неод,жр ед,рд* ***[магнитного ПРИЛ,кач ср,ед,рд поля* СУЩ,неод,ср ед,рд Земли СУЩ,неод,жр ед,рд]*** *, ЗПР полученным ПРИЧ,сов,перех,прош,страд ср,ед,тв летом СУЩ,неод,ср,sg ед,тв 2016 ЧИСЛО,цел г СУЩ,неод,мр,0,аббр ед,рд. ЗПР*

После каждого выделенного термина проверяется наличие имен существительных, не входящих в состав термина или буквенно-числовых последовательностей, которые морфологический анализатор размечает как неизв., т.е. неизвестный объект. В связи с тем, что некоторые части не входят в состав номенов, то в рассматриваемом случае до распознанной части речи, в рассматриваемом примере это предлог.

Приведены КР_ПРИЧ,сов,прош,страд мн результаты СУЩ,неод,мр мн,им реконструкции СУЩ,неод,жр ед,рд ***[вращательного ПРИЛ ср,ед,рд движения СУЩ,неод,ср ед,рд] [малого ПРИЛ,ср ед,рд спутника* СУЩ,неод,мр ед,рд]*** ***Аист СУЩ,од,мр ед,им -2Д НЕИЗВ*** *по ПР данным СУЩ,неод,хр,pl мн,дт* ***[бортовых ПРИЛ мн,рд измерений* СУЩ,неод,ср мн,рд]*** *векторов СУЩ,неод,мр мн,рд* ***[угловой ,ПРИЛ,мр ед,вн скорости*СУЩ,неод,жр ед,рд]*** *и СОЮЗ напряженности СУЩ,неод,жр ед,рд* ***[магнитного ПРИЛ,кач ср,ед,рд поля* СУЩ,неод,ср ед,рд Земли СУЩ,неод,жр ед,рд]*** *, ЗПР полученным ПРИЧ,сов,перех,прош,страд ср,ед,тв летом СУЩ,неод,ср,sg ед,тв 2016 ЧИСЛО,цел г СУЩ,неод,мр,0,аббр ед,рд. ЗПР*

На следующем этапе полученный кандидат-номенклатурное наименование проверяем по моделям номенклатурных наименований: *[малого ПРИЛ,ср ед,рд спутника* СУЩ,неод,мр ед,рд] Аист СУЩ,од,мр ед,им -2Д НЕИЗВ* совпадает с моделью «термин + слово + БЧП».

Оценка корректности извлечения терминов на основе предложенного метода произведена на основе 20 текстов научно-технических статей по космонавтике, опубликованных в журнале «Космические исследования» в 2018-2019 гг. Оценка эффективности метода извлечения терминов (Таблица 3) проведена путем сравнения с методом извлечения терминов на основе синтаксических шаблонов. Прочерк в таблице 3 означает отсутствие синтаксических шаблонов для терминов такой структуры. Учтены только уникальные термины независимо от их частотности вхождения в тексты научно-технических статей.

Таким образом, повышение эффективности извлечение терминов из научно-технических текстов происходит за счет добавления моделей терминов с правыми определениями, а также использованию дополнительных грамматических характеристик многокомпонентных терминов, заложенных в обновленные синтаксические шаблоны.

Таблица 3. Оценка качества метода извлечения терминов из научно-технических текстов, в %

Количество компонентов термина	Кол-во уникальных терминов	Синтаксические шаблоны			Усовершенствованные синтаксические шаблоны		
		Полнота	Точность	F-мера	Полнота	Точность	F-мера
1	235	93	60	72	93	60	72
2	293	91	72	80	91	72	80
3	209	60	67	63	**89**	**91**	**89**
4	58	-	-	-	**92**	**90**	**90**
5	25	-	-	-	**94**	**89**	**91**
6	17	-	-	-	**95**	**87**	**90**
Всего	837		Среднее	72		Среднее	83

В **разделе 3.5** рассмотрено два сценария выравнивания терминов в параллельных текстах: в первом сценарии сначала тексты были выровнены пословно, а затем применен предложенный выше метод извлечения терминов на основе синтаксических шаблонов. Во втором сценарии использовался метод «мешка терминов», а затем выравнивание полученных списков терминов на русском и английском языках. Для выравнивания лексических единиц использован пакет SimAlign, разработанный на основе нейронной сети BERT.

Пакет выравнивает пары слов на основе встроенного словаря, как показано на рис. 7 для следующего списка:

1. часть
2. шкала абсолютных потенциалов
3. двойственный характер
4. совместное использование
5. величина
6. поверхностный потенциал
7. ESa
8. существование промежуточных частиц
9. адатом
10. вакансия
11. абсолютный потенциал
12. eia
13. обратимая электродная реакция
14. использование абсолютного потенциала
15. водородный электрод

1. part
2. scale of absolute potentials
3. dual character
4. sharing
5. absolute
6. surface potentials ESa
7. existence of intermediate particles
8. adatoms
9. vacancies
10. absolute potentials
11. reversible electrode reactions
12. use of absolute potential
13. PZC
14. hydrogen
15. transition

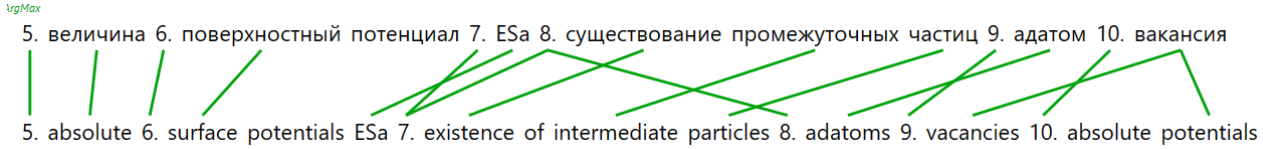

Рис. 7. Выравнивание элементов терминов в пакете SimAlign

Таким образом, список выровненных терминов на русском и английском языках имеет следующий вид:

1. часть	1. part
2. шкала абсолютных потенциалов	2. scale of absolute potentials
3. двойственный характер	3. dual character
4. совместное использование	4. sharing
5. поверхностный потенциал	5. surface potentials ESa
6. существование промежуточных частиц	6. existence of intermediate particles
7. адатом	7. adatoms
8. вакансия	8. vacancies
9. абсолютный потенциал	9. absolute potential
10. использование абсолютного потенциала	10. use of absolute potential

Повышение эффективности выравнивания многокомпонентных терминов достигается за счет изменения сценария извлечения терминов и охвата терминов с правыми определениями (Табл. 4).

Таблица 4. Оценка эффективности сценариев выравнивания терминов с использованием пакета SimAlign

	Сценарий	**Полнота**	**Точность**	**F-мера**
1	Выравнивание слов + применение шаблонов	61	71	65
2	«Мешок терминов» + выравнивание	**84**	**77**	**81**

Для языковой пары английский-русский более низкая эффективность первого метода объясняется тремя факторами: во-первых, разной длиной и формальной структурой терминологических единиц в русском и английском языках; во-вторых, разницей в синтаксической структуре англо- и русскоязычных предложений; в-третьих, широким использованием переводческих трансформаций.

В четвертой главе описаны методы разметки стилистических и переводческих особенностей параллельных текстов, а именно выявления и разметки машинно-сгенерированных и машинно-переведенных русскоязычных научно-технических текстов и их фрагментов.

В **разделе 4.1** обосновано, что особенности актуального членения предложения в русском языке являются основой для выявления машинно-сгенерированных текстов. Прямой порядок слов присущ всем синтаксическим конструкциям английского языка, а русскоязычное предложение строится на особенностях организации тема-рематических отношений в высказывании. Для анализа актуального членения предложения использован машинный переводчик DeepL, который переводит каждое русскоязычное предложение на английский язык, а затем на основе пакета SimAlign, реализовано выравнивание слов каждого предложения. Результаты работы современных машинных переводчиков показывают, что они способны обнаруживать в русскоязычных предложениях главные и второстепенные члены предложения и трансформировать русскоязычное предложение в соответствии с особенностями порядка слов в предложении на английском языке. На рис. 8 показано

соответствие порядка слов в русскоязычном предложении его переводном эквиваленте.

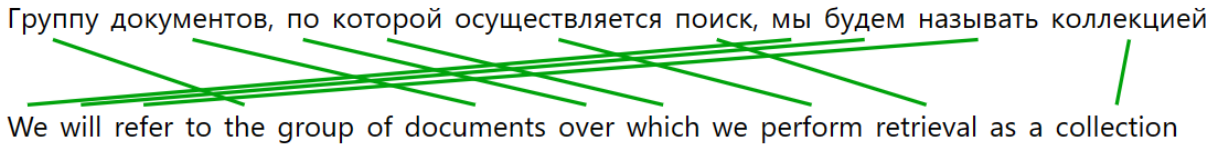

Рис. 8. Выравнивание слов в русскоязычно предложении и его переводном эквиваленте

При анализе синтаксических структур машинно-сгенерированных текстов выявлено, что русскоязычные предложения имеют прямой порядок слов, т.е. совпадают с порядком слов в английском предложении, как показано на рис. 9.

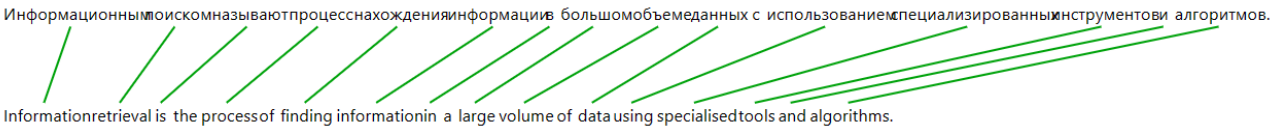

Рис. 9. Пример выравнивания слов в машинно-сгенерированном тексте

В таблице 5 представлен сравнительный анализ изменения порядка слов в машинно-сгенерированных и написанных человеком русскоязычных текстах. В таблице отражены только факты наличия в предложениях наибольших изменений в структуре одного предложения, при этом их количество не учтено. Иными словами, если в предложении есть несколько случаев изменения порядка слов, то учитывается наибольшее значение, например, для рис. 8 – это 6, а для рис. 9 – 0, то есть полное совпадение порядка слов.

Таблица 5. Сравнительный анализ изменения порядка слов в машинно-сгенерированных и написанных человеком текстах

		Всего предл.	Полное совпадение порядка слов, предл.	Наличие изменения структуры, предл.			
				1	2	3	4 и более
	Машинно-сгенери-рованный текст	200	108 (55%)	34 (18%)	22 (12%)	8 (5%)	18 (10%)
	Текст, написанный человеком	200	44 (22%)	42 (21%)	6 (3%)	28 (14%)	80 (40%)

В **разделе 4.2** предложен метод выявления машинно-сгенерированных текстов на основе особенностей актуального членения русскоязычного предложения. Для установления соответствия между порядком слов в русскоязычном предложении и его переводной версии вычисляется «расстояние» выровненных пар слов. Если слово в одном языке не выровнено со словом в другом языке, то его пропускают, а выровненные слова нумеруются. Изменением порядка слов будем считать несовпадение порядка следования пар

слов. Как следует из Рис. 10 слова *views* (3) и *рассматривается* (4) выровнены, а их позиции имеют расхождение в 1, так же как и пара слов just (6) и просто (5) также будет иметь измените порядка следования слов на 1.

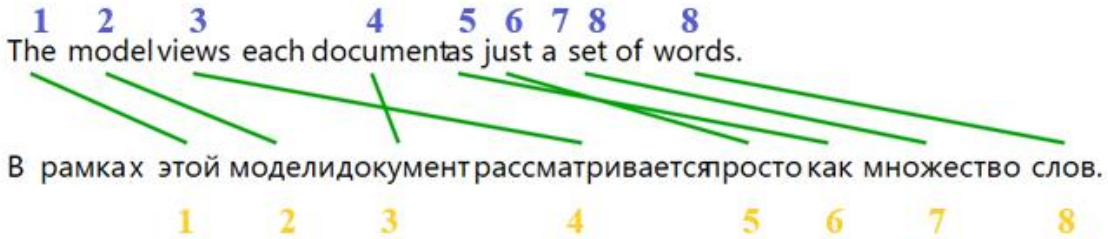

Рис. 10. Установление соответствия порядка слов в русском и английском языках

Метод выявления машинно-сгенерированных русскоязычных текстов или их фрагментов представлен на рис. 11 .

Наиболее яркими показателями машинно-сгенерированных текстов выступают полное совпадение порядка слов или незначительные изменения в порядке слов со сдвигом в 1-2 слова при машинном анализе, что в сумме составляет порядка 85% случаев. При этом изменения в порядке слов в русскоязычных предложениях со сдвигом в 3 более позиции слова – 54% случаев. Отсюда следует, что средства генеративного искусственного интеллекта не могут обрабатывать семантические аспекты русскоязычного высказывания, которые выражены в особенностях построения предложения, и поэтому при построении предложений порядок слов такой же, как и у языков с фиксированным порядком слов.

Рис. 11. Метод выявления машинно-сгенерированных фрагментов русскоязычных текстов

Раздел 4.3 посвящен анализу переводческих трансформаций как маркеров переведённых вручную текстов. Машинно-переведенные тексты с английского языка на русский также сохраняют прямой порядок. Кроме того, к различиям между машинным и ручным переводом относятся так называемые переводческие трансформации, которые используют лингвисты при выполнении ручного перевода. Использование переводческих трансформаций можно увидеть при выравнивании фрагментов англо-и русскоязычного предложения. На рис. 12 показаны 2 переводческие трансформации, грамматическая - «*морфонологическое расщепление =Morpheme split into*» и лексическая – добавление слова *distinct* в английском языке.

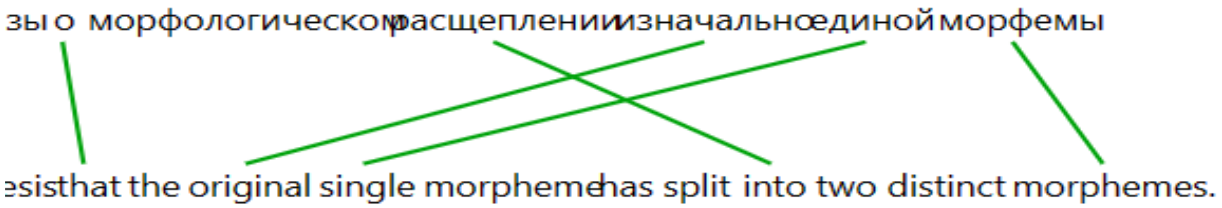

Рис. 12. Выявление непереведенных лингвистических единиц в параллельных текстах

При машинном переводе этого же предложения переводческие трансформации отсутствуют (Табл. 6). Стоит отметить, что эта особенность не в меньшей степени свойственна и переводу с русского языка на английский.

Таблица 6. Сравнительный анализ изменения порядка слов в машинно-сгенерированных и написанных человеком текстах

	Всего предл.	Полное совпадение порядка слов, предл.	Наличие изменения структуры, предл.				Не переведенная лексика
			1	2	3	4 и более	
Машинно-переведенный текст	200	86	44	30	19	21	0
Текст, переведенный человеком	200	44	42	6	28	80	82

В **разделе 4.4** предложен метод выявления машинно-переведенных русскоязычных текстов или их фрагментов, представленный на рис. 13.

Рис. 13. Метод выявления машинно-переведенных фрагментов русскоязычных текстов

В **разделе 4.5** описан статистический подход для выявления машинно-сгенерированных и машинно-переведенных текстов. Предложенный подход позволяет учитывать стилистические особенности научно-технических текстов разных видов, отраслей, жанров, написанных разными авторами. Так, например, тексты по математическим наукам в силу широкого использования формул могут быть по статистическим параметрам близки к машинно-сгенерированным текстам, а для текстов по гуманитарным наукам – значения для написанных человеком и машинно-сгенерированных текстов показывают выраженную корреляцию. Таким образом, целесообразно формировать нормальную генеральную совокупность текстов нужного типа, а затем проверять соответствие некоторой выборки текстов нормальной генеральной совокупности.

В таблице 7 приведены усредненные значения расхождений синтаксических структур в русскоязычных предложениях и их переводных эквивалентов из текстов научно-технических статей по космонавтике. С помощью χ^2 - критерия Пирсона с количеством интервалов *k*, равным *6* затем можно проверить соответствие выборки текстов нормальной генеральной совокупности.

$n = 49$;

Размах выборки: $R_n = \mathrm{x}_{\max} - \mathrm{x}_{\min} = 23.676 - 10.036 = 13.64$

Для числа интервалов $k = 6$:

Таблица 7. Значения расхождений синтаксических структур в научно-технических текстах

10.036	10.195	10.897	11.254	11.576	12.547	12.608	13.700	13.834
13.837	13.962	14.024	14.040	14.095	14.116	14.150	14.175	14.365
14.379	14.441	14.470	14.546	14.841	15.967	16.005	16.185	16.274
16.489	16.799	16.898	17.217	17.566	17.843	17.893	17.928	17.934
18.127	18.201	18.447	18.554	18.593	18.705	19.249	19.374	19.700
21.333	22.250	23.129	23.676					

Длина интервала

$$d = \frac{R}{k} = \frac{13.64}{6} = 2.27$$

Интервалы в общем виде:

$$\Delta_1 = (-\infty; x_{\min} + d);\ \Delta_2 = (x_{\min} + d; x_{\min} + 2d);\ \Delta_3 = (x_{\min} + 2d; x_{\min} + 3d);$$
$$\Delta_4 = (x_{\min} + 3d; x_{\min} + 4d);\ \Delta_5 = (x_{\min} + 4d; x_{\min} + 5d);\ \Delta_6 = (x_{\min} + 5d; \infty)$$

После подстановки:

$$\Delta_1 = (-\infty;12.306);\ \Delta_2 = (12.306;14.576);\ \Delta_3 = (14.576;16.846);$$
$$\Delta_4 = (16.846;19.116);\ \Delta_5 = (19.116;21.386);\ \Delta_5 = (21.386;+\infty)$$

Количество n_i элементов выборки, попавших в каждый интервал:

$$n_1 = 5 \qquad n_2 = 17 \qquad n_3 = 7 \qquad n_4 = 13 \qquad n_5 = 4 \qquad n_6 = 3$$

Поскольку $n_5 < 5$ и $n_6 < 5$ объединяем 5$^{\text{й}}$ и 6$^{\text{й}}$ интервалы, тогда:

$$\Delta_1' = (-\infty;12.306); \quad \Delta_2' = (12.306;14.576); \quad \Delta_3' = (14.576;16.846);$$
$$\Delta_4' = (16.846;19.116); \quad \Delta_5' = (19.116;+\infty)$$
$$n_1' = 5 \qquad n_2' = 17 \qquad n_3' = 7 \qquad n_4' = 13 \qquad n_5' = 7$$

Вероятность попадания в каждый интервал:

$$P_i = \Phi\left(\frac{h_i - \bar{x}}{S}\right) - \Phi\left(\frac{h_{i-1} - \bar{x}}{S}\right)$$

$h_0 = -\infty; \quad h_1 = 12.306; \quad h_2 = 14.576, \quad h_3 = 16.846, \quad h_4 = 19.116, \quad h_5 = +\infty$

$$P_1 = \Phi\left(\frac{12.306 - 16.05}{3.14}\right) - \Phi\left(\frac{-\infty - 16.05}{3.14}\right) = \Phi(-1.19) - \Phi(-\infty) = 0.116$$

$$P_2 = \Phi\left(\frac{14.576 - 16.05}{3.14}\right) - \Phi\left(\frac{12.306 - 16.05}{3.14}\right) = \Phi(-0.469) - \Phi(-1.19) = 0.204$$

$$P_3 = \Phi\left(\frac{16.846 - 16.05}{3.14}\right) - \Phi\left(\frac{14.576 - 16.05}{3.14}\right) = \Phi(0.254) - \Phi(-0.469) = 0.280$$

$$P_4 = \Phi\left(\frac{19.116 - 16.05}{3.14}\right) - \Phi\left(\frac{16.846 - 16.05}{3.14}\right) = \Phi(0.976) - \Phi(0.254) = 0.235$$

$$P_5 = \Phi\left(\frac{\infty - 16.05}{3.14}\right) - \Phi\left(\frac{19.116 - 16.05}{3.14}\right) = 1 - \Phi(0.976) = 0.165$$

Величины $C_i = np_i$

$C_1 = 49 \cdot 0.116 = 5.684; \quad C_2 = 49 \cdot 0.204 = 9.996; \quad C_3 = 49 \cdot 0.280 = 13.720;$

$C_5 = 49 \cdot 0.235 = 11.515; \quad C_5 = 49 \cdot 0.165 = 8.085$

Значение статистики:

$$\chi^2 = \sum_{i=1}^{5} \frac{(n_i - n_{pi})^2}{n_{pi}} = \frac{(5-5.684)^2}{5.684} + \frac{(17-9.996)^2}{9.996} + \frac{(7-13.72)^2}{13.72} + \frac{(13-11.515)^2}{11.515} + \frac{(7-8.085)^2}{8.085} =$$

$$= 0.082 + 4.907 + 3.29 + 0.19 + 0.145 = 8.614$$

Число степеней свободы: $\nu = k - l - 1; \quad k' = 5; \quad l = 2 \Rightarrow \nu = 2$

По таблице распределения χ^2 при уровне значимости $\alpha = 0.01$

$$\chi^2_{0.999}(2) = 9.21$$

Поскольку $\chi^2 > \chi^2_{0.999}(2)$, гипотезу о том, что данная выборка извлечена из нормальной генеральной совокупности можно принять на уровне значимости $\alpha = 0.01$.

Пятая глава посвящена описанию исследовательского прототипа системы управления корпусными данными параллельного корпуса научно-технических текстов.

В **разделе 5.1** описана система управления корпусными данными параллельного корпуса, в которой реализованы предложенные в диссертации модели и методы обработки русско- и англоязычных научно-технических текстов. Основные компоненты системы представлены на рис. 14.

Принципиальные отличия предложенной системы управления корпусных данных от существующих аналогов состоят в следующем:

- возможность видеть результаты разметок и вносить в них изменения;
- наличие средств автоматической разметки научно-технических текстов на разных языковых уровнях;
- возможность сохранения и дальнейшей обработки конкорданса;
- «золотой стандарт» терминологической разметки;
- создание датасетов с разными уровнями разметки в зависимости от решаемой задачи;
- наличие собственного словаря корпуса.

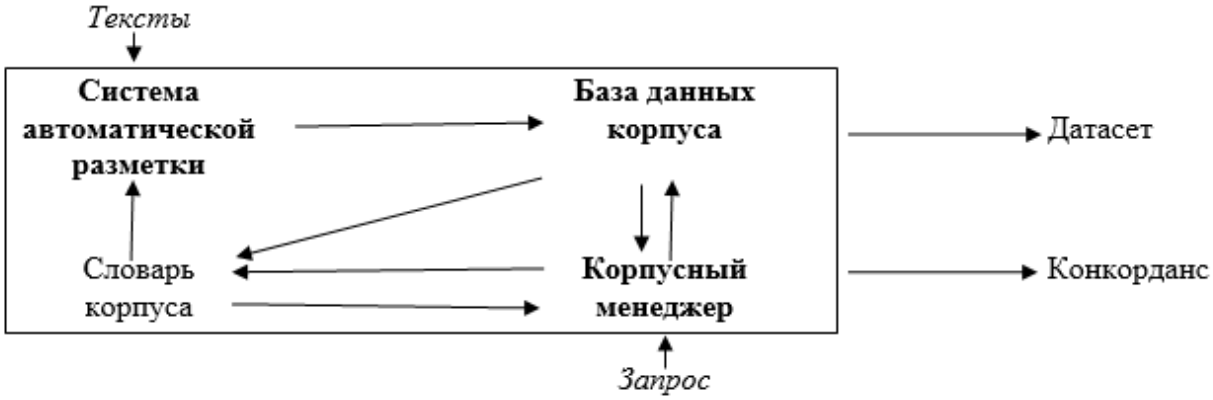

Рис. 14. Система управления корпусными данными параллельного корпуса научно-технических текстов

В **разделе 5.2** показана информационная технология обработки научно-технических текстов, размещаемых в параллельном корпусе.

Этапы обработки научно-технических текстов в параллельном корпусе показаны на рис. 15.

Рис. 15. Этапы обработки научно-технических текстов в параллельном корпусе

В **разделе 5.3** описаны программные средства автоматической разметки текстов, а именно: разметка и выравнивание терминов, структурная разметка и выравнивание, стилистическая разметка, семантико-синтаксическая разметка.

На рисунке 16 показаны результаты разметки научно-технических текстов, где каждый вид разметки представлен отдельным слоем.

В **разделе 5.4** показан исследовательский потенциал параллельного корпуса в изучении особенностей научно-технических текстов лингвистике и информатике. Показано, что обработка многокомпонентного термина как одной лексической единицы повышает естественность научно-технических текстов на 3-15%, то есть до показателей текстов других жанров и стилей. На основе анализа способов выражения модальности в английском языке и их перевода на русский в параллельных текстах научно-технических статей вывялены переводные эквиваленты, не зафиксированные в переводных словарях, а также случаи их некорректного перевода.

Структурная	Введение, абзац 2, предложение 1						Введение, абзац 2, предложение 2				
Стиличтическая	0						0				
Анафорическая				[боковая ручка* управления][1]			[она][1]				
Семантическая роль	агент	предикат	объект	инструменталис			объект	предикат			локутив
Терминологическая	[летчик]		[самолет]	[боковая ручка* управления]							[истребитель]
Морфологическая	летчик	пилотировал	самолет	боковой	ручкой	управления	Она	используется	в	современных	истребителях
	Летчик пилотировал самолет боковой ручкой управления.						*Она используется в современных истребителях.*				

Структурная	Введение, абзац 2, предложение 1								Введение, абзац 2, предложение 2						
Анафорическая							[side control stick][1]		[it][1]						
Семантическая роль		агент	предикат		объект		инструменталис								локутив
Терминологическая		[pilot]			[aircraft]		[side control stick*]								[fighter jet*]
Морфологическая	the	pilot	piloted	the	aircraft	with	side	control	stick	It	is	used	in	modern	fighter jets
	The pilot piloted the aircraft with the side control stick.								*It is used in modern fighter jets.*						

Рис. 16.

В шестой главе описаны варианты использования разработанных моделей и методов для решения практических задач в области лингвистики и информатики. В **разделах 6.1 и 6.2** показаны подходы к практическому использованию предложенных в диссертации методов обработки научно-технических текстов в сферах цифровой и учебной терминографии; в **разделе 6.3.** – выявления тенденций развития научных направлений на основе публикаций; **разделе 6.4** – формировании нормативных профилей при сертификации продукции, **разделе 6.5** – анализе студенческих работ на предмет использования систем генерации текстовых последовательностей, выявлении использования машинных переводчиков при переводе научно-технических текстов и др.

В **заключении** изложены итоги выполненного исследования, рекомендации и перспективы дальнейшей разработки темы.

В **приложении** представлены документы, подтверждающие использование результатов диссертации в науке и промышленности.

ВЫВОДЫ

В диссертационной работе была решена крупная научная проблема, имеющая важное хозяйственное значение в области информатики, которая заключается в необходимости автоматизации процедуры обработки научно-технических текстов в параллельном корпусе. Теоретические основы обработки научно-технических текстов расширены концепцией, базовыми принципами и стратегией обработки текстов как целостной структуры. Разработан комплекс информационных моделей лингвистических единиц разного уровня и методов их обработки, структурирования и интеграции. Основные результаты диссертационной работы перечислены в следующих пунктах:

1. Определены научные направления развития подходов и методов обработки научно-технических текстов при наполнении параллельного корпуса, выявлены основные проблемы, обоснована актуальность разработки. Сформулированы постановки основных задач исследования.

2. Развиты теоретические основы обработки научно-технических текстов, включающие в себя концепцию, базовые принципы и стратегию, отличающиеся новой научной идеей обработки языковых объектов как системы взаимосвязанных компонентов, что крайне важно для создания систем понимания текстов на естественном языке, а сам параллельный корпус за счет фиксации различий в плане выражения при одинаковом плане содержания может способствовать развитию подходов к фиксации смыслового содержания научно-технических текстов.

3. Выделены структурные модели русско- и англоязычных многокомпонентных терминов, при формировании перечня которых учтены ошибки, возникающие при осуществлении автоматической морфологической разметки корпусов текстов, а также предложен метод разметки англо- и русскоязычных многокомпонентных терминов в корпусе научно-технических текстов на основе структурных моделей терминологических единиц, который использует морфологические, синтаксические, лексические и семантические

ограничения при обработке англо- и русскоязычных научно-технических текстов, определяет ядерный элемент многокомпонентного термина. а при выравнивании терминов опирается на модели структурных трансформаций терминов в переводе.

4. Получены структурные модели англо- и русскоязычных номенклатурных наименований, которые учитывают не только самые разнообразные варианты структур номенклатурных наименований, но и вариации в их написании и возможностях обработки морфологическими анализаторами, а также разработан метод разметки англо- и русскоязычных номенклатурных наименований в параллельных научно-технических текстах, основой которого являются морфологическая и терминологическая разметки.

5. Предложены методы выявления машинно-сгенерированных и машинно-переведенных текстов, которые в отличие от существующих реализованы не на методах машинного обучения, а учитывают семантико-синтаксические особенности русского языка, а также предложен статистический подход к обработке научно-технических текстов, учитывающий их разные жанры и направления, в аспекте выявления машинных текстов.

6. Разработан прототип системы управления корпусными данными, который в отличие от существующих корпусных менеджеров позволяет управлять корпусными данными на разных этапах их обработки, а также формировать различные наборы данных для машинного обучения.

В целом совокупность полученных в диссертации теоретических и практических результатов позволяет сделать вывод о том, что цель исследований достигнута, сформулированная научная проблема решена. Перечисленные результаты получили высокую оценку научного сообщества при апробации и положительные рекомендации для внедрения в информационные процессы предприятий, учреждений и организаций различного профиля деятельности.

Основные результаты диссертации опубликованы в следующих работах:

Статьи в рецензируемых изданиях из списка ВАК РФ

1. Бутенко Ю. И. Метод выявления русскоязычных машинно-сгенерированных текстов по особенностям актуального членения предложения // Научно-техническая информация: Серия 1. Организация и методика информационной работы. 2025. №6. С. 19-26. DOI: 10.36535/0548-0019-2025-06-3. **(К-1**, переводная версия **Scopus, WoS)**

2. Бутенко Ю. И. Метод извлечения многокомпонентных терминологических единиц с правыми определениями из научно-технических текстов // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2024. Т. 22, № 3. С. 5-14. DOI: 10.25205/1818-7900-2024-22-3-5-14. **(К-1)**

3. Бутенко Ю. И. Извлечение номенклатурных наименований из англо- и русскоязычных научно-технических текстов // Искусственный интеллект и принятие решений. 2024. №3. С. 95-103. DOI:10.14357/20718594240309. **(RSCI)**

4. Бутенко Ю. И. Метод выравнивания многокомпонентных

терминологических единиц в параллельном корпусе научно-технических текстов // Научно-техническая информация: Серия 2. Информационные процессы и системы. 2024. №8. С. 29-38. DOI: 10.36535/0548-0027-2024-08-4. **(RSCI**, переводная версия **WoS)**

5. Бутенко Ю. И., Сидняев Н. И., Синева Е. Е. Стратегии поиска в пространстве состояний // Научно-техническая информация: Серия 2. Информационные процессы и системы. 2024. №6. С. 25-39. DOI:10.36535/0548-0027-2024-06-4. **(RSCI**, переводная версия **WoS)**

6. Бутенко Ю. И., Солошенко К. А. Лексический тренажер по иностранному языку для студентов технических специальностей МГТУ им. Н.Э. Баумана // Экономика. Информатика. 2024. №51(1). С. 189–200. DOI: 10.52575/2687-0932-2024-51-1-189-200. **(К-1)**

7. Бутенко Ю. И. Использование базы данных структурных трансформаций для извлечения многокомпонентных терминологических единиц // Системы и средства информатики. 2023.Т. 33, №1. С.35-44. DOI: 10.14357/08696527230104. **(RSCI)**

8. Бутенко Ю. И., Галетка М. Л. Синева Е. Е. Создание системы разметки семантических ролей в научно-технических текстах по авиации и космонавтике // Научно-техническая информация: Серия 2. Информационные процессы и системы. 2022. №10. С. 23-32. DOI:10.36535/0548-0027-2022-10-4. **(RSCI**, переводная версия **WoS)**

9. Бутенко Ю.И., Николаева Н.С., Карцева Е.Ю. Структурные модели англоязычных терминов для автоматической обработки корпусов научно-технических текстов // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2022. Т.14. №1. С. 80-95. DOI: 10.22363/2313-2299-2022-13-1-80-95. (**Scopus)**

10. Бутенко Ю. И. Модель научно-технической статьи для разметки в корпусе научно-технических текстов // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2022. №3 (20). С. 5-13. DOI: 10.25205/1818-7900-2022-20-3-5-13. **(К-1)**

11. Бутенко Ю. И. Строганов Ю. В. Сапожков А. М. Система извлечения многокомпонентных терминов и их переводных эквивалентов из параллельных научно-технических текстов // Научно-техническая информация: Серия 2. Информационные процессы и системы. 2022. №9. С. 12-21. DOI: 10.36535/0548-0027-2022-09-3. **(RSCI)**

12. Бутенко Ю.И., Лукьянова Г.О. Особенности разметки научно-технических текстов в аспекте создания специализированного корпуса // Филологические науки. Научные доклады высшей школы. 2022. №1. С.14-20. DOI 10.20339/PhS.1-22.014 **(WoS)**

13. Бутенко Ю. И., Тельнова И. Н., Гаража В. В. Методы выявления тенденций развития научных направлений (на материале анализа публикаций по газовому топливу) // Научно-техническая информация: Серия 2. Информационные процессы и системы. 2022. №1. С. 10-24. DOI: 10.36535/0548-0027-2022-01-2. **(RSCI**, переводная версия **WoS)**

14. Бутенко Ю. И. Использование онтологий для автоматизации

формирования нормативного профиля при сертификации программного обеспечения // Искусственный интеллект и принятие решений. 2021. №2. С. 55-65. DOI: 10.14357/20718594210206. **(RSCI**, переводная версия **Scopus, WoS)**

15. Бутенко Ю. И. Модель учебно-научного текста для разметки корпуса научно-технических текстов // Экономика. Информатика. 2021. № 48 (1). С. 123–129. DOI:10.52575/2687-0932-2021-48-1-123-129. **(К-1)**

16. Бутенко Ю. И., Строганов Ю. В., Сапожков А. М. Метод извлечения русскоязычных многокомпонентных терминов в корпусе научно-технических текстов // Прикладная информатика. 2021. №6. С.21-27. DOI: 10.37791/2687-0649-2021-16-6-21-27. **(RSCI)**

17. Бутенко Ю. И., Сидняев Н. И., Болотова Е. Е. Использование падежной грамматики при информационном поиске в базе знаний о конструкции летательных аппаратов // Системы и средства информатики. 2021. №3. С.75-82. DOI: 10.14357/08696527210307. **(RSCI)**

18. Бутенко Ю. И., Сидняев Н. И., Строганов Ю. В., Киселева А. Д. Предикативная симптоматика и биометрия речевого поведения // Научно-техническая информация: Серия 2. Информационные процессы и системы. 2021. №2. С. 22-33. DOI: 10.36535/0548-0027-2021-02-3. **(RSCI**, переводная версия **WoS)**

19. Бутенко Ю. И. Модель текста стандарта при информационном поиске в коллекции документов нормативной базы // Вестник компьютерных и информационных технологий. 2020. Т. 17, № 11. С. 23-32. DOI: 10.14489/vkit. 2020.11. pp.023-032. **(К-2)**

20. Бутенко Ю. И. Метод разрешения лексической многозначности поискового запроса на основе онтологий // Прикладная информатика. 2020. Т. 15, №5. С. 103-110. DOI: 10.37791/2687-0649-2020-15-5-103-110. **(RSCI)**

21. Бутенко Ю. И., Сидняев Н. И., Болотова Е. Е. Теории формальных грамматик в методах распознавания неизвестных объектов // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2020. №8. С. 1-12. DOI: 10.36535/0548-0027-2020-08-1. **(RSCI**, переводная версия **WoS)**

22. Бутенко Ю. И., Сидняев Н. И., Болотова Е. Е. Логическая модель требований информационно-системной надежности для баз знаний интеллектуальных систем // Программная инженерия. 2020. №4. С. 195-204. DOI: 10.17587/prin.11.195-204. **(RSCI)**

23. Бутенко Ю. И., Сидняев Н. И., Болотова Е. Е. Язык логики предикатов в системах обработки информации в базах знаний // Физические основы приборостроения. 2020. Т.9, №2 (36). С. 37-47. DOI: 10.25210/jfop-2002-037047 **(RSCI)**

24. Бутенко Ю. И. Онтологический подход к формированию нормативного профиля при сертификации программного обеспечения // Онтология проектирования. 2020. Т. 10, №2(36). С.190-200. DOI: 10.18287/2223-9537-2020-10-2-190-200. **(RSCI)**

25. Бутенко Ю. И., Семенова Е. Л. Влияние лингвистических особенностей текстов стандартов на информационный поиск // Филологические науки.

Научные доклады высшей школы. 2019. №6. С. 29-35. DOI: 10.20339/PhS.6-19.029. **(WoS)**

26. Бутенко Ю. И., Сидняев Н. И., Болотова Е. Е. Экспертная система продукционного типа для сознания базы знаний о конструкциях летательных аппаратов // Авиакосмическое приборостроение. 2019. №6. С. 38-52. DOI: 10.25791/aviakosmos.06.2019.676. **(RSCI)**

27. Бутенко Ю. И., Шостак И. В. Семантическая модель языковых объектов для автоматизации процесса сертификации систем критического применения // Инженерный журнал: наука и инновации. 2013. № 12(24). С. 51. **(К-2)**

Публикации в сборниках трудов конференций, индексируемых в базах данных Web of Science и\или Scopus

28. Butenko Iu. I. Disambiguation in information retrieval in scientific and technical texts // AIP Conference Proceedings: International conference on modeling in engineering 2021, Moscow, Russia, October, 26-27. 2023. Vol. 2833(1), P. 020038. doi.org/10.1063/5.0151708 **(Scopus)**

29. Butenko Iu. I., Stroganov Yu. V., Sapozhknov A.V. System for Extracting Multicomponent Terms and their Translated Equivalents from Parallel Scientific and Technical Texts // AIP Conference Proceedings: International conference on modeling in engineering 2021, Moscow, Russia, October, 26-27. 2023. Vol. 2833(1), P. 030015. doi.org/10.1063/5.0151707 **(Scopus)**

30. Butenko Iu. I., Garazha V. V., Sidnyaev N. I. Multidimensional scaling in the analysis of linguistic information // AIP Conference Proceedings: International conference on modeling in engineering 2020, Moscow, Russia, April, 1-2. 2022. Vol. 2383, P.030011 doi.org/10.1063/5.0074583. **(Scopus, WoS)**

31. Butenko Iu. I., Sidnyaev N. I., Kiseleva A. D. Predicative analytics and speech biometrics// AIP Conference Proceedings: International conference on modeling in engineering 2020, Moscow, Russia, April, 1-2. 2022. Vol. 2383. P. 030012. doi.org/10.1063/5.0074672 **(Scopus, WoS)**

32. Butenko I. I., Sidnyaev N. I. Fuzzy information on obtaining grammars for representative images // AIP Conference proceedings: XLIV Academic space conference: dedicated to the memory of academician S.P. Korolev and other outstanding Russian scientists – Pioneers of space exploration, Moscow, Russia, January, 28-31, 2020. Vol. 2318. – Moscow, Russia: American Institute of Physics Inc., 2019. – P. 120009. DOI: /10.1063/5.0036147 **(Scopus, WoS)**

33. Butenko I. I., Sidnyaev N. I., Bolotova E. E. The method of aviation systems diagnostics according to the admissible level of non-failure operation probability // IOP Publishing Ltd International Conference Aviation Engineering and Transportation (AviaEnT 2020) IOP Conf. Series: Materials Science and Engineering 1061 (2021) 012037.- P. 1 – 7. DOI: 10.1088/1757-899X/1061/1/012037. **(Scopus)**

34. Butenko I. I., Sidnyaev N. I., Bolotova E. E. Statistical and Linguistic Decision-Making Techniques Based on Fuzzy Set Theory // Advances in intelligent systems, computer science and digital economics: International Symposium on Computer Science, Digital Economy and Intelligent Systems (CSDEIS). Moscow,

Russia, October, 04-06, 2019. 2020. Vol. 1127. P. 165-174. DOI: 10.1007/978-3-030-39216-1_16. **(Scopus, WoS)**

35. Butenko I. I. Ontology approach to normative profiles forming at critical software certification // AIP Conference proceedings: XLIII Academic space conference: dedicated to the memory of academician S.P. Korolev and other outstanding Russian scientists – Pioneers of space exploration, Moscow, Russia, January, 28, 2019. Vol. 2171. – Moscow, Russia: American Institute of Physics Inc., 2019. – P. 110002. – DOI 10.1063/1.5133236. **(Scopus, WoS)**

36. Butenko J. I., Sidnyaev, N. I., Garazha, V. V. Mathematical apparatus for engineering-linguistic models // AIP Conference Proceedings: International Scientific and Practical Conference on Modeling in Education. Moscow, Russia, June, 19-21, 2019. Vol. 2195, No. 1, p. 020033. DOI: 10.1063/1.5140133 **(Scopus, WoS)**

Публикации в других изданиях

37. Бутенко Ю. И, Марченко Д. Е. Анализ возможностей современных информационных технологий манипулировать отзывами в сфере образования // Alma mater (Вестник высшей школы). 2023. №7. С.66-71. DOI: 10.20339/AM.07-23.066

38. Бутенко Ю.И., Николаева Н.С. Модели структурных трансформаций одно- и двухкомпонентных терминов предметной области «Виды сварки» в английском и русском языках // Теоретическая и прикладная лингвистика. 2022. № 8 (2). С. 21-31. DOI: 10.22250/24107190_2022_8_2_21

39. Бутенко Ю.И., Николаева Н.С., Маргарян Т.Д. Структурные модели терминологических словосочетаний для разметки корпуса научно-технических текстов // Вестник НГУ: лингвистика и межкультурная коммуникация. 2021. №3. С. 46-56. DOI 10.25205/1818-7935-2021-19-3-45-56

40. Бутенко Ю. И., Авагян Н. А. Способы выражения модальности в параллельных текстах стандартов (на примере нормативной базы программной инженерии) // Вестник ВГУ: лингвистика и межкультурная коммуникация. 2021. №2. С.46-55.

41. Бутенко Ю. И. Роль мультидисциплинарных исследований в автоматической обработке научно-технических текстов // Ключевые тренды развития искусственного интеллекта: наука и технологии: Международная ИТ-конференция, Москва, 21 апреля 2023 года. М.: Издательство МГТУ им. Н.Э. Баумана, 2023. С. 63-67.

42. Бутенко Ю.И. Технологический процесс создания параллельного корпуса научно-технических текстов // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2022): доклады XXI Международной научно-технической конференции, Минск, 17 ноября 2022 г. – Минск: ОИПИ НАН Беларуси. 2022. С. 122-126.

43. Бутенко Ю.И., Киселева А.Д. Базовый шаблон многоязычной словарной статьи предметной онтологии нм основе параллельного корпуса научно-технических текстов // Наука, технологии и бизнес : II Межвузовская заочная конференция аспирантов, соискателей и молодых ученых (Москва, 27–28 апреля 2022 г.) : сборник материалов конференции / ФГБОУ ВО «МГТУ

им. Н.Э. Баумана (национальный исследовательский университет)». М.: Издательство МГТУ им. Н.Э. Баумана, 2022. С. 38-42.

44. Бутенко Ю.И., Синева Е.Е., Строганов Ю.В., Виноградов И.А. Разметка семантических ролей с целью извлечения информации из баз знаний в области авиакосмического приборостроения // Королёвские чтения 2022: XLVI Академические чтения по космонавтике, Москва, 25–28 января 2022 года. – М.: Издательство МГТУ им. Н. Э. Баумана. 2022. С. 453-456.

45. Бутенко Ю.И., Гаража В.В., Сидняев Н.И. Алгоритм шкалирования при сборе и анализе интеллектуальной информации // XX Всероссийская научная конференция «Нейрокомпьютеры и их применение». Тезисы докладов. М.: МГППУ. 2022. С.177-179.

46. Butenko Iu.I., Kiseleva A.D. Key features of parallel corpora // Наука, технологии и бизнес. Сборник материалов III Межвузовской конференции аспирантов, соискателей и молодых ученых = Conference Proceedings and Papers III Interacademic Conference for Graduate Students and Young Researchers. Москва. 2022. С. 42-46.

47. Butenko Iu.I., Sineva E.E. Information search in the expert system knowledge base on aircraft structures // Наука, технологии и бизнес. Сборник материалов III Межвузовской конференции аспирантов, соискателей и молодых ученых = Conference Proceedings and Papers III Interacademic Conference for Graduate Students and Young Researchers. Москва. 2022. С. 131-135.

48. Бутенко Ю.И., Авагян Н.А. Parallel corpus of scientific and technical texts as a translator's tool // Языки и культуры: перспективы развития в 21 веке: Альманах, Москва. - М.: Цифровичок. 2021. С.16-20.

49. Бутенко Ю.И., Синева Е.Е. Application of the scientific and technical text corpus in linguistics and linguodidactics // Языки и культуры: перспективы развития в 21 веке: Альманах, Москва. - М.: Цифровичок. 2021. С.132-136.

50. Бутенко Ю.И., Сапожков А.М. Система извлечения многокомпонентных терминов из параллельных научно-технических текстов // Язык. Общество. Образование: сборник научных трудов II Международной научно-практической конференции «Лингвистические и культурологические аспекты современного инженерного образования»; Томский политехнический университет. – Томск: Изд-во Томского политехнического университета. 2021. С. 22-24.

51. Бутенко Ю.И. Строганов Ю.В., Бабаджанян Р.В. Исследовательский прототип параллельного корпуса научно-технических текстов // 4-я Международная научно-практическая конференция «Лингвистика и лингводидактика в неязыковом вузе»: Сборник трудов. 2021. Т1. С.205-209.

52. Бутенко Ю.И., Шершнева Е.А. Разрешение многозначности поискового запроса в корпусе научно-технических текстов // 4-я Международная научно-практическая конференция «Лингвистика и лингводидактика в неязыковом вузе»: Сборник трудов. 2021. Т1. С. 209-212.

53. Бутенко Ю.И., Болотова Е.Е. Проектирование базы знаний для перевода узкоспециализированных текстов // Материалы Международного молодежного научного форума «ЛОМОНОСОВ-2020» [Электронный ресурс]. –

Электрон. текстовые дан. (1500 Мб.) – М.: МАКС Пресс, 2020. – Режим доступа: https://lomonosov-msu.ru/archive/Lomonosov_2020/index.htm, свободный.

54. Бутенко Ю.И., Сидняев Н.И., Болотова Е.Е. Экспертная система продукционного типа для создания базы знаний о робототехнических системах специального назначения // Актуальные проблемы защиты и безопасности: Труды XXIII Всероссийской научно-практической конференции РАРАН, 2020. С. 171-177.

55. Бутенко Ю.И., Кочеткова Е.Л. Анализ средств автоматизации переводческой деятельности // Молодежь и наука: актуальные проблемы фундаментальных и прикладных исследований: материалы III Всерос. нац. науч. конф. студентов, аспирантов и молодых ученых, Комсомольск-на-Амуре, 06-10 апреля 2020 г. : в 3 ч. / редкол. : Э. А. Дмитриев (отв. ред.) [и др.]. Комсомольск-на-Амуре: ФГБОУ ВО «КнАГУ», 2020. Ч. 3. С. 247-250.

56. Бутенко Ю.И., Сидняев Н.И., Болотова Е.Е. Уровни представления обработки знаний экспертных технических систем при проектных оценках // Международная научная конференция «Фундаментальные и прикладные задачи механики», посвященная 100-летию со дня рождения Академика Константина Сергеевича Колесникова (Москва , 10–12 декабря 2019 г.): Тезисы докладов. Инженерный журнал: наука и инновации. 2020. Вып. 2. С.219-222.

57. Бутенко Ю.И., Сидняев Н. И., Болотова Е.Е. Алгоритм формирования требований информационно-системной надежности для баз знаний интеллектуальных систем // Материалы XVIII Всероссийской научной конференции «Нейрокомпьютеры и их применение». Тезисы докладов. М: ФГБОУ ВО МГППУ, 2020. С. 430-432.

58. Бутенко Ю. И., Сидняев Н.И., Оплетина Н.В., Болотова Е.Е. Новые решения и прогнозы в инженерном образовании будущего // Международный форум «Цифровые технологии в инженерном образовании: новые тренды и опыт внедрения» (Москва, 28-29 ноября 2019г.): сборник трудов / Московский государственный технический университет имени Н. Э. Баумана (национальный исследовательский университет). Москва: МГТУ им. Н. Э. Баумана, 2020. С.526-528.

59. Бутенко Ю.И., Киселева А.Д., Казанцева Е.С. Влияние полисемии на результаты информационного поиска // Информационные технологии в науке, бизнесе и образовании: сб. тр. X Международной науч.-практ. конф. студентов, аспирантов и молодых ученых. М.: ФГБОУ ВО МГЛУ, 2018. С. 36-40.

60. Бутенко Ю. И., Шостак И.В. Исследование свойств языка стандартов как экземпляра класса языков для специальных целей в контексте автоматизации процедуры сертификации // Интеллектуальные системы и прикладная лингвистика: тез. докл. IV Всеукр. научн.-практ. конф. Харьков, 2015. С. 20–23.