Федеральное государственное автономное образовательное учреждение
высшего образования «Московский государственный технический университет
имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

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»

Диссертация на соискание ученой степени
доктора технических наук

Москва – 2025

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ..... 5

1. ПРИНЦИПЫ И МЕТОДЫ ОБРАБОТКИ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ В ПАРАЛЛЕЛЬНОМ КОРПУСЕ..... 15

1.1. Обзор и анализ современных параллельных корпусов текстов 15

1.2. Этапы разработки параллельного корпуса научно-технических текстов 21

1.3. Обзор и анализ корпусов в аспекте автоматизации разметки и выравнивания параллельных научно-технических текстов..... 31

1.4. Анализ применимости современных методов и средств обработки текстов для создания параллельного корпуса..... 51

1.4.1. Методы и средства обработки структурных особенностей научно-технических текстов..... 51

1.4.2. Методы и средства обработки специальной терминологии 58

1.4.3. Методы и средства выявления машинных русскоязычных текстов..... 66

1.5. Выводы по главе..... 70

2. МОДЕЛИ КОМПОЗИЦИОННОЙ СТРУКТУРЫ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ В ПАРАЛЛЕЛЬНОМ КОРПУСЕ..... 72

2.1. Модель текста научно-технической статьи для структурной разметки научно-технических текстов в параллельном корпусе 72

2.2. Модель учебно-научного текста для структурной разметки в корпусе научно-технических текстов..... 79

2.3. Модель текста стандарта как иерархически-структурированного текста ... 84

2.4. Особенности обработки композиционной структуры научно-технических текстов в параллельном корпусе 96

2.5. Выводы по главе..... 101

3. МОДЕЛИ И МЕТОДЫ РАЗЕТКИ И ВЫРАВНИВАНИЯ АНГЛО- И РУССКОЯЗЫЧНЫОЙ СПЕЦИАЛЬНОЙ ЛЕКСИКИ В ПАРАЛЛЕЛЬНОМ КОРПУСЕ..... 103

3.1. Структурные модели русско- и англоязычных многокомпонентных терминов..... 103

2.2. Структурные модели англо- и русскоязычных номенклатурных наименований 112
3.3. Методы разметки и выравнивания специальной лексики на основе структурных моделей англо- и русскоязычных терминов 116
3.4. Метод разметки англо- и русскоязычных номенклатурных наименований в научно-технических текстах 121
3.5. Метод выравнивания многокомпонентных терминов и номенклатурных наименований в параллельных научно-технических текстах 126
3.6. Выводы по главе..... 139
4. МЕТОДЫ РАЗМЕТКИ РУССКОЯЗЫЧНЫХ МАШИННО-СГЕНЕРИРОВАННЫХ И МАШИНННО-ПЕРЕВЕДЕННХ ТЕКСТОВ .. 141
4.1. Актуальное членение предложения как маркер машинных текстов..... 141
4.2. Метод выявления русскоязычных машинно-сгенерированных текстов на основе особенностей актуального членения предложения 146
4.3. Метод выявления переводческих трансформаций как маркеров ручного перевода научно-технических текстов..... 151
4.4. Метод выявления русскоязычных машинно-переведенных текстов на основе особенностей актуального членения предложения..... 154
4.5. Статистическая обработка научно-технических текстов в аспекте выявления машинных текстов и их фрагментов..... 157
4.6. Выводы по главе..... 172
5. СИСТЕМА УПРАВЛЕНИЯ КОРПУСНЫМИ ДАННЫМИ ПАРАЛЛЕЛЬНОГО КОРПУСА НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ174
5.1. Концепция системы управления корпусными данными параллельного корпуса 174
5.2. Информационная технология обработки научно-технических текстов в параллельном корпусе 183
5.3. Программные средства разметки и выравнивания научно-технических текстов в параллельном корпусе 186

5.4. Использование параллельного корпуса в исследованиях по лингвистике и информатике 205
5.5. Выводы по главе 212
6. ИСПОЛЬЗОВАНИЕ РАЗРАБОТАННЫХ МОДЕЛЕЙ И МЕТОДОВ ДЛЯ РЕШЕНИЯ ПРАКТИЧЕСКИХ ЗАДАЧ ПО ОБРАБОТКЕ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ 215
6.1. Метод обработки учебных пособий для создания лексического тренажера по дисциплине «Иностранный язык» 215
6.2. Методы выявления тенденций развития научных направлений (на материале анализа публикаций по газовому топливу) 226
6.3. Метод формирования нормативного профиля требований к объекту сертификации 247
6.4. Метод информационного поиска в базе знаний о конструкции летательных аппаратов на основе падежной грамматики 257
6.5. Выводы по главе 266
ЗАКЛЮЧЕНИЕ 268
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 270

ВВЕДЕНИЕ

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 заключается в том, что предложены новые подходы и методы к построению систем автоматической обработки научно-технических текстов на английском и русском языках, которые автоматизируют рутинный процесс обработки параллельных текстов и позволяют увеличить объемы филологически компетентных баз данных размеченных научно-технических текстов. Практическая ценность работы подтверждается внедрением результатов диссертационной работы в ряд прикладных промышленных систем текстовой аналитики, о чем имеются акты о внедрении результатов диссертационного исследования.

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), Международной конференции «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 [1-8] соискателю лично принадлежит общий подход к извлечению англо- и русскоязычных многокомпонентных терминов на основе синтаксических шаблонов, подкрепленных морфологической информацией о каждой словоформе. В работах [9-10] соискателем предложен подход к созданию учебных и специальных словарей на основе параллельного корпуса научно-технических текстов, в работах [11-13] соискателем предложен общий подход к установлению семантических ролей в научно-технических текстах, принципы семантико-синтаксического анализа научно-технических текстов. В работах [14-16] соискателю лично принадлежит принципиальная постановка задачи анализа композиционной структуры научно-технических текстов и проработка основных подходов к их анализу. В работах [17-27] соискателем проработаны общие принципы информационного поиска в

сложно-структурированных научно-технических текстах на английском и русском языках. В работах [28-36] соискателем предложены различные подходы к использованию параллельного корпуса в лингвистике и лингводидактике. В работах [37-45] 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.

1. ПРИНЦИПЫ И МЕТОДЫ ОБРАБОТКИ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ В ПАРАЛЛЕЛЬНОМ КОРПУСЕ

1.1. Обзор и анализ современных параллельных корпусов текстов

В конце XX – начале XIX века сформировалась необходимость в создании инструментальных средств хранения и обработки больших объемов лингвистических данных. Одним из наиболее эффективных средств решения указанной проблемы представляются корпуса текстов [46]. Под корпусом текстов принято понимать большой, представленный в машиночитаемом формате, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач [47]. Параллельные корпуса, как особый вид лингвистических информационных ресурсов, представлены двумя основными типами: параллельный корпус как множество текстов-оригиналов, написанных на каком-либо исходном языке, и текстов — переводов этих же исходных текстов на один или несколько других языков; и корпуса, объединяющие тексты из одной и той же тематической области, независимо написанные на двух или нескольких языках, которые также в научной литературе называют сопоставимыми корпусами [48]. В рамках текущего исследования автор проводит исследования в рамках первого типа корпусов. Параллельные корпуса широко используются для решения целого спектра теоретических и практических задач в различных сферах человеческой деятельности, примерами которых являются:

- инженерия знаний – для составления тезаурусов по узким предметным областям и построения онтологий, создания лингвистических баз данных и баз знаний [49-50],

- машинный перевод – при создании статистических систем машинного перевода как источник для создания обучающей выборки переводов, выполненных людьми-переводчиками; при создании систем машинного перевода типа трансфер для реализации языковых трансформаций, которые необходимо производить при переводе с одного языка на другой [51-52];

- лингводидактика – при обучении иностранному языку для получения справок и статистических данных о языковых и речевых единицах, для составления упражнений для изучения родного и иностранного языков, обучении как общему, так и научно-техническому переводу, изучению специальной лексики по разным предметным областям [53-54];

- лингвистика – при проведении исследований по контрастивной лингвистике, терминоведении и терминографии, грамматике, лексикологии и лексикографии, переводоведению, оценке качества перевода [26, 55-56] и др.

Обзор параллельных корпусов, в которых одним из языков представлен русский, и анализ степени автоматизации обработки параллельных текстов целесообразно начать с параллельных подкорпусов Национального корпуса русского языка общим объемом более 150 млн словоупотреблений [48]. В настоящее время на сайте Национального корпуса русского языка размещены двуязычные пары параллельных подкорпусов для следующих языков, причем эти пары включают как переводы иноязычных текстов на русский, так и русских текстов на другой язык: английский, армянский, башкирский, белорусский, болгарский, бурятский, испанский, итальянский, китайский, латышский, литовский, немецкий, польский, португальский, румынский, украинский, финский, французский, чешский, шведский, эстонский. Параллельные тексты в составе Национального корпуса русского языка выравниваются при помощи программы HunAlign с возможностью ручной проверки и исправления результатов выравнивания. Тексты на русском и других языках сопровождаются автоматической морфологической разметкой с неснятой омонимией, а также семантической и метатекстовой разметками. Жанровое разнообразие параллельного корпуса Национального корпуса русского языка представлено художественными, публицистическими, научными, религиозными и юридическими текстами [57-58].

Параллельный многоязычный корпус InterCorp Parasol входит в состав Чешского национального корпуса и используется для контрастных и трансляционных исследований. Он содержит тексты в нескольких языковых

версиях, которые выровнены друг с другом по предложениям. Общий объем русской части составляет 13 млн словоупотреблений. InterCorp представляет собой версионный корпус, т.е. полностью доступен в отдельных версиях, которые добавляются примерно раз в год. Параллельный корпус InterCorp состоит из двух частей: ядра и коллекций. Ядро корпуса InterCorp состоит в основном из художественных текстов с ручной корректурой выравнивания. Коллекции состоят из текстов, полученных на нескольких языках, обработанных и выровненных автоматически [59].

Польско-русский параллельный корпус Варшавского университета является сбалансированным, размеченным параллельным корпусом со снятой омонимией. База данных параллельных текстов корпуса содержит 50% польских оригиналов, 33% - русских и 15% - переводы с других языков. Жанровое разнообразие корпуса представлено в подавляющем большинстве художественными (90%), религиозными (4%), юридическими текстами и документальной литературой (5%), публицистическими текстами (1%). Выравнивание параллельных текстов осуществляется при помощи программы ABBY Aligner, морфологическая разметка польских текстов выполнена с помощью программного средства TAKIPI, а для морфологической разметки русскоязычных текстов использован программный продукт Pantera [60].

Русско-китайский параллельный корпус научных текстов гуманитарной области содержит тексты гуманитарной направленности общим объемом 5 млн словоупотреблений: 14 монографий в области политики, международных отношений, лингвистики, литературоведения и переводоведения. Выравнивание корпуса выполняется при помощи программного средства Paraconc, точность выравнивания которого приблизительно 60-70%, что влечет за собой необходимость ручной проверки правильности результатов выравнивания. [61-62].

В китайско-русском параллельном корпусе официально-деловых документов реализована дискурсивно-структурная разметка. Ее наличие позволит выравнивать тексты не только по синтаксическим единицам – абзацам

и предложениям, но и по дискурсивным единицам, как лингвистическим единицам, образованным лексико-синтаксическим способом. Разметка в данном корпусе проводится вручную с использованием специального программного обеспечения, которое позволяет проводить выравнивание и разметку параллельных текстов [63].

Полистилевой русско-китайский и китайско-русский параллельный корпус по замыслу создателей должен включать подкорпуса официально-деловых, художественных, новостных, экономических и военных текстов. В настоящее время реализован подкорпус текстов по военной тематике общим объемом 168 тыс. русских слов и 283 тыс. иероглифов. Добавлена морфологическая разметка, в основе которой использованы принципы что и в Национальном корпусе русского языка. Выравнивание осуществлено с помощью алгоритма длины G-Clen, точность автоматического выравнивания которого составила свыше 95% при обработке официальных текстов [62].

Казахско-русский параллельный корпус, ориентированный на криминальную тематику, основан на сайтах новостных ресурсов и их перводных версиях. Объем корпуса насчитывает более 50 тыс. слов на русском и казахском языках вместе. Морфологическая разметка русскоязычных текстов реализована на основе пакета Python pymorhy2, а для казахского языка с учетом особенностей агглютинирующего языка, к которым он и относится, используется тэггер регулярных выражений, основанный на классе Regexp Tagger пакета nltk Python. Выравнивание элементов параллельных текстов осуществляется на основе словарного метода выравнивания, эффективность которого по данным создателей корпуса равна 60% [64].

Параллельный корпус ООН состоит из переведенных вручную документов ООН, собранных с 1990 по 2014 гг., для шести официальных языков ООН - английского, арабского, испанского, китайского, русского, французского и испанского. Каждый документ содержит метатекстовую разметку, включающую уникальный символ документа ООН, уникальный идентификатор документа для конкретного языка, дата публикации, место обработки и ключевые слова. Тексты

корпуса выровнены по языковым парам, доступен полностью выровненный подкорпус документов для всех шести официальных языков ООН [65].

Значительное внимание исследователей приковано к созданию параллельных корпусов текстов для разных языковых пар, причем довольно часто такие корпуса исследователи собирают и обрабатывают вручную. Таким образом, например, создан Русско-английский параллельный корпус экономических текстов, который представляет собой коллекцию из 22 текстов, взятых с русско- и англоязычной версии сайта Министерства экономического развития Российской Федерации. В результате было получено 169 пар предложений на языке оригинала и языке перевода [66].

Создатели Русско-тувинского параллельного подкорпуса электронного корпуса тувинского языка отмечают не только сложности со сбором иллюстративного материала, так как переводные произведения находятся в фондах редких книг библиотек, а некоторые из них имеются только в одном экземпляре, но и с разметкой и выравниванием параллельных текстов, которые на сегодняшний день они осуществляют в программе Microsoft Excel вручную [67].

В научной литературе также встречаются публикации, посвященные изучению параллельных корпусов научно-технических текстов с описанием сфер применения параллельных корпусов [68], однако при поиске более подробного описания или самих ресурсов дополнительной информации не найдено. Наиболее вероятное объяснение состоит в том, что авторы для проведения собственных исследований вынуждены создавать корпуса вручную.

В таблице 1.1 приведен сравнительный обзор параллельных корпусов, в которых одним из языков является русский, а также проведен анализ степени автоматизации разметки и выравнивания текстов при создании рассмотренных параллельных корпусов.

Таблица 1.1. Сравнение уровня автоматизации обработки текстов в параллельных корпусах

	Название корпуса	**Объем**	**Язык, образующий пару с русским**	**Выравнивание**	**Разметка**	
					Вид	Способ реализации
1	Параллельный корпус Национального корпуса русского языка	150 млн	23 языка	HunAlign с ручной подкоррекцией	морфологическая	Mystem
					семантическая	ручная
					метатекстовая	ручная
2	Параллельный многоязычный корпус InterCorp	13 млн	чешский	ручное	морфологическая	автоматическая
3	Русско-английский параллельный корпус экономических текстов	22 текста	английский	ручное	-	-
4	Русско-тувинского параллельного подкорпуса электронного корпуса тувинского языка	-	тувинский	ручное	-	-
5	Параллельный корпус русского и китайского языков	5 млн	китайский	Paraconc с ручной подкоррекцией	-	-
6	Китайско-русском параллельном корпусе официально-деловых документов	-	китайский	ручное	структурно-дискурсная	ручная
7	Полистилевой русско-китайский и китайско-русский параллельный корпус	-	китайский	алгоритм длины G-Clen	морфологическая	Mystem
8	Параллельный корпус ООН	335 млн	6 языков	По предложениям	метатекстовая	ручная
9	Польско-русский параллельный корпус Варшавского университета	30 млн	польский	ABBY Aligner	морфосинтакси-ческая	TAKIPI (польский), Pantera (русский)
					библиографическая	ручная
10	Русско-казахский параллельный корпус по криминальной тематике	50 тыс. слов	казахский	словарный метод выравнивания предложений	морфологическая	Pymorhy 2 (русский), теггер регулярных выражений (казахский)

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, а наибольшее число параллельных корпусов создано в парах русский-китайский и русский-чешский. 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, образующих пару с русским, а также средствах автоматической обработки текстов для наполнения корпусов. Кроме того, основываясь на результатах проведенного анализа, не было выявлено ни одного реализованного репрезентативного параллельного корпуса научно-технических текстов, где бы одним из одним из языков выступал русский [69-70], что говорит об актуальности проводимого исследования по созданию параллельного корпуса англо- и русскоязычных научно-технических текстов.

1.2. Этапы разработки параллельного корпуса научно-технических текстов

При проектировании параллельного корпуса научно-технических текстов прежде всего должен быть решен ряд вопросов, касающихся наполнения и структуры корпуса [71]. На основе анализа подходов к созданию специальных корпусов текстов, описанных в [72], технологический процесс создания параллельного корпуса научно-технических текстов [73] представлен на Рис. 1.1.

Потенциальными пользователями могут быть терминологи, терминографы [74], переводчики [35], [31], программисты, инженеры по знаниям, преподаватели иностранного языка для специальных целей, студенты и аспиранты, изучающие иностранные языки для специальных целей [32], программисты при создании программных средств по обработке естественного языка [75-76].

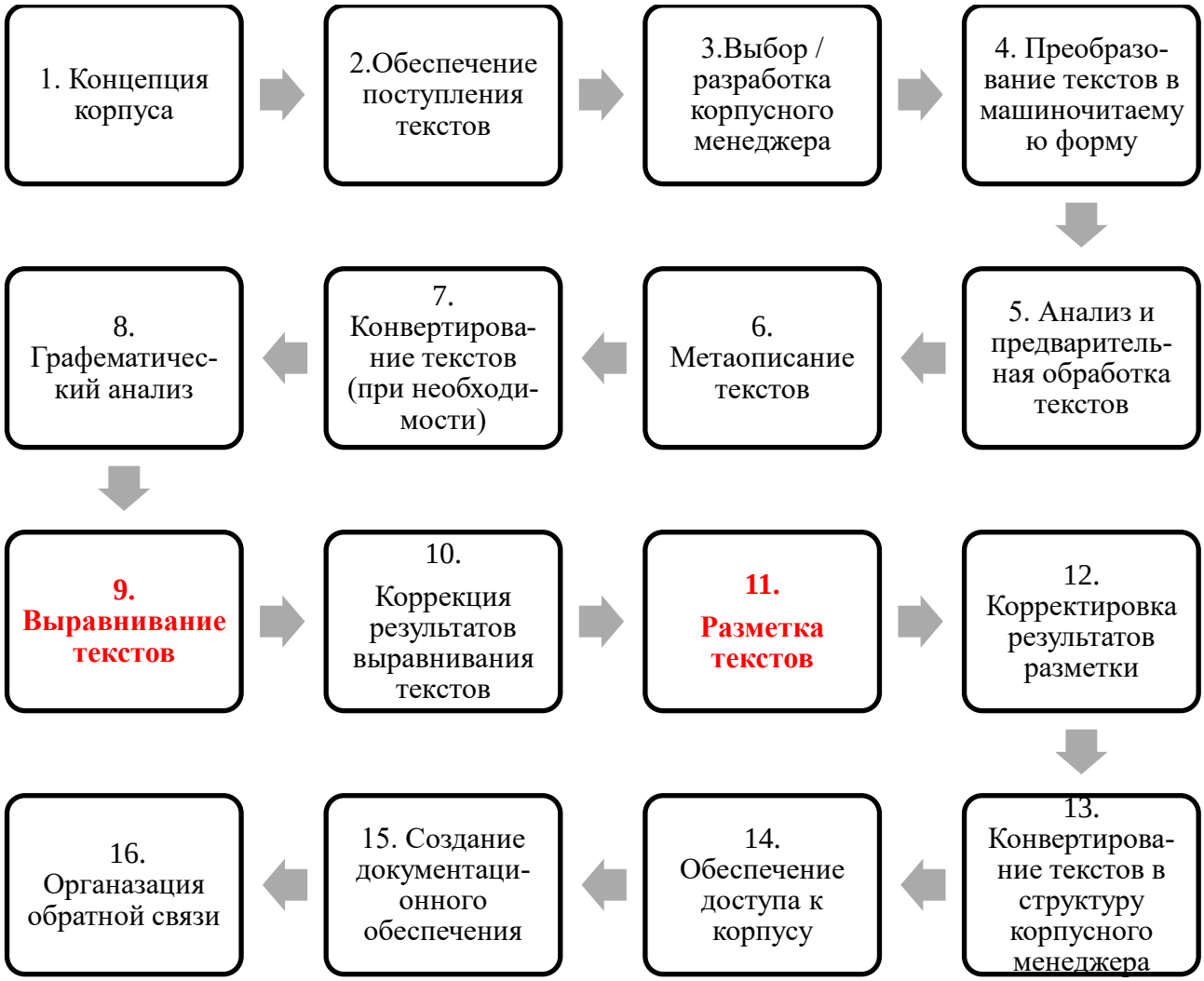


Рис. 1.1 – Этапы создания параллельного корпуса научно-технических текстов

2. Обеспечение поступления текстов. Одним из наиболее значимых аспектов, с которого начинается создание корпуса – это отбор языкового материала для наполнения корпуса, так как именно от отбора данных зависит репрезентативность корпуса, возможности его использования при решении теоретических и прикладных задач в рамках различных научных направлений, а также выбор или создание средств автоматической обработки научно-технических текстов.

Изучение параллельных текстов показало, что источниками наполнения параллельного корпуса англо-и русскоязычных научно-технических текстов могут быть: научно-технические статьи в изданиях, имеющих переводную

версию, переведенные с / на русский язык, научно-технические учебники и монографии, международные стандарты и др. В связи тем, что объем корпуса является важным и дискуссионным вопросом, необходимо оценить объемы доступных параллельных научно-технических текстов.

Так, по данным Электронной научной библиотеки (www.e-library.ru) по состоянию на лето 2022 года выпускается более 300 научно-технических журналов, которые имеют переводные версии, индексируемые в базах данных Scopus и Web of Science. В качестве примера рассмотрим объем параллельных текстов по предметной области «Информатика» за 2021 год. Переводные версии журналов по информатике имеются у 6 научно-технических журналов, а именно «Бизнес-информатика»/ Business Informatics (24 статьи), «Доклады Российской академии наук. Математика, информатика, процессы управления» / Doklady Mathematics (98 статей), «Моделирование и анализ информационных систем» / Automatic Control and Computer Sciences (31 статья), «Научно-техническая информация. Серия 1: Организация и методика информационной работы» (в переводную версию добавляются еще статьи из журнала «Искусственный интеллект и принятие решений») / Scientific and Technical Information Processing (54 статьи), «Научно-техническая информация. Серия 2: Информационные процессы и системы» / Automatic Documentation and Mathematical Linguistics (48 статей), «Проблемы передачи информации» / Problems of Information Transmission (26 статей), что в общем объеме составляет 281 статью. Архивы указанных журналов доступны за несколько лет, а также каждый квартал или чаще выходят новые выпуски. К предметной области информатики могут относиться научно-технические статьи из смежных предметных областей кибернетики, вычислительной техники, а также междисциплинарных исследований.

Информацию о ежегодно издаваемых переводных изданиях можно взять из Доклада Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации совместно с журналом «Книжная индустрия» [77]. Из доклада следует, что в 2020 году были издано 16 061 переводная брошюра и

монография, из них 9 917 книг – это переводы с английского языка. Основываясь на материалах доклада о том, что научно-техническая литература составляет порядка 20% от общего объема печатных изданий, то количество переводных версий научно-технических изданий составляет порядка 2 000 печатных изданий только за 2020 год.

Международные стандарты выпускаются международными организациями по стандартизации такими как International Standard Organization (Международная организация стандартизации), International Electrotechnical Commission (Международная электротехническая комиссия), International Telecommunication Union (Международный Союз Электросвязи). В российской системе классификации стандартов они имеет специальные приставки ИСО для стандартов, выпущенных Международной организацией стандартизации, МЭК – Международной электротехнической комиссией, ITU – Международным союзом электросвязи. Международные стандарты ИСО имеет переводные версии, начинающиеся с приставки ГОСТ Р ИСО, и в настоящее время насчитывают около 23 000 изданий. Объем стандарта не ограничен и может варьировать от нескольких страниц до нескольких сотен страниц. Полный перечень таких стандартов можно найти на Информационном портале по международной стандартизации Федерального агентства по техническому регулированию и метрологии [78].

Заранее оценить и предусмотреть репрезентативность создаваемого параллельного корпуса научно-технических текстов сложно, однако его можно будет сбалансировать в процессе создания [72], а проведенный выше анализ потенциальных источников научно-технических текстов для размещения в параллельном корпусе показывает возможность реализации такого подхода.

3. Выбор или разработка корпусного менеджера. На следующем этапе создания параллельного корпуса научно-технических текстов необходимо осуществить выбор корпусного менеджера из доступных в настоящее время или создать свой собственный. Корпусным менеджером принято называть «специализированную систему, включающую программные средства для поиска

данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме» [72]. Выбор корпусного менеджера зависит как от самих текстов, которые планируется размещать в тексте, так и соответствовать цели и задачам создания корпуса.

В настоящее время существует несколько корпусных менеджеров WordSmith Tools [79], MonoConc Pro [80] и AntConc [81], corpus.byu.edu [82], CQPweb [83], Sketch Engine [84], и Wmatrix [85]. Наибольший недостаток WordSmith Tools, MonoConc Pro и AntConc состоит в том, что они плохо работают с корпусами большого объема. Корпус-менеджеры corpus.byu.edu, CQPweb, Sketch Engine и Wmatrix обеспечивают возможность работы с большим объемом размеченных текстов за счет хранения корпуса на сервере с предварительной индексацией данных для быстрого поиска.

Однако использование корпус-менеджеров ограничено рядом факторов: во-первых, сложностью доступа к корпус-менеджерам, которые зачастую требуют ежемесячной абонентской платы за пользование ресурсом. Во-вторых, в связи с тем, что данные хранятся на внешнем сервере нет возможности обратиться к исходным данным непосредственно, чтобы соотнести их с результатами поиска в корпусе. В-третьих, некоторые из этих программ не доступны пользователям из России даже на платной основе. В-четвертых, ограничения доступа к базе данных корпуса ставит невозможным использование корпуса при разработке различных систем обработки естественного языка [72].

Снять указанные выше ограничения можно путем использования системы управления корпусными данными. Корпусный менеджер может быть частью системы управления корпусными данными, которые могут работать автономно или объединены для загрузки и предварительной обработки текстов, автоматической разметки и выравнивания текстов, поиска информации в корпусе. Кроме стандартных функций создаваемые системы могут реализовывать концептуальное аннотирование – приписывание корпусам текстов определенной предметной области релевантных концептов. К таким системам относится платформа концептуального аннотирования,

представленная в работе [86], а сущность концептуального аннотирования представлена в [87]. Такие комплексы на практике реализованы крайне редко, однако актуальность разработки систем управления корпусными данными подтверждается наличием ряда публикаций по этой теме [88-90].

4. Преобразование текстов в машиночитаемую форму. Этот этап используется в тех случаях, когда источники текстов представлены на бумажных носителях или цифровых форматах, с которыми корпусные менеджеры не работают, например, когда речь идет о создании корпуса древнерусских или старославянских текстов. Все тексты, которые планируется размещать в корпусе доступны в форматах doc, docx, pdf.

5. Анализ и предварительная обработка текстов. Научно-технические тексты зачастую содержат большое количество рисунков, диаграмм, математических формул, таблиц и т.д., которые также являются элементами текста, однако могут потребовать предварительной обработки перед размещением текста в корпусе.

6. Метаописание текстов. Одним из ключевых аспектов проектирования параллельного корпуса научно-технических текстов является также метаописание текстов – процесс приписывания тексту различных характеристик, описывающих обстоятельства его создания, автора, соотнесенность с определенным жанром и стилем изложения, предметной областью. Основное назначение метаописания – дать возможность пользователям параллельного корпуса настроить внешние параметры поиска текстов: например, осуществлять поиск по текстам, созданным авторами определенного года рождения, страны происхождения, гендерной принадлежности, предметной области.

7. Конвертирование текстов. Некоторые научно-технические тексты проходят через один или несколько этапов предварительной машинной обработки, в ходе которых осуществляется перекодировка, если требуется, а также удаление из текста переносов, «жестких концов строк» (тексты из MS-DOS), обеспечение единообразного написания тире и т. д.

8. Графематический анализ. Графематический анализ предполагает

разделение входного текста на элементы (слова, разделители и т. д.), удаление нетекстовых элементов, обработку специальных текстовых элементов (имен, написанных инициалами, иностранных лексем, записанных латиницей, названий рисунков, примечаний, страниц форзаца, зачеркиваний, титульных листов, списков литературы и т. д.). Как правило, эти операции выполняются в автоматическом режиме. Обычно на этом же этапе осуществляется сегментирование текста на его структурные составляющие. На этом же этапе необходимо решить проблему использования разных алфавитов в одном слове, что связано с наличием совпадающих букв, с одной стороны, и использования букв других алфавитов как частей термина, например, *E-слой*, где первая буква английская, а остальные русские. Также необходимо решить вопрос буквенно-числовых последовательностей, а также различать римские цифры и буквы латинского алфавита [72].

9. Выравнивание текстов. Выравнивание — установления соответствий между фрагментами научно-технического текста на языке оригинала и текста на языке перевода. Для решения этой задачи используются различные методы автоматического выравнивания текстов: по предложениям, клаузам (грамматическим конструкциям), словосочетаниям и словам [91]. Для выравнивания научно-технических текстов также часто применят метод транскрибирования, так как многие научные термины имеют один источник происхождения (греческий, латинский, английский, французский), что позволяет использовать термины «точками опоры» для дальнейшего выравнивания. Однако этот метод имеет недостатки, главным из которых можно назвать наличие различий в порядке между языком оригинала и перевода [92].

Методы выравнивания текстов также включают статистические методы, методы лингвистического выравнивания и методы выравнивания цепочек слов. Статистические методы выравнивания параллельных текстов направлены на то, чтобы найти наиболее вероятный вариант выравнивания для двух заданных параллельных текстов. Такие методы могут учитывать длину предложения на языке оригинала и языке перевода в символах, и подходят для выравнивания

параллельных текстов на сходных языках и буквального перевода. Метод пословного выравнивания представляется как вероятность того, что эквивалент слова в предложении на языке оригинала окажется в предложении на языке перевода. Для каждого выравнивания делаются упрощения, что каждое на языке оригинала имеет один переводной эквивалент на языке перевода. Однако такой метод не учитывает связи между словами и фразами, а также для некоторых языковых пар не может быть применим из-за необходимости использования большого числа переводческих трансформаций. Лингвистические методы могут основываться на грамматической информации, например, метод выравнивания по словам, относящимся к значимым частям речи, а служебные части речи не учитываются. Реализация такого метода требует наличия частеречной разметки параллельного корпуса. Однако сложность использования таких методов возникает при попытках выровнять слова внутри устойчивых выражений или переводческих трансформациях [51]. Семантические аспекты выравнивания параллельных текстов основаны на использовании трансфем, являющихся единицами когнитивного переноса, которые устанавливают функционально-семантические соответствия между структурами на языке оригинала и языке перевода.

Выравнивание параллельных текстов может быть реализовано не только по предложениям, но и по фразовым структурам – синтаксически значимым единицам в составе предложения, которые рассматриваются парадигматическом и синтагматическом аспектах. Выравнивание таких единиц реализуется в двух режимах: сопоставления на уровне трансфем и сопоставление на уровне концептов и отношений. Типология наиболее частотных трансфем в параллельных текстах приведена в работе [93]. Дистрибутивно-семантический подход к выравниванию параллельных текстов позволяет выравнивать тексты по словам [94]. Стоит отметить, что такой подход к обработке научно-технических текстов плохо применим в силу структурных особенностей терминологических единиц, которое чаще все представляют собой словосочетания, состоящие из 2 - 12 элементов.

10. Коррекция результатов выравнивания. Корректировка результатов выравнивания элементов текста представляет собой исправление ошибок и снятие неоднозначности. Коррекция может быть выполнена с использованием автоматических средств или вручную.

11. Разметка текстов. Тексты корпусов размечают для удобства использования, т.е. приписывают источникам наполнения корпуса и их компонентам специальные метки: внешние, экстралингвистические, содержащие информацию об авторе, названии, годе и месте издания, жанре, структурные и, собственно, внутрилингвистические, описывающие лексические и грамматические характеристики элементов текста: морфологическая и синтаксическая. В зависимости от целей создания корпуса в него включают дополнительные виды разметки: терминологическую, структурную, семантическую, форматную, ситуационную, редакторскую и пр. [95]. Разметка позволяет сделать корпус гораздо удобнее в использовании и является главной отличительной особенностью корпуса по сравнению с любыми другими коллекциями текстов. Таким образом, создание корпуса научно-технических текстов предполагает наличие лингвистической разметки, которая описывает сугубо лингвистические характеристики языковой выборки корпуса и представляет собой сложный процесс, требующий длительной и кропотливой работы над каждой лексической единицей, представленной в корпусе [96].

Данные лингвистической разметки можно добавлять к текстовым элементам разных уровней. Например, код класса слов или код частеречной принадлежности может быть привязан к каждому слову (токену) или группе токенов, которая может быть неразрывной или разрывной. Синтаксический код может быть закреплен за предложением или за синтаксическим отношением. С точки зрения технологии, разметка может быть автоматической, ручной или автоматической с ручной правкой [47].

12. Корректировка результатов разметки. Корректировка результатов разметки элементов текста представляет собой исправление ошибок и снятие неоднозначности (вручную или полуавтоматически).

13. Конвертирование текстов в структуру корпусного менеджера. Конвертирование размеченных текстов в структуру корпусного менеджера, обеспечивающего поиск и статистическую обработку.

14. Обеспечение доступа к корпусу. Корпус может быть доступен на локальном компьютере, в пределах дисплейного класса, может распространяться на машинных носителях и может быть доступен в режиме глобальной сети. Различным категориям пользователей могут предоставляться разные права и разные возможности.

15. Создание документного обеспечения. Создание документационного обеспечения, в котором описываются различные аспекты создания и использования корпуса, в частности приводятся сведения о разметке, позволяющие искать по метаданным, язык запросов корпус-менеджера и т. д.

16. Организация обратной связи. Любой пользователь сайта может обратиться в техподдержку для решения какого-либо вопроса, или же сообщить об ошибках, неточностях или необходимости расширить функционал разрабатываемого параллельного корпуса научно-технических текстов.

Таким образом, широкий спектр способов применения параллельных корпусов, значительное число публикаций по разным аспектам создания многоязычных лингвистических баз данных на основе параллельных корпусов, а также в значительной мере ручной характер обработки текстов при создании и наполнении параллельных корпусов, свидетельствуют об актуальности темы исследования – создании моделей, методов, алгоритмов и программных средств обработки научно-технических текстов, на их основе создания параллельного корпуса. При этом, некоторые этапы разработки параллельного корпуса либо выполняются один раз, например, концепция корпуса, организация обратной связи, либо используются для ограниченного числа текстов, например, преобразование текстов в машиночитаемую форму. Наиболее рутинными и время затратными этапами создания параллельного корпуса научно-технических текстов являются этапы выравнивания и разметки научно-технических текстов. Решением указанных проблем может стать автоматизация рутинных процедур

по разметке и выравниванию параллельных научно-технических текстов, первым этапом которой является анализ корпусов в аспекте автоматизации разметки и выравнивания научно-технических текстов как источников наполнения параллельного корпуса, а на его основе – выделение видов разметки, а также способов их автоматической реализации.

1.3. Обзор и анализ корпусов в аспекте автоматизации разметки и выравнивания параллельных научно-технических текстов

Для решения автоматизации обработки научно-технических текстов в параллельном корпусе следует решить проблему их автоматической разметки и выравнивания. С этой целью необходимо прежде всего проанализировать степень автоматизации ранее созданных как параллельных корпусов для разных языковых пар, так и одноязычных корпусов специальных текстов, которые как как правило, небольшие по размеру, подчиненные определенной исследовательской задаче и предназначенные для использования преимущественно в целях, соответствующих замыслу составителя [47].

Основные характеристики и средства автоматического выравнивания и разметки текстов в многоязычных параллельных корпусах представлены в Таблице 1.2., а двуязычных параллельных корпусах – в Таблице 1.3.

Проведенный обзор и анализ 80 параллельных корпусов для разных языковых пар показал, что тематика текстов в значительной степени определена доступностью электронных версий параллельных текстов: многоязычные параллельные корпуса чаще всего основаны на многоязычной документации международных организаций в сфере законодательства, финансов и международных отношений, а источниками для наполнения двуязычных параллельных корпусов чаще всего являются художественные и юридические тексты или тексты из сети интернет. Довольно широко представлено языковое разнообразие параллельных корпусов, широко распространены корпуса, где одним из языков выступают чешский и английский.

Таблица 1.2. Сравнение уровня автоматизации разметки текстов в многоязычных параллельных корпусах

	Название корпуса	**Объем**	**Языковые пары**	**Выравнивание и разметка**	**Тематика текстов**
1	ACCURAT balanced test corpus for under resourced languages [97]	4.6 тыс. предложений	7 языков	Выравнивание: по предложениям Разметка: лингвистич., фонетич., семантич.	корпус включает новости, разговорную речь, технические тексты, тексты из социальных медиа и другие типы текстов для недостаточно ресурсных языков.
2	Civitas Gentium [97]	31 статья	3 языка	Выравнивание: ручное и автоматическое Разметка:XML	содержит выборку документов, статей, судебных решений, комментариев и других материалов, связанных с этой тематикой
3	CRATER 2 Corpus [97]	4 млн	3 языка	Выравнивание: по предложениям Разметка: токенизация	тексты из телекоммуникационной сферы
4	CsEnVi Pairwise Parallel Corpora [97]	31 млн	3 языка	Выравнивание: по предложениям Разметка: токенизация	выступления TED и субтитры из корпуса CLUVI на вьетнамском, чешском и английском языках.
5	DGT-Acquis [97]	20 млн	23 языка	Выравнивание: по предложениям Разметка:XML-разметка	семейство многоязычных параллельных корпусов, извлеченных из Официального журнала Европейского Союза с 2004 по 2011 года.
6	DGT-TM-2016 [97]	373 млн	~30 языков	Выравнивание: по предложениям Разметка: XML-разметка	правовые акты Европейского союза
7	EAC Translation Memory [97]	320 тыс.	26 языков	Выравнивание: по предложениям, абзацам, блокам, предложениям. Разметка: TMX	коллекция предложений и их профессионально выполненных переводов по образованию и культуре. TMX (Translation Memory eXchange)
8	ECDC Translation Memory [97]	320 тыс.	~20 языков	Выравнивание: по предложениям Разметка: токенизация	тексты по общественного здравоохранения и эпидемиологии: инфекционные заболевания, эпиднадзор, меры профилактики и контроля и др.
9	EMEA Corpus [97]	31 млн	~20 языков	Выравнивание: по предложениям Разметка: морфол., синтаксич., семантич.	новостные статьи, исследования, отчеты, рецензии, академические работы и другие тексты, которые отражают разнообразие событий в странах EMEA (Европа, Ближний Восток и Африка)

10	EUR-Lexparallel corpus [97]	37 млн	24 языка	Выравнивание: по абзацам Разметка: -	тексты европейских директив, регламентов, международных договоров, конвенций, соглашений; нормативных актов и правил.
11	Europarl Parallel Corpus [97]	650 млн	21 язык	Выравнивание: алгоритм Гейла и Черча Разметка: -	тексты законопроектов, правовых документов и речей, связанных с процессом законотворчества и правоприменения в Европейском парламенте
12	EuroParl-UdS [97]	1.9 млн предложений	3 языка	Выравнивание: по предложениям Разметка: XML	тексты дебатов Европейского Парламента.
13	European Central Bank parallel corpus [97]	757 млн	19 языков	Выравнивание: по предложениям Разметка: токенизация	тексты, связанные с деятельностью Европейского Центрального Банка, его политиками и мероприятиями, а также общие экономические и финансовые новости и аналитические материалы.
14	European Parliament Interpretation Corpus (EPIC) [97]	177 тыс.	3 языка	Выравнивание: по предложениям Разметка: частеричная	тексты дебатов Европейского Парламента на итальянском, английском и испанском с переводами во всех возможных сочетаниях.
15	GLOSSOLOGIA [97]	-	4 языка	Выравнивание: Разметка:	образцы естественного языка с целью их анализа и исследования.
16	JRC-Acquis Multilin-gual Parallel Corpus [97]	1 млн	22 языка	Выравнивание: по предложениям Разметка: токенизация	сборник законодательных текстов ЕС, написанных в период с 1950-х годов по настоящее время.
17	MLCC Multilingual and Parallel Corpora [97]	10.2 млн	9 языков	Выравнивание: параллельное Разметка: XML	тематика текстов в корпусе включает политику, экономику, спорт, науку, культуру, искусство, здоровье, технологии и др.
18	MULCOLD - Multilingual Corpus of Legal Documents [97]	1.2 млн	4 языка	Выравнивание: по абзацам Разметка: лемматизация, частеричная	тексты юридического характера, такие как законы, судебные решения, договоры, правила и регуляции, содержащиеся в различных странах и на различных языках.
19	MULTEXT-East "1984" annotated corpus 4.0 [97]	1.06 млн	11 языков	Выравнивание: по предложениям Разметка: лемматизация, мофрол., синтаксич.	тексы, которые аннотированы с помощью методологии MULTEXT-East. Содержатся тексты, на которых основана аннотированная версия романа "1984" Джорджа Оруэлла.

20	MultiJur: Multilingual Parallel Corpus of Legal Texts [97]	1.2 млн	5 языков	Выравнивание: по абзацам Разметка: -	тексты из различных видов юридических документов, таких как законы, судебные решения, контракты, уставы и другие.
21	OpenSubtitles2011 [97]	87 млн	54 языка	Выравнивание: по предложениям и словам Разметка: токенизация	субтитры к фильмам и телевизионным шоу на различных языках
22	OPUS [98]	100 млн	90 языков	Выравнивание: по предложениям, Разметка: MaltParser v1.4.1	Тексты художественных произведений, публицистика, наука и техника, туризм, политика, религия и др.
23	PANACEA English-French and English-Greek parallel corpus	2 млн предложений	3 языка	Выравнивание: GIZA++ или ручное Разметка: XML	экологические и законодательные тексты на английском языке и их переводы на французский и греческий языки.
24	ParaCrawl Corpus version 1.0 [97]		11 языков	Выравнивание: симметричное Разметка: TMX	Корпус представляет собой коллекцию параллельных текстов новостных статей, блогов, форумов и других публикаций из Интернета, где каждый текст представлен на нескольких языках.
25	Parallel Bible Corpus [97]	-	1169 языков	Выравнивание: сопоставление отрывков текстов Разметка: синопсис	включает полные тексты или фрагменты, относящиеся к Ветхому и Новому Заветам, а также книгам и писаниям, связанным с Христианством.
26	Parallel corpus of KDE4 localization files [97]	60 млн	92 языка	Выравнивание: по предложениям Разметка: токенизация	тексты, относящиеся к локализации среды рабочего стола KDE4, что позволяют пользователям взаимодействовать с KDE4 на любом языке.
27	Parallel Global Voices [97]	800 тыс.	~50 языков	Выравнивание: по предложениям Не размечен	тексты о всевозможных политических событиях и международных взаимоотношений в различных странах мира.
28	Parallel sense-annotated corpus ELEXIS-WSD 1.1	345 тыс.	10 языков	Выравнивание: - Разметка: лемманизация, частеречная	набор текстов на нескольких языках, с каждым словом в тексте размеченным смысловым значением.
29	PELCRA multilingual parallel corpora [97]	143 млн	25 языков	Выравнивание: по предложениям Разметка: токенизация	корпус содержит параллельные тексты на различные темы, с целью поддерживать исследования в области обработки естественного языка на многоязычных данных
30	Polish-Bulgarian-Russian Parallel	55 текстов	3 языка	Выравнивание: параллельное	тексты различных тематик, таких как литература, наука, новости, экономика и другие.

	Corpus [97]			Разметка: XML и JSON	
31	SETimes [97]	43 млн	9 языков	Выравнивание: частичное по предложениям Разметка: XML	Тексты о политике, экономике, культуре, спорте, науке, образовании и других важных областях жизни стран Южной Европы.
32	SPC - Stockholm Parallel Corpora [97]	1.32 млн	4 языка	Выравнивание: по предложениям Разметка: токенизация	Тексты из различных источников: новостные статьи, литературные произведения, официальные документы и техническая документация.
33	Tatoeba [99]	7,5 млн	117 языков	Выравнивание: по предложениям Разметка: -	коллекция предложений и переводов, которая может быть использована любым человеком, разрабатывающим приложение для изучения языка.
34	The DPC – Dutch Parallel Corpus [97]	10.8 млн	3 языка	Выравнивание: по предложениям Разметка: токенизация	художественные, журналистские, поучительные и административные тексты на английском, голландском и французском языках.
35	UFAL Parallel Corpus of North Levantine 1.0 [97]	844,2 тыс. предложений; 6,2 млн слов	6 языков	Выравнивание: по предложениям Разметка: -	тексты различных жанров на североливанском диалекте
36	UMC 0.1: Czech-Russian-English Multilingual Corpus	1,8 млн	3 языка	Выравнивание: по предложениям Разметка: токенизация	новостные статьи и комментарии на чешском, русском и английском языках с сайта Project Syndicate с 1995 по 2008 год.

Таблица 1.3. Сравнение уровня автоматизации разметки текстов в двуязычных параллельных корпусах

	Название корпуса	**Объем**	**Языковые пары**	**Выравнивание и разметка**	**Тематика текстов**
1	Aformes	376,250 токенов	Английский, греческий	Выравнивание: - Разметка: токенизация	Содержит статьи из журнала творческого письма студентов факультета английского языка в Греции.
2	Amharic-English bilingual corpus [100]	500 млн токенов	Амхарский, английский	Выравнивание: Billingual Sentence Aligner Разметка: -	Юридические тексты и новостные статьи.
3	Catalan-Spanish Parallel Corpus [101]	100 млн токенов	Каталанский, испанский	Выравнивание: Billingual Sentence Aligner Разметка: -	Двуязычные статьи из "El Periódico de Catalunya" (1997–2007) о политике, экономике, культуре, здоровье и спорте.

4	COMPARA: Portuguese – English parallel translation corpus	-	Португальский, английский	Выравнивание: по предложениям Разметка: -	Содержит художественные тексты и научные, газетные и туристические статьи.
5	Croatian-English parallel corpus hrenWaC 2.0 [102]	55 млн лов 1,6 млн предл.	Хорватский, английский	Выравнивание: Bitextor, по предложениям Разметка: -	Хорватско-английские тексты, найденные в домене верхнего уровня .hr для Хорватии.
6	Czech and English abstracts of ÚFAL papers [103]	200тыс. слов 12000 предл.	Чешский, английский	Выравнивание: по предложениям Разметка: токенизация	Английские и чешские аннотации научных статей, из Института формальной и прикладной лингвистики Карлова университета в Праге и представленных в системе Biblio.
7	Czech-English Manual Word Alignment [104]	113 тыс. токенов, 2500 предл.	Чешский, английский	Выравнивание: ручное, по предложениям Разметка: токенизация	Корпус выровненных вручную чешско-английских параллельных предложений.
8	Czech-Slovak Parallel Corpus [105]	5,7 млн предлож.	Чешский, словацкий	Выравнивание: Hunalign Разметка: автоматическая морфологическая, токенизация	Чешско-словацкий параллельный корпус, состоит из нескольких свободно доступных корпусов (Acquis, Europarl, Official Journal of European Union и часть корпуса OPUS - EMEA, EUConst, KDE4 и PHP) и загруженного сайта Европейской комиссии.
9	CzEng 1.6 [106]	206 млн	Чешский, английский	Выравнивание: по предложениям, словам Разметка: морфологическая, семантическая	Содержит тексты на английском и чешском языке.
10	English-Czech Corpus from Wikipedia [107]	7,5 млн токенов	Английский, чешский	Выравнивание: по предложениям, Hunalign Разметка: частеречная	Тексты статей из Википедии на английском и чешском языках
11	English-Persian parallel Corpus [108]	3,5 млн	Английский, персидский	Выравнивание: алгоритм Гейла и Черча Разметка: токенизация	субтитры к известным фильмам, с доступными одним или несколькими переводами
12	English-Urdu Religious Parallel Corpus	14,371 предлож.	Английский, урду	Выравнивание: по предложениям Разметка: токенизация	Содержит религиозные тексты (Библию и Коран).
13	English-Vietnamese Parallel Corpus [109]	5 млн слов 500 тыс. предл.	Английский, вьетнамский	Выравнивание: по абзацам, предложениям и словам	Тексты разбиты на категории: экономика, искусство и спорт, здоровье, наука, общество и политика, технологии, художественные тексты.

				Разметка: частеречн. Stanford parser, синтаксич.	
14	EnTam: An English-Tamil Parallel Corpus (EnTam v2.0) [123]	169,871 предлож.	Английский, тамильский	Выравнивание: по предложениям, Hunalign Разметка: морфологическая	Корпус содержит новостные статьи и тексты, связанные с кино.
15	Estonian Open Parallel Corpus 2012. Estonian-English	2,5 млн токенов	Эстонский, английский	Выравнивание: - Разметка: токенизация	содержит библейские и юридические тексты, был собран в рамках национальной программы по технологии
16	Estonian-English parallel corpus [123]	307,000 предложений	Английский, эстонский	Выравнивание: по предложениям Разметка: морфологическая	Корпус содержит эстонские законы и их переводы на английский язык, а также законодательство ЕС, переведенное на эстонский язык.
17	European Parliament Proceedings Parallel Corpus 1996-2011 [123]	1,2 млн предложений	Английский, греческий	Выравнивание: по предложениям, по абзацам, алгоритм Гейла и Черча Разметка: -	Корпус содержит дебаты Европейского парламента с 1996 по 2011 год.
18	Finnish-English parallel corpus fienWaC 1.0 [110]	2,9 млн	Английский, финский	Выравнивание: по предложениям Bitextor Разметка: токенизация	финско-английские тексты из домена верхнего уровня .fi для Финляндии.
19	FREL [123]	701,401 tokens	Греческий-французский	Выравнивание: - Разметка: токенизация	корпус содержит художественные тексты, переведенные с французского на греческий.
20	Greek-Bulgarian Bul-TM parallel corpus [123]	10 млн слов	Греческий, болгарский	Выравнивание: по предложениям Разметка: синтаксическая, частеречная	социальные и политические тексты, фольклорные тексты.
21	HindEnCorp 0.5 [111]	132 тыс. предло-жений	Английский, хинди	Выравнивание: по предложениям, Hunalign Разметка: морфологическая теггером Morče	тексты содержат TED talks, новостные статьи, статьи из Википедии
22	INTERA Corpus the Greek-English part [123]	4 млн токенов	Английский, греческий	Выравнивание: по предложениям Разметка: частеречная	Корпус содержит тексты из области права, образования, окружающей среды, туризма и здравоохранения.
23	Interliingual Perspectives [123]	18 статей	Английский, греческий	Выравнивание: - Разметка: -	Содержит статьи, опубликованные с 2010 года и далее, посвященные взаимодействию греческого языка с другими языками посредством перевода.

24	Kacenka [123]	3,3 млн токенов	Английский, чешский	Выравнивание: по абзацам, по предложениям, hunalign Разметка: токенизация	Параллельные тексты художественных произведений мировой классики на английском и чешском языках.
25	LILA parallel corpus [123]	8 млн токенов	Литовский, латышский	Выравнивание: по предложениям Разметка: токенизация	содержит художественные и нехудожественные тексты с 1991 по 2012 год
26	Manually aligned CES Polish-English parallel corpus [123]	1,4 млн токенов	Польский, английский	Выравнивание: по предложениям Разметка: токенизация	содержит отчеты CES
27	PaGeS [123]	38 млн токенов	Немецкий, испанский	Выравнивание: по предложениям, Разметка: лемматизация, частеречная	В корпусе находятся тексты на немецком и испанском языках. Также 6% текстов переведены с третьего языка на немецкий и испанский.
28	Parallel corpus newsletters IFT FR-GR [123]	-	Греческий, французский	Выравнивание: - Разметка: -	Этот корпус содержит информационные бюллетени IFT.
29	Parallel English-Irish corpus of legal texts [123]	-	Английский, ирландский	Выравнивание: по предложениям Разметка: -	Содержит юридические тексты
30	ParCor – A Parallel Pronoun-Coreference Corpus [123]	-	Английский, немецкий	Выравнивание: по местоимениям Разметка: -	Содержит выступления TED и публикации в книгах в Европейском Союзе.
31	ParFin [123]	360,000 токенов	Русский, финский	Выравнивание: по предложениям Разметка: токенизация	Корпус содержит художественные тексты за период с 1990 по 2010 год.
32	ParIce [112]	3,6 млн предло-жений	Исландский, английский	Выравнивание: по предложениям Hunalign Разметка: частеречная	тексты с 11 разных источников, в основном из доступных параллельных корпусов (Opus, Tilde, ELRC), с веб-сайтов
33	ParRus [123]	5,9 млн токенов	Русский, финский	Выравнивание: по абзацам Разметка: токенизация	Содержит тексты из классической литературы и литературы 20-го века.
34	QTLP English Greek Corpus for the automotive domain	2,946 тыс. пар предлож.	Английский, греческий	Выравнивание: по предложениям Разметка: токенизация	Содержит параллельные тексты с многоязычных сайтов компаний, производящих автомобили, автомобильные аксессуары и автозапчасти.

35	QTLP English-Greek Corpus for the MEDICAL domain	62 тыс предло-жений	Английский, греческий	Выравнивание: по предложениям Разметка: -	тексты, полученные из Интернета: тезисы научных работ; и тексты сайтов государственных и частных организаций, которые связаны с медициной.
36	QTLP German-Greek Corpus for the MEDICAL domain	2,752 пар предложений	Греческий, немецкий	Выравнивание: по предложениям Разметка: -	корпус содержит медицинские тексты. Почти все тексты были получены с официального сайта Европейского союза
38	QTLP Portuguese-Greek Corpus for MEDICAL domain	62,608 предлож.	Португальский, греческий	Выравнивание: по предложениям Разметка: -	Содержит медицинские тексты. Почти все тексты были получены с официального сайта Европейского союза.
39	QTLP Portuguese-Greek Corpus for the AUTOMOTIVE domain [123]	59,297 sentence pairs	Португальс-кий, греческий	Выравнивание: по предложениям Разметка: -	Содержит параллельные тексты с многоязычных сайтов компаний, производящих автомобили, автомобильные аксессуары и автозапчасти.
40	Serbian-English parallel corpus for the domain of management [123]	600 тыс.	Сербский, английский	Выравнивание: по предложениям Bilingual Sentence Aligner Разметка: морфологич, терминологич.	тексты по менеджменту из журнала *Journal for Theory and Practice of Management* и его переводной версии на сербском языке за 2008–2012 года.
41	Serbian-English parallel corpus srenWaC 1.0 [113]	23,1 млн токенов	Сербский, английский	Выравнивание: - Разметка: токенизация	Содержит тексты, полученные с сербских доменов верхнего уровня .rs. Корпус был построен с помощью Spidextor.
42	Slovak-English Parallel Corpus [123]	556 млн токенов	Словацкий, английский	Выравнивание: Hunalign Разметка: лемматизация, морфологическая	содержит тексты из учебников по языкам
43	Slovene-English parallel corpus slenWaC 1.0 [123]	718,315 токенов	Словенский, английский	Выравнивание: по предложениям, Spidetextor Разметка: токенизация	корпус содержит тексты, полученные с доменов верхнего уровня Slovenia.si. Корпус был создан с помощью Spidextor,
44	SzegedParalell: angol-magyar párhuzamos korpusz [123]	-	Венгерский, английский	Выравнивание: - Разметка: -	содержит художественные тексты и тексты о Европейском союзе
45	Text Corpus – EMEL	43 тыс. токенов	Английский, французский	Выравнивание: по предложениям Разметка: токенизация	Корпус состоит из докладов конференций, подготовленных сотрудниками "Лаборатории перевода и обработки речи", занимающейся обработкой естественного языка.

46	The Corpus of Free Trade Agreement [123]	3 млн токенов	Английский, испанский	Выравнивание: Translation Corpus Aligner 2 Разметка: токенизация	Содержит тексты по Соглашению о свободной торговле.
47	The English-Nepali Parallel Corpus [114]	1,2 млн	Английский, непальский	Выравнивание: по предложениям Разметка: токенизация	содержит тексты на непольском языке
48	The English-Slovak Parallel corpus [123]	556 млн токенов	Английский, словацкий	Выравнивание: Hunalign Разметка: лемматизация, TreeTagger, автоматическая морфологическая	Содержит юридические тексты, парламентские дебаты (из Europarl corpus), статьи из официального журнала Европейского союза и тексты из корпуса
49	The English-Swedish Parallel Corpus [123]	3,5 млн токенов	Английский, шведский	Выравнивание: по абзацам Разметка: токенизация,	Корпус состоит из художественной и научно-популярной литературы. Все тексты опубликованы с 1980 по 2000 год.
50	The KOTUS Finnish-Swedish Parallel Corpus [123]	4,3 млн токенов	Финский, шведский	Выравнивание: по предложениям Разметка: токенизация	пресс-релизы, обзоры, отчеты, законы и нормативные акты, а также правительственные предложения за период с 1993 по 2004 год.
51	The NAACL 2003 English-Romanian corpus [123]	1,6 млн токенов	Английский, румынский	Выравнивание: по словам, Princeton WordNee Разметка: морфологич., синтаксич.	Содержит тексты, датированные 2003 годом
52	The Norwegian-Spanish Parallel Corpus [115]	6 млн	Норвежский, испанский	Выравнивание: по предложениям Translation Corpus Aligner 2 Разметка: морфологическая Oslo Bergen Tagger	Корпус современных норвежских письменных текстов, переведенных на испанский язык и опубликованных в период с 2000 по 2009 год. Каждый текст классифицирован по жанру, полу автора, полу и родному языку переводчика.
53	The TRIS corpus [123]	1,76 млн токенов	Немецкий, испанский	Выравнивание: по предложениям Разметка: токенизация	содержит тексты Европейской комиссии с 1997 по 2010 год.
54	Tourism English-Croatian Parallel Corpus 2.0 [116]	140 тыс токенов	Английский, хорватский	Выравнивание: по предложениям Разметка: токенизация	Тексты из 25 веб-сайтов из области туризма.

При разметке и выравнивания двуязычных корпусов текстов используются автоматические средства выравнивания и разметки текстов. Чаще всего реализована частеречная разметка. Большинство параллельных корпусов выровнены по предложениям с использованием следующих программных средств: Billingual Sentence Aligner, Bitextor, Hunalign и Translation Corpus Aligner 2, а средства для лемматизации и частеричной разметки зависят от типа обрабатываемого языка.

С целью поиска средств автоматической разметки англо- и русскоязычных научно-технических текстов также целесообразно провести обзор и анализ средств автоматической разметки специальных одноязычных корпусов академических текстов. В данную группу также отнесены сопоставимые корпуса, отличительной особенностью которых является сбор текстов на разных языках, объединенных одной темой, жанром, временными или другими аспектами, т.е. размещаемые в корпусе тексты не являются оригиналами и их переводными эквивалентами.

Основные характеристики указанных специальных корпусов, а также виды разметок и средства автоматической обработки текстов, представлены в Таблице 1.4.

Объем специальных корпусов имеет разброс от полумиллиона до триллиона словоупотреблений, а жанровое разнообразие представлено практически всеми видами академических текстов, при этом наибольшие число корпусов в качестве источников наполнения содержат научные статьи.

Стоит отметить, что по сравнению с параллельными корпусами специальные корпуса имеют большее разнообразие видов разметки, в том числе широкий инструментарий их автоматической обработки. Таким образом, при создании параллельного корпуса научно-технических текстов стоит проанализировать и отобрать виды разметок, которые отражают все особенности размещаемых в корпусе текстов.

Обобщенные результаты степени автоматизации обработки текстов в параллельных и специальных корпусах представлены в таблице 1.5.

Таблица 1.4. Анализ степени автоматизации разметки специальных корпусов текстов

	Название корпуса	Объем	Язык	Разметка	Тематика текстов
1	Academic Corpus [123]	3,5 млн	Английский	-	Корпус содержит журнальные статьи, главы из книг, рабочие тетради, лабораторные пособия и конспекты занятий по искусству, коммерции, праву, биологии.
2	Academic texts – humanities / social science [123]	15,5 млн / 10,8 млн	Шведский	-	Корпус содержит академические тексты по гуманитарным / социальным дисциплинам, опубликованные в период с 1997 по 2012 гг.
3	ACL Anthology Reference Corpus [117]	75 млн	Английский	частеречная, метаданные автора / текста	Корпус содержит научные статьи по вычислительной лингвистике, опубликованные в период с 1979 по 2015 год.
4	Annotated Corpus of Czech Case Law for Reference Recognition Tasks [123]	350 статей	Чешский	метаразметка	Корпус состоит из 350 вручную аннотированных решений чешских судов высшей инстанции.
5	Bononia Legal Corpus (BoLC) [123]	18 млн слов	Английский Итальянский	Лемматизация, частеречная	Корпус состоит из итальянских и английских текстов. Представлены две различные правовые системы.
6	Chambers-Le Baron Corpus of Research Articles [118]	1 млн	Французский	-	Научные работы по дисциплинам: культура, литература, лингвистика и изучение языков, социальная антропология, право, экономика
7	Corpus Juridisch Nederlands [123]	5,856 текстов 100 млн слов	Нидерландский	Лемматизация, частеречная	Данный корпус содержит юридические тексты с 1814 по 1989 год, составленные по годам.
8	Corpus of academic Lithuanian [119]	9 млн	Литовский	Лемматизация, частеречн., синтаксич.	Учебники, научные монографии, журнальные статьи, рефераты, предисловия, научные отчеты, диссертации.
9	Corpus of Academic Slovene KAS 2.0 [120]	1,5 трлн	Словацкий	Лемматизация, морфологич., терминологич. KAS-term tool, структурн.	Корпус содержит бакалаврские, магистерские и докторские диссертации по гуманитарным, социальным и естественным наукам, опубликованные в период с 2000 по 2018 год.
10	Corpus of Estonian law	11 млн	Эстонский	-	Этот корпус содержит эстонские законы (1,8

	texts [123]				млн.), а также европейское законодательство (9,6 млн.), переведенное на эстонский язык.
11	Corpus of Estonian scientific texts [121]	5 млн	Эстонский	-	Данный корпус содержит научные статьи и кандидатские диссертации.
12	Corpus of Romanian Academic Genres – ROGER (bilingual, student papers) [122]	3,3 млн	Английский, румынский	-	Корпус содержит учебные работы, написанные румынскими студентами на румынском и английском языках с 2018 по 2021 гг.
13	Corpus of Slovene linguistic scientific writing JezKor [123]	9.3 млн токенов	Словенский	Лемматизация, морфологич. синтаксич.	Является институциональным хранилищем научных публикаций, содержащих различные типы текстов, от кандидатских диссертаций до научных и профессиональных статей.
14	Corpus scientific texts from the Open Science Slovenia Portal OSS [123]	326 млн токенов	Словенский	Лемматизация, морфол., синтаксич., семантич, метаданные	Более 150 тысяч монографий, статей, дипломных, магистерских и докторских диссертаций, учебников, обзоров с 2000 по 2022 гг.
15	Czech and English abstracts of UFAL papers [123]	2 млн слов	Чешский, английский	Лемматизация, морфологич. синтаксич., метаданные	Содержит рефераты научных работ по формальной и прикладной лингвистке. Для каждой публикации авторы представляют как оригинал аннотации, так и ее перевод.
16	Czech Court Decisions Corpus (CzCDC 1.0)	460 млн	Чешский	-	Судебные решения трех высших судов Чехии, опубликованных в период с 1993 по 2018 год.
17	Czech Legal Text Treebank [123]	40,950 токенов.	Чешский	Синтаксич., семантич., ручная	два документа: Закон о бухгалтерском учете и Постановление о бухгалтерском учете для предприятий с двойной записью
18	Czech Sociological Review [123]	3 млн	Чешский	-	Избранные научные статьи и эссе, опубликованные в Czech Sociological Review с 1993 по 2016 год.
19	English Scientific Text Corpus [123]	35 млн токенов	Английский	Токенизация, морфол., струк-турная, метаразметка	содержит журнальные статьи по компьютерной лингвистике, информатика, биология, машиностроение и электротехника
20	FiRuLex, Russian-Finnish Comparable Corpus of Legal Texts	2740917 токенов	Русский Финский	Лемматизация	Двуязычный сопоставимый русско-финский корпус юридических текстов

21	GENIA corpus [124]	437 тыс	Английский	Частерич., синтакс., терминол., семантич. и кореференции; метаразметка	содержит 1 999 рефератов Medline, отобранных с помощью запроса PubMed по трем терминам "человек", "клетки крови" и "факторы транскрипции".
22	German Legal Corpus (GeLeCo) [123]	23030293 токена	Немецкий	Лематизация, частерич., метатекстовая	Корпус законов, административных постановлений и судебных решений, изданных в Германии на федеральном уровне.
23	German Legal Decision Corpora (GLDC) [123]	32748 предл.	Немецкий	Метразметка	Корпус состоит из решений суда федеральной земли Германии Бавария.
24	German Legal Judgement Corpus (GLJC) [123]	200 предл.	Немецкий	Метаразметка, частеречн., семантич.	юридический эксперт аннотировал компоненты заключения, определения и суждения немецкого юридического стиля письма Urteilsstil.
25	IGC-Laws-21.05 [123]	27 дел 132537 слов	Исландский	Метаразметка, частеречн., семантич.	IGC-Laws - это подкорпус The Icelandic Gigaword Corpus. Содержит исландские законы, пояснительные отчеты и замечания.
26	Inglise-eesti ja eesti-inglise paralleelkorpus [123]	12,8 млн токенов	Эстонский Английский	Морфологич., синтаксич.	Эстонско-английский параллельный корпус правовых текстов, который содержит 2 подкорпуса
27	Legal Documents from Norwegian Nynorsk Municipialities [123]	127 млн слов	Норвежский (букмол и нюнорск)	Метаразметка	содержит 50 000 юридических документов и протоколов заседаний, собранных с помощью веб-краулера Veidemann.
28	META-NORD Acquis Treebank [123]	12 364 слов	7 языков	Синтаксич., морфол.	В рамках совместной работы INESS и META-NORD созданы две параллельные базы данных: Sofie Parallel Treebank и Acquis Parallel Treebank.
29	Modern Greek Dialects: scientific papers [125]	113 тыс	Греческий	Не размечен	Корпус содержит научные тексты по лингвистике и диалектологии.
30	MuchMore Springer Bilingual Corpus [126]	1 млн токенов	Английский, немецкий	Синтаксич., морфол.	Сопоставимый корпус из англо-немецких научных медицинских рефератов, взятых из 41 медицинского журнала,
31	MULCOLD, Multilingual Parallel Corpus of Legal Texts	1056508 токенов	4 языка	Метатектовая, лемматизация	Многоязычный сопоставимый корпус юридических текстов. Создан 26.04.2016 в Хельсинки, Финляндия
32	Old Bailey Corpus [123]	24.4 млн слов	Английский	Социолингвистич.	тексты "Трудов Олд Бейли", которые в период с 1674 по 1913 г. являлись центральным уголовным

					судом Лондона.
33	OROSSIMO Corpus [127]	2,5 млн	Греческий	Размечены термины-кандидаты, частичная структурная	учебные тексты по общественным наукам, информатике, экономике, лингвистике, фотографии, праву, инженерии, истории, астрономии геологии, медицине и биологии.
34	Reading Academic Text corpus [128]	-	Английский	-	кандидатские диссертации по сельскому хозяйству, психологии, пищевой науке, технологиям, метеорологии и истории.
35	Scientext corpus [129]	20 млн	Английский, французский	Морфологич., синтаксич., тематич.	Корпус содержит научные тексты и аргументированные эссе по гуманитарным, и техническим наукам.
36	Spanish-English Research Article Corpus	5,7 млн	Анлийский, испанский	-	Корпус содержит журнальные статьи, опубликованные в период с 2000 по 2010 год.
37	The KIAP corpus [123]	3,9 млн	Английский, французский, норвежский	Частеречная	Сопоставимый корпус содержит научные статьи по экономике, лингвистике и медицине, опубликованные в период с 1992 по 2003 год.
38	The Language of Literature and the Language of Translation	48.300 тыс слов	Греческий	Морфологич., синтаксич., сематич.	Данный корпус содержит журнальные статьи по литературоведению и переводоведению на греческом языке.
39	The Royal Society Corpus [123]	32 млн токенов	Английский	Морфологич., синтаксич., семантич.	Корпус охватывает период с 1665 по 1996 год, основан на философских трудах и трудах Лондонского королевского общества.
40	UH's English E-thesis corpus [130]	200 млн	Английский	Частеричная, синтаксич.	Данные корпуса содержат магистерские и кандидатские диссертации, опубликованные в период с 1999 по 2016 год. UH = University of Helsinki'
	UH's Finnish E-thesis corpus	12,5 млн	Финнский	Частеричн., лемматизация	
	UH's Russian E-thesis corpus	1,1 млн	Русский	Не размечен	
	UH's Spanish E-thesis corpus	2,3 млн	Испанский	Не размечен	
	UH's Swedish E-thesis corpus	105 млн	Шведский	Не размечен	

Таблица 1.5. Анализ степени автоматизации обработки текстов в параллельных и специальных корпусах

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

Таким образом, на основе проведенного обзора и анализа корпусов, степени автоматизации разметки текстов, а также проводимых исследований по разметке [95] в таблице 1.6 выделены виды разметок и обоснована их значимость для параллельного корпуса научно-технических текстов.

Анафорическая разметка подразумевает фиксацию референтных связей, т.е. отнесенность включённых в речь имён, именных выражений или их эквивалентов к объектам действительности. Референция – это соответствие высказывания с действительностью, при этом с индивидуальными и каждый раз новыми объектами и ситуациями. Она определяет логичность текста и не может осуществляться без общих знаний в конкретной области. В научно-технических текстах чаще всего рассматривают референции терминов. При автоматической обработке научно-технических текстов учитывается только термин в его единственном варианте употребления (например, аббревиатура «КА»), снижается частотность употребления данного понятия, однако если учитывать еще и референции данного понятия (например, «космический аппарат»/ «КА»/ «аппарат»), то очевидно, что частотность употребления понятия повышается, за счет учета других вариантов его наименования.

Таблица 1.6. Анализ видов разметки для параллельного корпуса научно-технических текстов

	Вид разметки	**Суть разметки**	**Целесообразность реализации в параллельном корпус**
1	**анафорическая**	фиксирование референтных связей	для повышения точности информационного поиска, классификации и рубрикации текстов
2	грамматическая	приписывание частей речи	для поиска в корпусе по разным словоформам
3	дискурсная	обозначение коммуникатив-ных актов, оговорок, пауз, повторов, реплик, хезитаций.	отсутствие таких явлений в научно-технических текстах
4	жанровая	отнесение текста к стилю и жанру	размещаются определенные жанры научно-технических текстов
5	издательская	оригинал-макет	особенности форматирования текста могут быть использованы при реализации других видов разметок
6	**метатекстовая**	паспортизация источников	приписывание автора, года издания, тематическая классификация
7	**морфологическая**	приписывание морфологических характеристик слов	основа для поиска в корпусе по лингвистическим параметрам
8	синтаксическая	описывает синтаксические связи между лексическими единицами и различные синтаксические конструкции	сложность формальной структуры многокомпонентных терминов, которые необходимо обрабатывать как целое
9	**семантическая**	приписывание единицам текста один или несколько семантических и словообразовательных признаков	спецификация значения слов, разрешение омонимии и синонимии, категоризацию слов
10	служебная	отображает стиль и информацию об аудитории	отсутствие существенных различий в стиле и аудитории текстов
11	**структурная**	отражение логической структуры источника	необходимость обработки текста как целого путем его представления как упорядоченного набора взаимосвязанных элементов
12	**терминологическая**	фиксация терминов и особенности использования общеупотребительной лексики	сложность обработки терминов разной формальной структуры в научно-технических текстах

Антецедент, в роли которого обычно выступает именное словосочетание, берется в пронумерованные скобки, а рядом со словом-заместителем ставится особый знак, отсылающий к антецеденту с соответствующим номером.

Грамматическая и морфологическая разметки подразумевают приписывание лексическим единицам категорий рода, числа, падежа, вида, времени и т.д. для реализации поиска словоформ в корпусе по различным лингвистическим параметрам. В современных корпусах реализуется автоматически (см. п. 1.3).

Дискурсную разметку целесообразно реализовывать при обработке текстов художественного и разговорного стилей, где наличие оговорок, пауз, повторов, реплик, хезитаций может значительно повлиять на значение высказывания или модифицирует его. Так как научно-техническим текстам данные явления не свойственны, то реализация дискурсной разметки в параллельном корпусе научно-технических текстов нецелесообразна.

Жанровая разметка в разрабатываемом корпусе реализована при отборе текстов для наполнения параллельного корпуса: учебные пособия, научно-технические статьи и стандарты, и соответственно реализуется в метатекстовой разметке при выборе жанра загружаемого текста.

Издательская разметка отражает особенности форматирования текста, которые могут быть использованы при реализации других видов разметок. Особенности издательской разметки отражены в цифровых форматах docx и pdf, в которых доступны научно-технические тексты для наполнения параллельного корпуса.

Метатекстовая разметка подразумевает под собой приписывание к научно-техническому тексту внетекстовых характеристик: авторов, год, издательство, журнал, жанр и другие особенности текста или условия его создания. Реализуется вручную, но требует незначительных временных затрат на реализацию. Может быть реализована автоматически для текстов научно-технических статей при внесении авторов, их аффилиации и электронной почты, журнала, его номера, года издания и страниц.

Синтаксическая разметка описывает синтаксические связи между лексическими единицами и различные синтаксические конструкции. Синтаксическая разметка является результатом синтаксического анализа. Чаще всего в его основе лежит грамматика структур непосредственно составляющих. Графически синтагматические отношения между членами предложения изображаются, как известно, в виде дерева, а в тексте они представлены парами из открывающейся и закрывающейся квадратных скобок, которые обрамляют различные синтаксические конструкции – именные, глагольные и предложные словосочетания, придаточные предложения. Тексты, получившие синтаксическую разметку, известны как treebanks.

Семантическая разметка состоит в приписывании единицам текста один или несколько семантических категорий. Семантические тэги чаще всего обозначают семантические категории, к которым относится данное слово или словосочетание, и более узкие подкатегории, специфицирующие его значение. Семантическая разметка корпусов предусматривает спецификацию значения слов, разрешение омонимии и синонимии, категоризацию слов (разряды), выделение тематических классов, признаков каузативности, оценочных и деривационных характеристик и т.д.

Служебная разметка используется для приписывания текстам информации о стиле и целевой аудитории текстов, размещенных в корпусе. Реализация служебной разметки в разрабатываемом параллельном корпусе нецелесообразна в силу отсутствия значимых различий в стиле текстов и аудитории.

Структурная разметка эксплицирует логическую структуру источника и предназначена для выделения структурных элементов текста. Для проведения структурной разметки необходимо обосновать процедуру сегментации текста, то есть его членение на необходимые и достаточные сегменты, релевантные для его компьютерной обработки, позволяющие пользователю определить лексические, грамматические и другие свойства изучаемых лингвистических единиц [131]. Структурная разметка предполагает деление текста на главы, абзацы и

предложения. Для сложно структурированных текстов указывают главы и параграфы [132].

Терминологическая разметка позволяет фиксировать терминологические единицы разных предметных областей. Терминологическая разметка является одной из наиболее значимых видов разметок специальных текстов, а ее ручная реализация в корпусе представляет собой рутинную процедуру, требующую значительных временных и человеческих ресурсов на ее реализацию.

В настоящее время ни в одном из корпусов не реализована **разметка машинных текстов** – текстов, которые по своей сути являются не продуктом речевой деятельности человека, а сгенерированы на основе текста на другом языке – машинно-переведенные тексты; или же созданы средствами генеративного искусственного интеллекта – машинные тексты.

Таким образом, на основе проведенного анализа видов разметки при создании параллельного корпуса научно-технических текстов необходимо реализовать следующие виды разметки: анафорическую, метатекстовую, морфологическую, грамматическую, семантическую, структурную и терминологическую разметки, а также разметку машинных текстов или их фрагментов. При этом, как показал, проведенный анализ морфологическая и грамматическая разметки уже успешно реализованы в существующих параллельных и одноязычных специальных корпусах. Метатекстовая разметка реализуется вручную, но требует незначительных временных затрат на реализацию. В виду сложности выявления и большого разнообразия семантических особенностей семантическая разметка реализуется вручную, однако целесообразно создание программных средств для ее реализации в автоматизированном режиме. Наиболее ресурсозатратными видами разметки, и соответственно требующими разработки специальных моделей и методов автоматической обработки текстов являются структурная и терминологическая разметки, а также разметка машинных текстов.

Таким образом, с целью автоматизации процедуры разметки научно-

технических текстов при терминологической, структурной разметок и разметки машинных текстов необходимо проанализировать современные подходы, методы и программные средства в аспекте их применимости для обработки текстов в параллельном корпусе.

1.4. Анализ применимости современных методов и средств обработки текстов для создания параллельного корпуса

1.4.1. Методы и средства обработки структурных особенностей научно-технических текстов

Научно-технический текст – это текст, представленный в виде письменного документа, которому присущи следующие особенности: предварительное обдумывание высказывания, строгий отбор языковых средств, нормированная речь, логическая последовательность изложения, упорядоченная система связи между частями высказывания, стремление к точности, сжатости, однозначности при сохранении насыщенного содержания [133].

Научно-технический текст, относясь к научному стилю, обладает рядом особенностей, которые отличают его от текстов других стилей. Наиболее ярко эти особенности выражены в наличии структуры текста, когда за каждым элементом текста закреплено его место и определен характер передаваемой информации; соединение вербальных и невербальных элементов, а также могут содержать как первичные и вторичные тексты одновременно: например, аннотация к статье будет вторичным текстом, а сама статья – первичным.

В качестве источника наполнения корпуса выбраны учебники, научно-технические статьи и стандарты, при этому учебники относятся к научно-учебному жанру, а остальные – к собственно-научному. На рис. 1.2 приведена классификация текстов, а также жирным выделены особенности, свойственные рассматриваемым научно-техническим текстам.

Научно-технический текст представляет собой совокупность вербальных и невербальных элементов, при этом невербальные элементы могут передавать значительно больше информации нежели вербальные. В языке вербальные

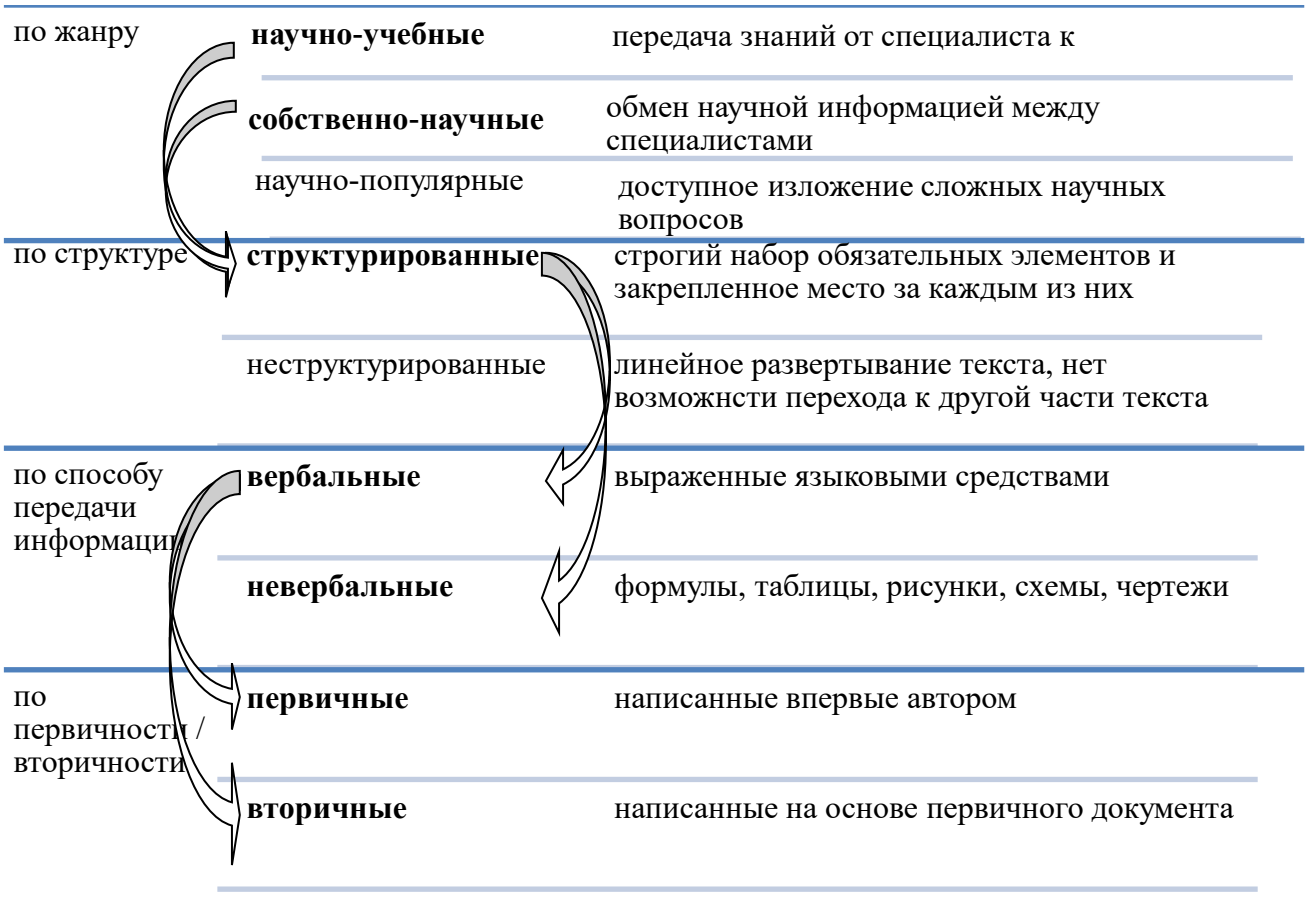


Рис. 1.2. Классификация научно-технических текстов

особенности отображаются на лексическом, грамматическом и семантическом уровнях, а также структурной организации научно-технического текста. Композиция научно-технических текстов проявляется в строении, соотношении и взаимном расположении частей текста [134].

Сложность и целесообразность сегментации научно-технических текстов связаны с особенностями их построения и синтаксической организации, что во многом обусловлено жанровой принадлежностью научно-технического текста. В лингвистике часто используют описание жанра при моделировании структуры текста. Параметризация текста, предполагающая оценку совокупности параметров, их специфики и речевой репрезентации, позволяет идентифицировать жанровую принадлежность документа, которая в общем виде определяет структуру текста, в первую очередь, его композицию и формуляр, образуемый реквизитами и речевыми единицами. Речевые единицы могут быть разной длины – от одного слова до фрагмента текста в несколько абзацев [131].

Помимо вербальных компонентов научно-технические тексты содержат большое число математических формул, графиков, таблиц, рисунков, схем, диаграмм, являющихся невербальными компонентами научно-технического текста. С одной стороны, такие элементы должны быть удалены из корпуса, так как зачастую не содержат лингвистической информации [72], а с другой стороны, могут быть более значимыми в плане информативности, чем вербальные компоненты текста [135], в связи с чем должна быть обеспечена возможность просмотра и разметки невербальной информации в параллельном корпусе. Накопление такой информации может стать основой для реализации поиска, например, по математическим или химическим формулам [136], поиск геологической структуры на карте одновременно с выполнением полнотекстового запроса в этом же документе [137]. Особенности разметки математических выражений, при котором пользователю предлагается формулировать поисковый запрос на поиск математической формулы в форме ключевых слов или словосочетаний, извлекаемых из анализируемого научно-технического текста представлен в [138]. В научно-технических текстах выделяются виды сущностей: естественнонаучные термины, символьные условные обозначения терминов (переменные), математические фрагменты (формулы). Для них определяются отношения: «термины – переменные» и «переменные – формулы». Первое отношение есть текстовое определение значения символа в некотором контексте с помощью терминов, второе отношение указывает на вхождение символа в формулу. Предполагается, что появление текстового определения переменной в окрестностях её символьного представления указывает на семантическую связь между ними [138]. Аналогичным образом прописываются другие невербальные элементы научно-технических текстов.

Исследованию композиционных особенностей как целых научно-технических текстов, так и их отдельных частей посвящено значительное количество работ. Так, в ходе анализа текста академической лекции выделены ее основные композиционные элементы: предтекстовая часть (заголовочный

комплекс), текстовая часть (вступление, основная часть, заключение) и послетекстовая часть (список литературы, выражение благодарности за внимание). Композиционно-смысловая структура лекции фиксирует движение от «старого» знания к «новому». В тексте лекции наблюдаются сильные и слабые позиции. Это означает, что ряд элементов в тексте лекции играет более важную роль по сравнению с остальными, при этом сильная позиция не обязательно жестко связана со структурой текста. Сильными позициями считаются заголовок лекции, названия ее подразделов, начало и окончание подразделов, вступительная и заключительная части лекции, выводы, а также семантические повторы ключевой информации, вопросно-ответная часть, внутритекстовые ссылки на литературу [139].

Композиция научных медицинских статей отличается достаточным разнообразием, что обусловлено отраслью медицины, спецификой описываемого метода (хирургический или терапевтический), социальной значимостью проблемы, авторским стилем, социокультурными особенностями [140]. Среди типичных разделов представляется возможным выделение следующих: «Аннотация», «Введение», «Материалы и методы», «Результаты», «Обсуждение». Для научной медицинской статьи характерно эксплицитное выражение средств рубрикации и композиционных приемов. Отдельные элементы научно-технический текстов также представляют интерес исследователей. В работе [141] автор представляет обобщенную дискурсивную структуру аннотации, в частности, сочетание характерных дискурсивных микросегментов и композиционно-речевых форм в каждом разделе аннотации.

Научно-популярные тексты имеют следующие структурные компоненты: предисловие, введение, основная часть, при этом основная часть является обязательным элементом, а остальные – факультативными, их наличие зависит от автора. Тексты делятся на главы, а главы на разделы и далее подразделы; они все пронумерованы, озаглавлены и внесены в оглавление [142].

В таблице 1.7 представлен анализ подходов и программных средств обработки научно-технических текстов или их элементов.

Таблица 1.7. Анализ подходов и программных средств обработки научных текстов и их элементов

	Исследуемый текст или его элемент	**Описание сути подхода, метода, алгоритма, программного средства**
1	Академическая лекция	Выделены структурные элементы текста академической лекции, семантические повторы ключевой информации
2	Научные статьи	Структурный анализ текста, выделение основных частей «Аннотация», «Введение», «Материалы и методы», «Результаты», «Обсуждение».
3	Научно-популярные тексты	Структурные компоненты: предисловие, введение, основная часть; основная часть является обязательным элементом, а остальные – факультативными, наличие зависит от автора.
4	Математические формулы	Организация поиска по формулам на основе их описаний в тексте, расположенном рядом
5	Химические формулы	Разработка методов поиска названий химических соединений в научно-технических текстах на основе методов машинного обучения.
6	Библиографические документы	Формирование коллектива решающих правил для классификации документов по их библиографическим документам: названию, аннотации и ключевым словам [143].
7	Аннотации	Деление на микросегменты и композиционно-речевые формы в каждом разделе аннотации
8	Учебные работы	Выделена структура каждого вида студенческих работ: реферат, курсовая работа, курсовой проект, выпускная квалификационная работа, охарактеризован каждый элемент в аспекте анализа на плагиат [144]

Несмотря на то, что научно-технические тексты представляют определенный интерес у исследователей, однако в научной литературе отсутствует комплексное представление научно-технического текста как иерархически-структурированной совокупности вербальных и невербальных элементов.

Таким образом, при реализации структурной разметки параллельного корпуса научно-технических текстов необходимо проанализировать каждый компонент научно-технического текста с целью определения оптимальных вариантов их разметки. Например, оценить необходимость включения текстов упражнений из научно-учебных текстов в основную часть корпуса, так как при составлении упражнения могут быть использованы термины из разных предметных областей или их комбинации, при этом исследуемая предметная

область в силу специфики текста и средств его автоматической обработки не будет отражена [145].

Также стоит отметить, что установление между структурными элементами научно-технических текстов, т.е. отношений между названием текста и его пунктами, а затем между названием пункта и абзаца позволяет для научно-технических текстов формировать минимальный контекст, необходимый для его восприятия, например, понять содержание требования стандарта: *В данном документе должны быть отмечены все расхождения между проектом и реализацией, обнаруженные в процессе тестирований* можно только обратившись к названию раздела, а затем стандарта.

Стандарт	*Атомные электростанции. Системы контроля и управления, важные для безопасности. Программное обеспечение компьютерных систем, выполняющих функции категории А*
Раздел	*Отчёт о тестированиях ПО*
Требования	*8.2.3.1.3.1 В отчёте о тестированиях ПО должны быть представлены результаты верификации, описанные в спецификации тестирований ПО и устанавливающие, работает или нет ПО в соответствии со спецификацией проекта ПО.* *8.2.3.1.3.2 В данном документе должны быть отмечены все расхождения между проектом и реализацией, обнаруженные в процессе тестирований.* *8.2.3.1.3.3 Отчёт о тестированиях ПО должен включать следующие пункты, как на уровне модуля, так и на уровне основного проекта…*

Таким образом, научно-технические тексты обладают ярко выраженной иерархической структурой, которая проявляется в том, что за каждым элементом установлено строго закрепленное место в научно-техническом тексте. Научно-технические тексты имеют четко выраженную композиционную структуру, которая просматривается в четком распределении структурных элементов научно-технического текста, его деление на разделы, подразделы, пункты, подпункты. В свою очередь, такое разделение порождает особенность: значения нижних элементов научно-технического текста можно установить однозначно

только с учетом значения верхних элементов. На рис. 1.3 представлена обобщенная структура научно-технических текстов, являющихся источниками наполнения параллельного корпуса.

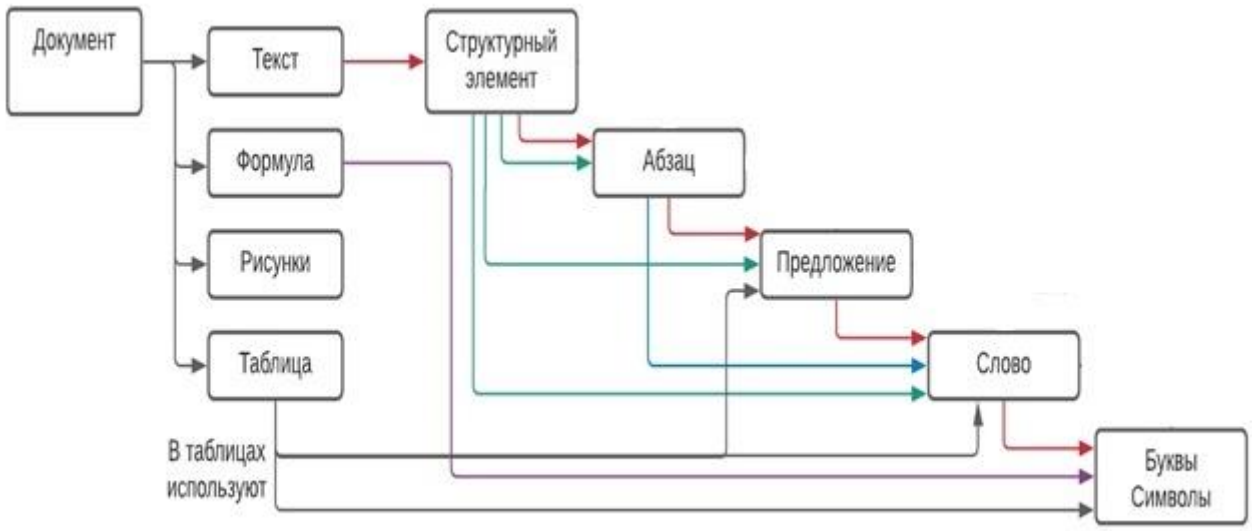


Рис. 1.3 – Обобщенная структура научно-технических текстов

Научно-технический текст представляет собой совокупность вербальных, т.е. текстовых элементов, и невербальных элементов. К последним принято относить формулы, рисунки, схемы, диаграммы и другие графические элементы, а также таблицы. Формулы, таблицы и рисунки могут содержать вербальные элементы, выраженные буквами (символами), словами или предложениями на естественном языке. Сам текст, как вербальный компонент, имеет строго выраженную иерархическую структуру, в которой можно выделить следующие уровни: структурный элемент, абзац, предложение, слово, буква (символ). Некоторые структурные элементы могут состоять только из предложений или слов. Например, аннотация не делится на абзацы, и чаще всего выражена некоторой совокупностью связанных по смыслу предложений, а структурный элемент ключевые слова состоит из только из слов.

Для автоматизации структурной разметки научно-технических текстов целесообразно провести инвентаризацию структурных элементов каждого жанра

научно-технического текста, которые являются источниками наполнения параллельного корпуса, а на их основе предложить метод автоматической разметки элементов научно-технических текстов.

1.4.2. Методы и средства обработки специальной терминологии

К настоящему времени разработано множество методов автоматического извлечения терминов, и число работ новых работ в последние годы только увеличивается. Общая схема для большинства методов извлечения терминов имеет следующий вид: сбор кандидатов: фильтрация слов и словосочетаний, извлекаемых из коллекции документов, по статистическим или лингвистическим принципам; подсчет признаков: перевод каждого кандидата в вектор признакового пространства; вывод на основе признаков: оценка вероятности быть термином для каждого кандидата на основе значений признаков. С целью снижения шума при сборе терминов кандидатов производится дополнительная фильтрация: по частоте, по содержанию в составе термина-кандидата стоп-слов, по длине или содержанию в термине-кандидате особых символов [146].

В настоящее время существует множество схем классификации методов извлечения терминологии из текстов [146-149]. Традиционно среди наиболее распространенных методов выявления терминов в текстах используют лингвистические и статистические методы, а также методы, основанные на машинном обучении и использовании различных информационных ресурсов [150]. В основе лингвистических методов лежит использование грамматики лексико-синтаксических шаблонов, представляющих собой структурные модели лингвистических конструкций [151]. Лингвистические методы – наиболее гибкие и точные, так как учитывают особенности языка, но сложны в реализации так как требуют активного участия лингвиста-эксперта, что в свою очередь порождает дилемму: лингвист не обладает навыками в области программирования, но от него требуется представление своих знаний в виде некоторых формализмов. С другой стороны, программисты в большинстве случаев не обладают достаточными знаниями в области лингвистики [152]. В

настоящее время лексико-грамматические шаблоны многокомпонентных терминов исследователи создают для каждого специализированного корпуса отдельно, например, для корпусной лингвистики, информационной безопасности [153], нанотехнологий [154] и др. В работе [152] предложен прототип программной системы, позволяющий лингвисту описывать синтаксические шаблоны терминов без знания формального языка.

Статистический подход заключается в нахождении n-грамм по заданным частотным характеристикам [155-157]. При статистическом подходе появление слова в тексте рассматривается как случайное событие. Данная трактовка позволяет оперировать языковыми единицами как в контексте математического аппарата теории вероятностей и статистики, то есть оценивать уровень случайности употребления определенных слов в контекстах определенного размера.

Методы на основе машинного обучения, опирающиеся на контекстуальные и внутренние свойства слов, показывают сравнительно высокую точность, однако их реализация требует огромных массивов обучающих данных, которые могут отсутствовать для определенной предметной области [158-160]. Кроме того, такие методы не обладают гибкостью, и в случае необходимости внесения изменений в систему, с высокой долей вероятности возникнет необходимость ее переобучения.

Извлечение терминологических единиц может быть также реализовано на основе информационных ресурсов таких как корпуса текстов, словари, энциклопедии, тематические коллекции текстов. В работе [161] предложена методология построения терминологического ядра предметной области на базе электронных энциклопедических источников данных. Особенностью предлагаемого похода является тщательный анализ структуры термина, распознавание ошибок на базе их лингвистической классификации, автоматическая генерация лексико-синтаксических шаблонов, представляющих многокомпонентные термины, и использование набора эвристических методов обработки «особых» терминов. В работе [162] для предложена многофакторная

модель для извлечения терминов, использующая три типа признаков: признаки, построенные на основе текстовой коллекции предметной области; признаки, полученные на основе информации глобальной поисковой машины; признаки, полученные на основе заданного тезауруса предметной области.

Стоит отметить, что в настоящее время для извлечения терминов в той или иной степени используют комбинацию двух-трех методов, что вызывает сложности при их классификации. Так, например, подход для выделения терминологических сочетаний, объединяющий лингвистический и статистический методы, заключается в предварительном описании моделей, по которым могут быть построены термины, для последующего нахождения их в корпусе. Внутри множеств однотипных синтаксических конструкций выполняется ранжирование в соответствии с той или иной статистической мерой [163].

Многие работы посвящены автоматическому извлечению терминов определенной длины: однословных [164, 158], двухсловных [150] и многословных терминологических единиц, извлечению специальных наименований, например, названий химических реакций в текстах научных публикаций [136]. В некоторых системах реализована возможность обработки текстов на нескольких естественных языках [165].

В зависимости от приложения, для которого требуются термины, также выделяют категории сценариев извлечения терминов [146], представленные на Рис. 1.4.

В рамках обработки научно-технических текстов для создания параллельного корпуса целесообразно использовать сценарии, рассматривающие каждое вхождение термина, сценарии, в которых число извлекаемых терминов определяется алгоритмом для каждой входной коллекции, а также сценарии извлечения терминов любой длины.

Рис. 1.4. Сценарии извлечения терминов

В таблице 1.8 проведен анализ применимости методов извлечения терминов для реализации терминологической разметки параллельного корпуса с учетом особенностей, отраженных в этих сценариях.

Анализ групп методов извлечения терминов, проведенный в таблице 1.8, показал ряд типовых ограничений. Так, статистические методы, будучи независимыми от языка, извлекают не все термины и их употребления в тексте; методы на основе синтаксических шаблонов – зависят от языка, кроме того, в них велика доля шума, так как модели терминов совпадают по форме с моделями общеупотребительных словосочетаний естественного языка. Кроме того, современные исследования в области терминоведения свидетельствуют об изменениях в способах образования терминов в русском языке [180]. Например, изучение только названий научных статей высокорейтингового журнала «Космические исследования» за 2018 и 2019 гг. позволило выявить следующие литеральные термины: *К метод, Е-слой, N тело, точка либрации L2, F-область*. Данный факт свидетельствует о жизнеспособности данного способа образования терминов и соответственно о необходимости поиска способов их автоматического извлечения [3].

Таблица 1.8. Обзор и анализ подходов и программных средств автоматического извлечения терминов для разметки параллельного корпуса

	Критерии / Методы	год	Зависимость от конкретного языка	Однокомпонентные термиты	Двукомпонентные термиты	3 и более компонена в термине	Номенклатурные нимменования	Каждое вхождение термина?	Все термины или определенные кол-во
	Методы на основе статистики вхождений								
1	Математическая модель русскоязычного текстового документа для решения задачи автоматического извлечения терминов из текста [166]	2017	+/-	+	+	-	-	-	-
2	Метод извлечения технических терминов с использованием меры странности [167]	2014	-	+	+	-	-	-	-
3	Метод извлечения технических терминов с использованием усовершенствованной меры странности [155]	2015	-	+	?	?	-	-	-
4	NFM-based approach to automatic term extraction [168]	2022	-	+	+	+	-	-	Не все
5	Estimating term domain relevance through term frequency, disjoint corpora frequency-tf-dcf [169]	2016	-	+	+	?	-	-	Не все
6	A survey on deep matrix factorizations [170]	2021	-	+	+	-	-	-	+
	Методы на основе синтаксических шаблонов								
7	Подход к извлечению многословных терминов из текстов на естественном языке с применением синтаксических шаблонов [152]	2021	+	+	+	?	-	+	+
8	Automatic term extraction in technical domain using part-of-speech and common-word features [151]	2018	+	+	+	-	-	+	+
9	Проблемы извлечения терминологического ядра предметной области из электронных энциклопедических словарей [161]	2018	+	+	+	+	-	-	Не все
10	Методы и программные средства извлечения терминологической информации из научно-технических текстов [171]	2013	+	+	+	-	-	-	?
	Методы на основе тематического моделирования								
11	Метод извлечения однословных терминов на основе статистического	2017	-	+	-	-	-	?	?

	распределения слов внутри контекста [164]								
12	Метод контрастного извлечения редких терминов из текстов на естественном языке [157]	2017	-	+	+	?	-	-	Не все
13	Topical word importance for fast keyphrase extraction [172]	2015	-	+	+	-	-		-
14	Salience rank: Efficient keyphrase extraction with topic modeling [173]	2017	-	+	+	-	-		-
15	Using latent semantic analysis for automated keyword extraction from large document corpora [174]	2017	-	+	+	-	-		-
16	BERT for arabic topic modeling: An experimental study on BERTopic technique [175]	2021	-	+	+	-	-		-
	Методы на основе информационных ресурсов								
17	Автоматическое извлечение двухсловных терминов по тематике «Нанотехнологии в медицине» на основе корпусных данных [150]	2013	+	-	+	-	-	-	-
18	In no uncertain terms: a dataset for monolingual and multilingual term extraction from comparable corpora [165]	2020	+	+	+	+	-	-	+
19	TermSuite: terminology extraction with term variant detection [176]	2016	-	+	+	-	-	-	-
20	Модели и методы автоматической обработки неструктурированной информации на основе базы знаний онтологического типа [162]	2014	-	+	+	?	-	-	не все
21	Автоматическое извлечение терминов из коллекции текстов предметной области с помощью Википедии [159]	2018	-	+	+	?	-	-	?
22	Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов [177]	2009	-	+	+	?	-	-	-
23	Метод извлечения терминов предметной области на базе электронных энциклопедических источников [161]	2015	-	+	+	-	-	-	Не все
	Методы на основе машинного обучения								
24	Метод автоматического извлечения терминов из научных статей на основе слабоконтролируемого обучения [160]	2021	-	+	+	+	-	?	-
25	Извлечение однословных терминов из текстовых коллекций на основе методов машинного обучения [158]	2013	-	+	-	-	-	?	Не все
26	Transforming Term Extraction: Transformer-Based Approaches to Multilingual Term Extraction Across Domains [178]	2021	-	+	+	+	-	?	-
27	Biomedical term extraction: overview and a new methodology [179]	2016	-	+	+	+	-	?	+

Методы на основе машинного обучения требуют наличия большого количества размеченных данных. Потенциально, такими ресурсами являются специальные корпуса, но их анализ в пункте 1.2 показал, что у них или маленький объем размеченных данных или существенные ограничения предметных областей текстов: чаще всего это тексты документов различных международных организаций.

Кроме того, при анализе результатов обработки терминологии программными средствами также выявлен следующий фактор. Обобщенная модель многокомпонентной терминологической единицы представлена на рис. 1.5. В структуре термина лингвисты выделяют ядерный элемент, левые и правые определения.

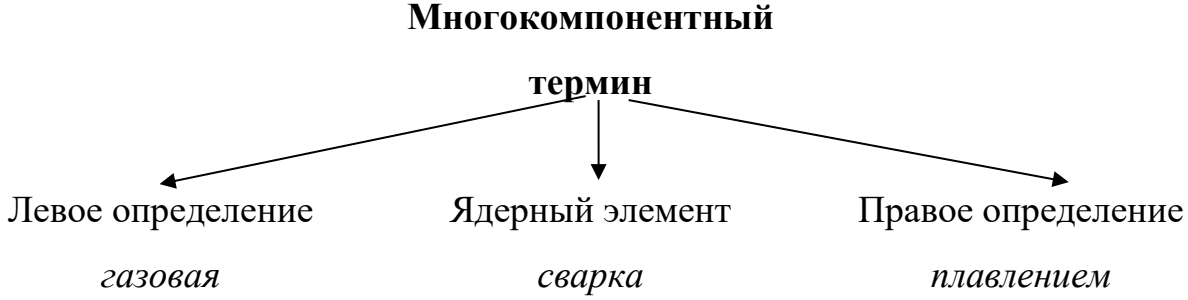


Рис. 1.5. Обобщенная модель многокомпонентного термина

Ядерный элемент – главный элемент в структуре многокомпонентного термина, который вступает в словоизменительную парадигму в предложении. Левое и правое определение уточняют значение ядерного элемента. При этом левое определение чаще всего выражено 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.

Анализ научных работ по терминоведению свидетельствует о частотности употребления такого рода конструкций. Обобщенные данные о количественном составе терминологических единиц представлены в таблице 1.9.

Таблица 1.9. Количественный состав терминологических единиц

Количество компонентов	**1**	**2**	**3**	**4**	**5**	**6**
%	35	28	25	7	3	2

При реализации вышеописанных в таблице 1.8 методов извлечения терминов на первом этапе проводится лемматизация, суть который состоит в приведении всех слов в начальную форму. Такой подход приводит к тому, что правые определения выделяются как отдельные термины как показано на рис. 1.6.

Левое определение	**Ядерный элемент**	**Правое определение**		
*прил, **ж.р., ед.ч., р.п***	*сущ, **ж.р., ед.ч., р.п.***	*сущ, ж.р., ед.ч.р.п*	*сущ,ср.р., ед.ч.р.п*	*сущ,мн.ч.р.п*
диалоговой	*системы*	*поддержки*	*принятия*	*решений*
*прил, **ж.р., ед.ч.им.п.***	*сущ, **ж.р., ед.ч.им.п**.*	*сущ, ж.р., ед.ч.им. п*	*сущ, ж.р., ед.ч.им.п.*	*сущ, ж.р., ед.ч.им.п.*
диалоговый	*система*	*поддержка*	*принятие*	*решение*

Рис. 1.6. Результаты лемматизации компонентов термина

Таким образом, для повышения эффективности методов извлечения многокомпонентных терминов за счет обработки терминов с правыми определениями необходимо разработать специальный подход, учитывающий их лингвистические особенности.

Анализ современных средств извлечения терминологии также показал, что они обрабатывают только один класс специальной лексики. В то же время развивающиеся терминологии компьютерных наук, авиации и космонавтики, нанотехнологий и других предметных областей используют новые способы терминообразования, таким образом расширяя структурные модели терминов [181]. К таким моделям словообразования можно отнести номенклатурные наименования, имеющие отличительную структуру от других пластов

специальной лексики [182]. Номенклатурой называют терминологическое обозначение частного специального понятия какой-либо предметной области. Номенклатурные наименования обозначают единичные вещи, предметное значение в них преобладает над понятийным – это знаки промышленных товаров, наименования механизмов и машин, видов животных, сортов растений, медикаментов и т.п., например, *пистолет «Браунинг»*, *самолет «Боинг 747»*, *насадки Kärcher 310* и т.п. Каждый номенклатурный список включает в себя совокупность однородных предметов, обладающих общими существенными признаками и различающихся второстепенными. По утверждению В.М. Лейчика, в настоящий момент человечество переживает «номенклатурный взрыв»: современное общество живет в мире номенклатуры – названий сортов, марок, артикулов [183].

Таким образом, для решения задачи разметки терминов из научно-технических текстов необходима разработка собственного решения, способного извлекать терминологические единицы разной формальной длины на русском и английском языках, при этом учитывать все термины и их вхождения в некоторой предметной области, при этом учитывать тот факт, что термин может состоять не только из слов на одном языке, но и включать буквенно-числовые последовательности в своем составе.

1.4.3. Методы и средства выявления машинных русскоязычных текстов

В настоящее время стремительную популярность набирают различные программные средства автоматической генерации текстов, результаты работы которых приближены к текстам, созданным человеком. В работе [184] поставлен эксперимент, в ходе выполнения которого случайным испытуемым предложили из списка выбирать машинно-сгенерированные фрагменты текстов. По его результатам лишь в 23% случаев испытуемые справлялись с этой задачей.

Использование технологий генерации текстов получило широкое распространение при решении целого ряда задач человеко-машинного

взаимодействия [185]. Алгоритмы генерации текстовых последовательностей успешно встраиваются в диалоговые системы, которые обеспечивают взаимодействие с пользователями в формате чата. Современные чат-боты способны предоставить ответ на любой корректно поставленный вопрос, а также обладают большим количеством дополнительных возможностей от перевода текста до написания программного кода и др. [186]. Такой широкий спектр возможностей в сфере генерации текстов стал пристальным объектом внимания недобросовестных авторов в разных сферах человеческой деятельности [187]. Однако особую обеспокоенность научного сообщества вызывают факты генерации научных и студенческих работ, в том числе и их использования для обхода проверки на антиплагиат [188]. Это обеспечивает актуальность решения задачи создания надежного средства обнаружения машинно-сгенерированных текстов в научных или учебных документах.

Современные исследования по выявлению машинно-сгенерированных текстов чаще всего основаны на использовании нейронных сетей, например: рекуррентных нейронных сетей для обнаружения человеческих и спам-бот аккаунтов в Twitter [189]; глубоких нейронных сетях на основе контекстной долговременной памяти [190]; модели BERT для обнаружения фальшивых новостей в Twitter [191] и бинарной классификации машинно-сгенерированных и написанных человеком текстов [192].

На сегодняшний день нельзя с полной уверенностью заподозрить сгенерированный текст при наличии только одного или нескольких вышеуказанных признаков. Но если их будет больше, можно с высокой долей вероятности определить «искусственный текст». Одно можно сказать с точностью – система «Антиплагиат» с новой функцией определения генерации текста (2023–2024 гг.) носит лишь рекомендательный характер в выявлении текста «с участием ИИ» и «без участия ИИ», последнее слово в процессе этой идентификации пока остается за проверяющим – за человеком [193].

Тельпов Р. Е., Ларцина С.В. исследуя типовые различия естественных и сгенерированных текстов, отмечают, что в сгенерированных текстах, слова,

включенные в заголовок, встречаются по тексту значительно чаще, чем в естественных текстах. Сгенерированный текст не раскрывает вопрос в полной мере, включает общую информацию, имеет более простую структуру и поверхностные примеры. Нейросеть очень часто повторяет сама себя, при этом она не способна оформить в тексте отсылку к своим словам. Никогда не будет оборотов - *как уже было сказано, повторимся, вернусь к*. Можно сделать вывод, что связанность текста присутствует, но связей между частями текста не так много. Также в особенностях ChatGPT указано, что используется авторегрессия. Авторегрессия – нейросеть генерирует текст, основываясь на предыдущих словах и последовательностях [194].

А. С. Никонов в своей статье предоставляет обширный анализ современных подходов, применяемых в задачах обработки естественного языка (NLP). Автор рассматривает архитектуры нейронных сетей, включая сверточные нейронные сети (CNN), рекуррентные нейронные сети (RNN) и трансформеры, а также их применение в задачах классификации текстов, машинного перевода [193].

Одним из перспективных направлений в выявлении машинно-сгенерированных текстов является использование методов глубокого обучения. Как отмечает А. С. Никонов, современные нейросетевые архитектуры, такие как сверточные нейронные сети (CNN), рекуррентные нейронные сети (RNN), а также модели типа Transformer, позволяют решать широкий спектр задач в области обработки естественного языка — от классификации до синтаксического анализа и генерации текста. Сверточные нейронные сети применяются в основном для извлечения локальных признаков текста, особенно в задачах классификации. Их достоинство заключается в способности фиксировать устойчивые шаблоны (например, фразы или устойчивые лексические конструкции), что делает такие модели полезными при построении фильтров для выявления повторяющихся фрагментов, характерных для текстов, созданных средствами искусственного интеллекта [193].

Рекуррентные нейронные сети (в частности, их модификации LSTM и GRU) демонстрируют высокую эффективность при работе с последовательностями. Эти модели сохраняют контекст и обучаются на длинных связных текстах, что позволяет им выявлять слабую связанность между частями текста — типичный признак, по которому можно отличить машинно-сгенерированный текст от написанного человеком. Особый интерес представляют трансформеры, как наиболее актуальное достижение в области NLP. Архитектура Transformer (в том числе её реализация в виде моделей GPT) позволяет учитывать глобальные зависимости в тексте, одновременно обрабатывая весь вход. Как показано в исследовании, такие модели обеспечивают высокую точность в задачах машинного перевода, резюмирования и извлечения отношений. Однако именно предсказуемость и шаблонность, присущая обученным трансформерам, может быть использована в задачах по выявлению ИИ-текстов [193].

Сравнительный анализ указанных моделей показывает, что каждое архитектурное решение обладает собственными достоинствами и недостатками в контексте задачи детекции искусственных текстов. CNN хорошо фиксируют локальные шаблоны, но не способны обрабатывать длинные зависимости. RNN обеспечивают лучшее понимание контекста, но подвержены проблемам затухающих градиентов. Трансформеры преодолевают эти ограничения, но из-за своей обучающей парадигмы склонны к генерации стилистически предсказуемых и тематически обобщенных текстов, что и делает их уязвимыми к алгоритмической идентификации [193].

Таким образом, нейронные сети используются для выявления ими же сгенерированных текстов. Кроме того, данные исследования ограничены как в языковом (чаще всего английский язык), так и жанровом (художественные и интернет-тексты) разнообразии текстов. Таким образом, использование нейронных сетей для выявления машинно-сгенерированных научных текстов ограничено в силу отсутствия больших объемов русскоязычных научных

текстов, заранее классифицированных на машинно-сгенерированные или написанные человеком.

1.5. Выводы по главе

Выявлено, что подавляющее большинство параллельных корпусов, в которых одним из языков выступает русский, разрабатываются авторами вручную или с ограниченным использованием средств автоматизации, а наиболее рутинными и времязатратными процедурами создания параллельного корпуса являются разметка и выравнивание научно-технических текстов.

Получены следующие результаты:

1. Анализ научно-технических текстов показал, что их лингвистические особенности проявляются прежде всего на композиционном, лексическом и стилистическом уровнях. Существующие средства автоматизации обработки научно-технических текстов слабо применимы для их автоматической обработки так как не учитывают указанные особенности научно-технических текстов. 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.

2. На лексическом уровне ключевой особенностью научно-технических текстов является использование специальной терминологии. Однако несмотря на значительные успехи в аспекте автоматического извлечения терминов из специальных текстов, существующие подходы и программные средства, с одной стороны, не рассчитаны на извлечение терминов из текстов на нескольких языках одновременно, а с другой стороны, не учитывают некоторые виды специальной лексики, такие как номенклатурные наименования, которые в своем составе содержат буквенно-числовые последовательности с использованием разных алфавитов.

3. Повсеместное распространение средств генеративного искусственного интеллекта привело к появлению нового вида текстов – машинных текстов,

которые обладают собственными лингвистическими особенностями. Обоснована необходимость добавления нового вида разметки – разметки машинно-переведенных и машинно-сгенерированных текстов, которые могут значительным образом повлиять на репрезентативность и сбалансированность параллельного корпуса.

4. Резюмируя вышесказанное, существующие средства машинной обработки научно-технических текстов малоприменимы для решения задачи автоматизации обработки научно-технических текстов при наполнении параллельного корпуса, так как не учитывают ряд лингвистических особенностей научно-технических текстов. С целью получения нужного уровня автоматизации обработки научно-технических текстов в аспекте создания параллельного корпуса необходимой является разработка специальных моделей представления структуры научно-технических текстов, моделей и методов извлечения многокомпонентных терминов и номенклатурных наименований из параллельных научно-технических текстов, методов разметки машинных текстов и / или их фрагментов, а также методов автоматического выравнивания параллельных научно-технических текстов.

Основные результаты к разделу опубликованы в работах [3, 29, 30-32, 35, 55-56, 69, 71, 74-75].

2. МОДЕЛИ КОМПОЗИЦИОННОЙ СТРУКТУРЫ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ В ПАРАЛЛЕЛЬНОМ КОРПУСЕ

2.1. Модель текста научно-технической статьи для структурной разметки научно-технических текстов в параллельном корпусе

Анализ целесообразно начать с определения присущих текстам нормативной базы характеристик. Стилистика, как раздел языкознания, изучает неодинаковые для разных условий языкового общения принципы выбора и способы организации языковых единиц в единое смысловое и композиционное целое (текст), а также определяет различиями в этих принципах и способах разновидности употребления языка (стили) и их систему. Каждый из лингвистических стилей имеет свои особенности, которые отличают его от других стилей, и в то же время облегчают автоматизированную обработку текстов [195].

В качестве источников наполнения параллельного корпуса научно-технических текстов отобраны научно-технические статьи, учебники и учебные пособия, нормативная документация по направлениям подготовки МГТУ им. Н.Э. Баумана. Обязательным условием для каждого текста, размещаемого в корпусе, является наличие переводной версии с русского на английский или английского на русский язык. Научно-технические статьи, учебники и учебные пособия относятся к научному стилю. Стоит отметить, что стилевая принадлежность текстов стандартов однозначно не определена, так как они совмещают черты официально-делового и научного стилей, что может потребовать их более детального изучения и разработки специальных методов обработки.

Научно-технический текст является структурированной и организованной по определенным правилам лингвистической единицей, которая несет когнитивную, информационную, социальную и психологическую нагрузку коммуникации. Неотъемлемым свойством научно-технических текстов является структурность, которая выражает отношения, между частями текста. Логико-

композиционная структура научно-технических текстов отражает строгую последовательность расположения смысловых частей в тексте, то есть предполагает расположение элементов текста по определенной схеме. Композиционная структура научно-технического текста отражает строение, соотношение и взаимное расположение его частей, членение на смысловые элементы, степень и характер выраженности этих элементов, порядок их следования и взаимосвязь между ними. Независимо от отрасли науки любой научно-технический текст строится по схеме: введение – основная часть-заключение, где каждый элемент выполняет собственную функцию в структуре целого текста. Во введении излагается тема, постановка проблемы, конкретные шаги по ее рассмотрению. Основная часть содержит развитие концепции, намеченное во вступлении: раскрывается тема исследования, решаются поставленные проблемы, сообщаются основные сведения и результаты. В заключении подводятся итоги, формируются выводы, прогнозируются возможные пути применения полученных результатов [196].

Все элементы научно-технического текста можно отнести к макроуровню, который включает главы, разделы и подразделы, т. е. те части письменного произведения, которые визуально отделяются (выделяются) при помощи заголовков и подзаголовков; мезоуровню, к которому относятся абзацы, перечни, таблицы, графики и цитаты, где каждый из перечисленных элементов считается самостоятельной единицей мышления; и микроуровню, включающему все грамматические конструкции в широком смысле, а также орфографию [196].

Композиционная структура научно-технических текстов подразумевает стандартизацию, однако несмотря на это требование, в композиционной матрице научно-технических текстов выделяются обязательные и факультативные элементы. При этом не существует единой схемы научно-технического текста: текстам различных жанров присуща своя специфическая структура. В связи с вышесказанным возникает необходимость анализа текстов каждого жанра, которые предполагается размещать в параллельном корпусе. Таким образом, необходимо установить общую структуру рассматриваемых научно-

технических текстов как совокупности элементов.

Структура совокупности может быть выявлена в результате проведения тщательной классификации исследуемых единиц, то есть их иерархического распределения по определенному признаку на основные разделы, которые далее распадаются на подразделы, пункты, подпункты, требования, которые в свою очередь разбиваются на отдельные более мелкие базисные единицы - предложения. Все это вместе образует номенклатуру - полный подробный перечень отдельных элементов изучаемой совокупности [23].

Любую систему можно представить как некоторую совокупность взаимосвязанных элементов. Каждая из таких систем S_j является отделенной системой (научный / официально-деловой стиль) и может быть представлена как некоторая часть (подсистема) более общей системы S (суперсистема – русский язык) $S_j \in S$.

Взаимосвязь между системами S_j и S построено по иерархическому принципу, который предусматривает подчиненность подсистемы S_j суперсистеме S, в плане своего структурного расположения, так и в плане функционально-коммуникативной направленности составных частей. Отсюда вытекает, что любую систему (совокупность) S можно разделить на подсистемы разных рангов $S_1 - S_2 - S_3$ определенный естественный язык – стиль – жанр), проводя процесс членения по определенным признакам до получения составных элементов. При этом каждая операция членения системы порождает отдельные подсистемы, что обеспечивает построение некоторого дерева метасистем S, на котором выделены отдельные подсистемы ($S_1; S_2; S_3$), которые относятся к разным уровням ($S_{12}; S_{121}; S_{1211}$). Членение системы может быть произведено рядом способов, про этом генерируется разное количество частей (подсистем, базисных элементов) [197].

Научно-техническая статья – это первичный письменный жанр научного дискурса, задачей которого является постановка и решение одной научной проблемы, имеет средний объем, конвенциальную структуру, системы ссылок и

выходные данные [198]. Научно-техническим статьям присущи все стилевые особенности научного стиля: точность, логичность изложения материала, эмоциональная нейтральность, наличие специальной терминологии. Результаты анализа текстов статей представлены в Таблице 2.1.

Таблица 2.1 – Результаты анализа текстов научно-технических статей

Сфера функционирования и типовая ситуация общения	Наука и техника. Обмен информацией о решении научно-технических проблем и задач
Участники речи	Профессиональное сообщество ученых и инженеров. Специалисты, владеющие обширными знаниями о специализированной предметной области статьи
Функция речи	Профессиональная коммуникация специалистов разных специализированных предметных областей.
Тип содержания и предмет речи (тема)	Содержание – конкретное. Тема – результаты определенного исследования или обзор текущего состояния предмета исследования.
Коммуникативная цель	Обмен ясно, кратко и достоверно изложенной информацией о результатах исследования в некоторой предметной области
Стилеобразующие признаки	Ясность, точность, логичность, объективность и точность. Обезличенность информации. Шаблонность, четкое закрепление места за элементами.
Форма бытия	Письменная, в виде текстов установленной формы в научно-технических или специализированных периодических изданиях определенных предметных отраслей.

К ключевым элементам структуры научно-технической статьи с точки зрения их функциональных и лексико-грамматических особенностей относят [199, 200]:

1. Код УДК.

2. Название. Название отражает содержание статьи, обычно состоит не более одиннадцати слов без учета союзов и предлогов, возможны варианты названий, включающие от двух до пятнадцати слов.

3. Информация об авторах. Информация об авторах включает имя и фамилию автора, место работы автора и контактные данные. Зачастую количество авторов не ограничено, но некоторые специализированные журналы

могут устанавливать ограничения, например, не более пяти авторов. Место работы автора включает название организации с указанием города и страны, реже с полным адресом места работы автора. Контактные данные чаще всего представлены адресом электронной почты. Если авторов несколько, то указывается имя автора ответственного за переписку.

4. Аннотация и ключевые слова. Стандартная аннотация содержит информацию об объекте, цели и методах исследования, основных результатах и выводах. Ключевые слова используют для быстрого поиска информации в больших коллекциях документов.

5. Введение. Во введении акцентируют внимание на новизне и актуальности работы, приводят обзор литературы предметной области исследования. Элемент «Введение зачастую заканчивается формулировкой цели работы.

6. Основная часть. Способы организации данного раздела варьируют в зависимости от цели исследования, обычно выделяют разделы «Материалы и методы», «Обсуждение», «Результаты».

7. Заключение. Отражает факты, сделанные на основе проведенной работы.

8. Слова благодарности. Благодарят тех, кто оказал помощь в проведении работы.

9. Ссылки на литературу. Представляет собой список литературы, процитированной в тексте статьи.

В результате анализа композиционной структуры текстов научно-технических статей, выявлено, что они имеют ярко-выраженную структуру, содержат определенный набор элементов, за каждым из которых закреплено свое место в тексте документа, взаимоувязанную систему заголовков разделов и подразделов.

Для решения задачи моделирования текста научно-технической статьи необходимо разработать формальные средства композиционной структуры научно-технических статей, использование которых позволит осуществлять структурную разметку в корпусе научно-технических текстов.

В результате будет получена модель формального представления текстов

научно-технических статей, которая даст возможность при разметке корпуса научно-технических текстов учитывать их композиционную структуру.

Композиционная структура научно-технической статьи в первом приближении состоит из трех частей. Первая часть включает реферативный раздел, под котором понимается совокупность основных конвенциональных элементов, используемая для формальной идентификации первичного документа с учетом его природы, элементов и порядка внешних признаков, которые отличают его от других. Компонентами реферативного раздела являются: код УДК, название статьи, информация об авторах, место работы авторов, аннотация (реферат), ключевые слова. Код УДК не является обязательным элементом статьи. Реферативный раздел зачастую имеет переводную версию на английском языке. Вторая часть статьи представляет корпус научно-технической статьи, который обычно состоит из пяти основных структурных элементов: введение, материал и методы, результаты, обсуждение результатов, заключение. Стоит отметить, что названия разделов, кроме «введения» и «заключения» могут отличаться, однако их содержание соответствует по смыслу заявленным. В связи с тем, что разные научные журналы формируют собственные требования к рукописям статей, структурные элементы корпуса научно-технической статьи могут быть разбиты на пункты и подпункты, которые в свою очередь разбиваются на абзацы и затем предложения. Третья часть статьи является вспомогательным аппаратом публикации, который включает примечание и ссылки на источники [201].

В нотациях Бекуса-Наура композиционную структуру текстов научно-технических статей можно задать следующим образом [202]:

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

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.2.1 представлена полученная структурная схема элементов текста научно-технической статьи.

Рис. 2.1 – Структурные элементы текста научно-технической статьи

На основе проведенного анализа композиционной структуры текстов научно-технических статей модель текста научно-технической статьи *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 $L = \{l_1, \dots l_5\}$, де l_1 – раздел, l_2 – пункт, l_3 – подпункт, l_4 - абзац, l_5 – предложение.

Такое представление научно-технической статьи порождает дальнейшую возможность анализировать с помощью математических методов как вероятностную, так и логическую структуру всего исследуемого текста научно-технической статьи в целом. Таким образом, модель композиционной структуры текста научно-технической статьи – это граф, вершинами и ребрами которого являются только полноценные единицы – разделы, пункты, подпункты, то есть наиболее значимые структурные элементы. Наличие структурной разметки текста научно-технической статьи при создании корпуса научно-технических текстов значительно расширит исследовательский потенциал корпуса, что в свою очередь позволит при разработке систем обработки естественного языка учитывать композиционные особенности научно-технических текстов, в целом, и их отдельных структурных компонентов, в частности.

2.2. Модель учебно-научного текста для структурной разметки в корпусе научно-технических текстов

Под учебно-научным текстом принято понимать книгу, в которой систематически изложены основы знаний в определенной предметной области на уровне современных достижений науки и техники [203]. К учебно-научным текстам выдвигают такие же требования, как и к научным текстам, а именно: логичность, краткость, ясность, последовательность изложения материала, абстрактность [204]. В работе [205] Тюрина Л.Г. описывает педагогическую модель учебной книги, в которой выделяет три подсистемы: предметную, дидактическую и аксиологическую. Связь между подсистемами представлена следующим образом: сначала излагается вербально или наглядно система знаний (предметная подсистема) о некоторой предметной области, затем идут

материалы, формирующие необходимые навыки и умения – вопросы, задания, упражнения (дидактическая подсистема). При этом элементы двух указанных подсистем строятся с учетом мировоззренческого, воспитательного воздействия на читателя – аксиологическая подсистема.

Стоит отметить, что композиционная структура учебно-научного текста также отражает выше описанную модель и включает в себя текст, как главный компонент, так и внетекстовые, вспомогательные компоненты, к которым относят аппарат организации усвоения (вопросы и задания, памятки или инструктивные материалы, таблицы и шрифтовые выделения, подписи к иллюстративному материалу и упражнения); собственно иллюстративный материал; аппарат ориентировки, включающий предисловие, примечание, приложения, оглавление, указатели [206]. Структура учебно-научного текста показана на рис.2.2. Как видно из рисунка учебно-научные тексты имеет сложную многокомпонентную структуру.

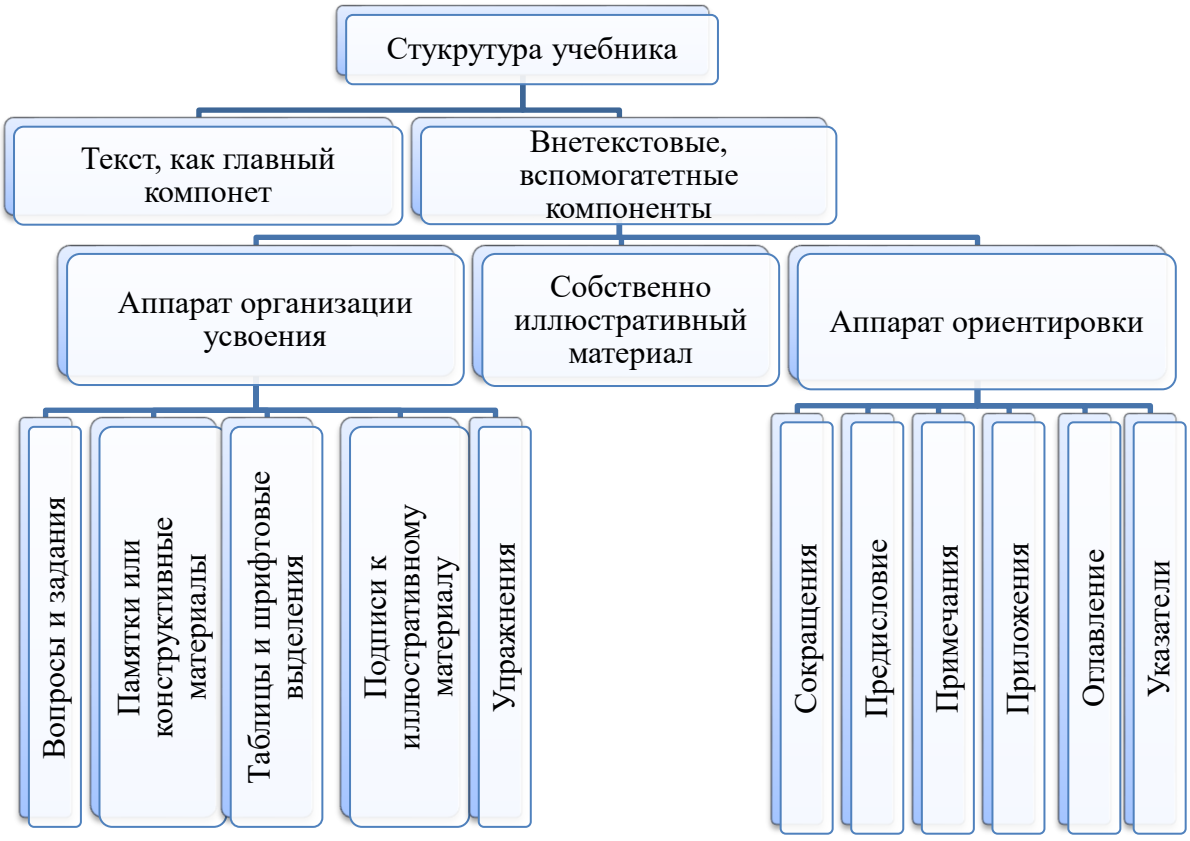


Рис. 2.2 – Структура учебно-научного текста по любой предметной области

При реализации структурной разметки корпуса научно-технических текстов необходимо проанализировать каждый компонент учебно-научного текста с целью определения оптимальных вариантов их разметки. Например, необходимость включения текстов упражнений в основную часть корпуса, так как при составлении упражнения могут быть использованы разные предметные области или их комбинации, при этом исследуемая предметная область в силу специфики не будет отражена.

Ярким примером могут служить учебники и учебные пособия по научно-техническому переводу, где тематика текстов упражнений зачастую кардинально отличаются от предмета содержания учебно-научного текста.

В результате анализа композиционной структуры учебно-научных текстов, выявлено, что они имеют ярко-выраженную структуру, содержат определенный набор элементов, за каждым из которых закреплено свое место в тексте документа.

Композиционная структура учебно-научного текста в первом приближении состоит из реферативного раздела, корпуса научно-технической статьи и информативного раздела. При этом структурные элементы научно-учебных текстов можно разделить на обязательные и факультативные, то есть те, которые приводятся в зависимости от необходимости. К обязательным элементам относят название, автор(ы), оглавление, введение, основной текст и ссылки на источники. Несмотря на то, что название и авторы являются элементами метатекстовой разметки, они также несут значимую информацию при автоматической обработке научно-технических текстов информацию и являются полноценными объектами лингвистического исследования.

Оглавление является важным элементом учебно-научного текста, дающим общее представление о структуре и проблематике учебного пособия, отражает взаимосвязи всех компонентов учебника и является средством навигации по научно-учебному тексту. Если у разных разделов учебника разные авторы, то вместо структурного элемента «Оглавление» используют элемент «Содержание» [207]. Введение в учебно-научном тексте обычно представляет читателю

информацию о текущем состоянии проблем и явлений в некоторой предметной области, обзор взглядов и литературных источников, базовую терминологию и др. Основной текст раскрывает содержание, обеспечивает последовательное, полное и аргументированное изложение материала и служит основным источником учебно-научной информации, обязательный для изучения и усвоения. Структурный элемент «Ссылки на источники» содержит основные или рекомендуемые литературные источники для углубленного или самостоятельного изучения определенный тем некоторой предметной области.

К факультативным структурным элементам учебно-научных текстов относят «Предисловие», «Вопросы», «Задания и упражнения», «Примечания» и «Приложения». Предисловие представляет собой текст, предваряющий изложение основного материала, содержит цель и особенности издания, отражает структуру и краткую характеристику всех разделов. В зависимости от типа литературы и вида издания выделяют ряд разновидностей «Предисловия»: «От автора», «От редактора», «Вместо предисловия» и др. Структурные элементы «Вопросы» и «Задания и упражнения» относят к аппарату организации усвоения материала, призванные стимулировать познавательную деятельность в процессе усвоения материала. Структурный элемент «Примечания» являются краткими дополнениями, пояснениями и уточнениями к основному учебно-научному тексту, бывают внутритекстовые, подстрочные и затекстовые. Авторы используют этот структурный элемент с целью дополнения основного учебно-научного текста [208]. В «Приложения» включают материал, служащий дополнением основного текста, куда входят официальные и справочные материалы – таблицы, схемы, словари, чертежи, списки, вклейки, иллюстрации, карты, рисунки.

В нотациях Бекуса-Наура композиционную структуру учебно-научных текстов 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. 2.3 представлена полученная структурная схема элементов учебно-научного текста.

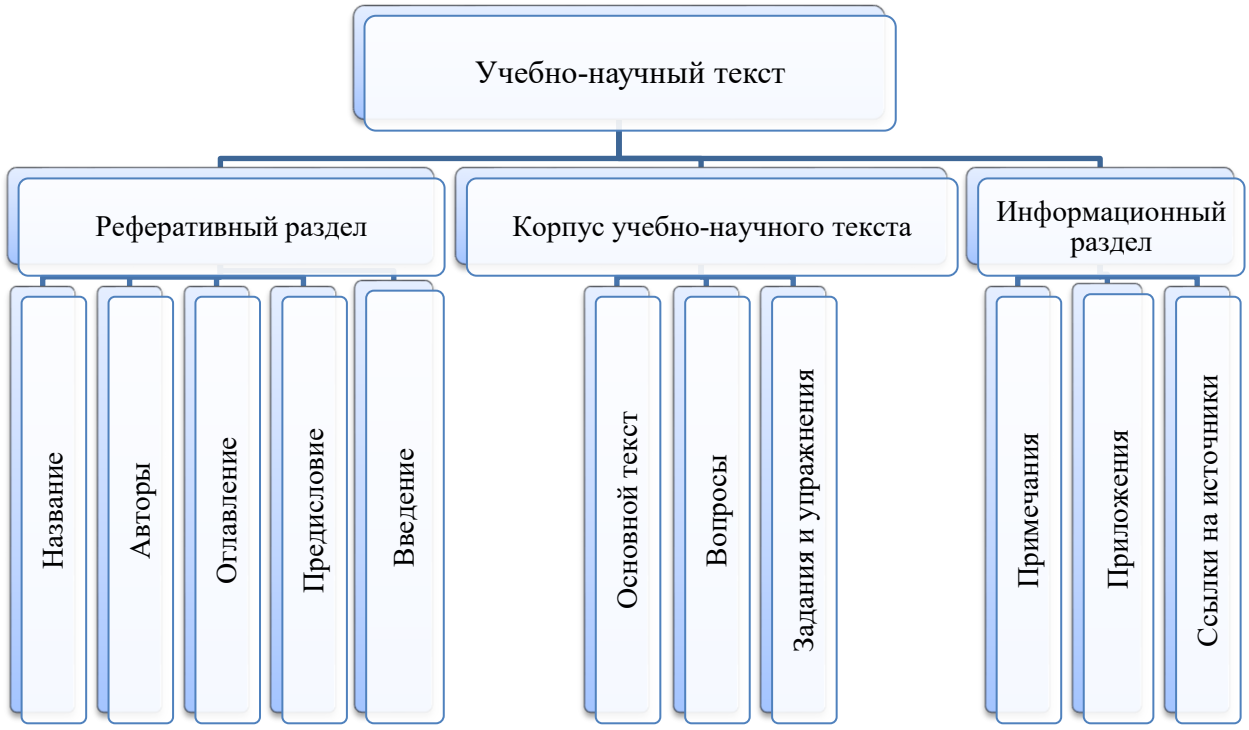


Рис. 2.3 – Структурные элементы учебно-научных текстов

На основе проведенного анализа композиционной структуры учебно-

научных текстов модель учебно-научного текста *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 $L = \{l_1, \dots l_5\}$, где l_1 – раздел, l_2 – пункт, l_3 – подпункт, l_4 - абзац, l_5 – предложение.

Представление текста в виде упорядоченного набора структурных элементов дает возможность анализировать с помощью математических методов как вероятностную, так и логическую структуру всего учебно-научного текста. Таким образом, модель композиционной структуры учебно-научного текста – это граф, вершинами и ребрами которого являются только полноценные единицы – разделы, пункты, подпункты, то есть наиболее значимые структурные элементы. Наличие структурной разметки учебно-научных текстов при создании корпуса научно-технических текстов значительно расширит исследовательский потенциал корпуса, что в свою очередь позволит при разработке систем обработки естественного языка учитывать композиционные особенности научно-технических текстов, в целом, и их отдельных структурных компонентов, в частности.

2.3. Модель текста стандарта как иерархически-структурированного текста

В современной лингвостилистике статус стилевой принадлежности текстов нормативной базы, в частности стандартов, является неопределенным. Именно поэтому, необходимо провести анализ текстов стандартов, определить принципиальные для автоматизированного информационного поиска характеристики, и выделить именно те, которые будут иметь существенное влияние на качество информационного поиска в коллекции таких текстов.

Анализ текстов языка стандартов показал, что его основными чертами является максимальная ясность изложения, краткость и лаконичность высказываний, краткость и четкость формулировок, не допускающих различных

толкований, определенность информации, динамичность и экспрессивность ее передачи, однозначность ее восприятия.

С целью соблюдения требований лаконичности, краткости и максимального отображения контента в заголовках и подзаголовках в текстах языка стандартов можно встретить такие подзаголовки, на примере подзаголовков [209]:

7.1.1. Общие сведения

7.2.1. Общие требования

7.3.1 Общие положения

7.4. Документация

Основной особенностью организации языковых средств в языке стандартов является их обобщенно-отвлеченный характер на лексическом и грамматическом уровнях языковой системы. Лексику языка стандартов составляют три основных пласта: общеупотребительные слова, общенаучные слова и термины.

Стремлением к информационной насыщенности обусловлен отбор наиболее емких и компактных синтаксических конструкций. В языке стандартов преобладают простые распространенные и сложноподчиненные предложения. Среди первых наиболее употребительны неопределенно-личные с прямым дополнением в начале предложения, синонимичные пассивным конструкциям. Абстрактность и обобщенность языка стандартов на синтаксическом уровне выражается, прежде всего, в широком использовании пассивных конструкций, безличных предложений разных типов.

Частотность употребления тех или иных типов сложных предложений определяется такой специфической чертой языка стандартов как логичность. Среди сложных предложений преобладают сложносочиненные и сложноподчиненные с четко выраженной синтаксической связью между отдельными частями. Преобладание союзных предложений над бессоюзными

объясняется тем, что с помощью союзов связь между частями сложного предложения выражается более точно, однозначно. Среди союзных предложений наиболее употребительны сложноподчиненные, так как при подчинении взаимоотношения между отдельными предложениями выражаются более четко. Наиболее распространенным и типичным для языка стандартов видом связи предложений является повторение существительных, часто в сочетании с указательными местоимениями *этот, тот, такой* [14].

Проведенный анализ текстов показал, что синтаксические структуры предложений в текстах стандартов в большинстве случаев однотипны. Так, например, был проанализирован стандарт [210]; в тексте настоящего стандарта содержится 513 предложений (вступительная часть и приложения не были приняты к рассмотрению), из которых 294 являются простыми предложениями, 220 - сложными, в том числе к сложным предложениям относим и перечисление, которых в тексте стандарта - 59.

Простые предложения могут быть осложнены причастными и деепричастными оборотами, которые модифицируют значение того члена предложения, к которому они относятся. Сложносочиненные предложения встречаются реже, чем сложноподчиненные. Так как сложносочиненное предложение состоит из двух и более простых предложений, то их разбор идентичен разбору простых предложений, что дает возможность в процессе информационного поиска использовать семантическую модель языковых объектов.

Сама структура простых предложений в составе сложноподчиненных идентична простому предложению, но следует учитывать особенности придаточного предложения. В текстах стандартов существует три вида придаточных предложений в составе сложноподчиненных, а именно: часть-целое, причина-следствие, условие-причина.

Каждое требование отражает степень соблюдения требования, что на лексическом уровне выражено использованием модальных глаголов. Одна часть рассматриваемых глаголов выражает реальную, утвердительную модальность,

достоверность, бесспорность; другая - гипотетическую модальность, предположение. Во второй группе можно выделить модальные слова, выражающие возможность, вероятность т. п. В Табл. 2.2 приведены результаты анализа текстов стандартов, а подчеркиванием выделены особенности изучаемого языка для специальных целей, которые должны быть учтены во время решения задачи автоматизированного информационного поиска в коллекции текстов нормативной базы программной инженерии.

Таблица 2.2 – Результаты анализа текстов стандартов
(на примере нормативной базы программной инженерии)

Сфера функционирования и типовая ситуация общения		Программная инженерия. Регулирование деятельности в сфере программной инженерии.
Участники речи		Специалисты, владеющие обширными знаниями о предметной области
Функция речи		Нормативное регулирование качества программного обеспечения
Тип содержания и предмет речи (тема)		Содержание – конкретное. Тема – программная инженерия, в частности, сертификация программного обеспечения
Коммуникативная цель		Регулирование сферы деятельности программной инженерии. Представление требований к программному обеспечению. Контроль над соблюдением качества программного обеспечения
Стилеобразующие признаки		Объективность и точность. Обезличенность информации. Шаблонность, четкое закрепление места за элементами.
СПЕЦИФИЧЕСКИЕ ХАРАКТЕРИСТИКИ	▪ Композиционная структура – структурное членение используется как важное средство логического членения текста ▪ Требование, как правило, представлено 1-2 предложениями. ▪ Лексика – общенаучная и узкоспециализированная терминология предметной области программной инженерии ▪ Глагольная форма настоящего времени имеет абстрактное значение ▪ Наличие глаголов с ослабленным лексическим значением (быть, служить). ▪ Качественные прилагательные чаще употребляются в сокращенной форме ▪ Употребление слов и предложений с различными модальными значениями. ▪ Прямой порядок слов в предложении. ▪ Пассивные конструкции. Перечисления. ▪ Однородные члены предложения с обобщающими словами. ▪ Преобладают простые предложения. ▪ Выражены условно-следственные и причинно-следственные отношения.	
	Форма бытия	Письменная, в виде текстов

Тексты стандартов языка имеют четко выраженную композиционную структуру, которая просматривается в четком распределении структурных элементов текста стандартов, его деление на разделы, подразделы, пункты, подпункты. В свою очередь, такое разделение порождает особенность: значения нижних элементов стандартов можно установить однозначно только с учетом значения верхних элементов. В качестве примера приведен пример из стандарта МЭК 60880 Атомные электростанции. Системы контроля и управления, важные для безопасности. Программное обеспечение компьютерных систем, выполняющих функции категории А [211]:

МЭК 60880 АТОМНЫЕ ЭЛЕКТРОСТАНЦИИ. Системы контроля и управления, важные для безопасности. Программное обеспечение компьютерных систем, выполняющих функции категории А

10. Программные аспекты валидации системы

10.3 Программные аспекты отчета о валидации системы

10.3.5 В данном отчете должна быть дана оценка соответствия системы всем требованиям.

В приведенном примере содержание требования 10.3.5 можно уточнить через название подразделения, к которому оно относится. В свою очередь значение заголовка подраздела 10.3 и раздела 10 уточняются через заголовок раздела 10 и названия самого стандарта, соответственно.

Сложность автоматизации задачи анализа иерархически структурированных текстов обусловлена такими их свойствами:

–как правило, разметка заголовков и маркеров (с помощью стилей, тегов и т. д.) в документе присутствует лишь частично или вообще отсутствует;

–заголовки разных уровней иерархии могут быть отличаются по виду;

–заголовок и ссылка на него в тексте могут иметь одинаковый вид;

–разнообразие конфигураций непрерывных текстовых фрагментов: предложение может состоять из нескольких таких фрагментов, один фрагмент

может включать несколько предложений, группа предложений может быть вложена в предложение в виде комментария.

В процессе анализа нормативной базы автором выявлено несколько типов стандартов в соответствии с их иерархическими особенностями, а именно:

– документы с суровой сквозной нумерацией, например: ЧП 306.5.02/3.035-2000. Требования по ядерной и радиационной безопасности к информационным и управляющим системам, важным для безопасности атомных станций [212];

– документы без нумерации, но четкой структурой с возможностью определения заголовков, подзаголовков и другое, например: NUREG 6303. Method for Performing Diversity and Defense-in-Depth Analyses of Reactor Protection Systems [213];

– смешанный тип – как имеющаяся нумерация и заголовки выделены жирным шрифтом, но не пронумерованы или нумерация по разделам - в каждом новом разделе новая нумерация, например: NS-G-1.1. ПО для систем, базирующихся на компьютерах, важных для безопасности атомных станций. Руководство по безопасности [214].

Структурная формализация – это формализация самого способа изложения материала. Она проявляется во всей номенклатурной организации языка стандартов от принципов ее классификации до внешнего оформления композиционной структуры текстов.

Одним из средств структурной формализации текстов стандартов выступает рубрикация, что является внешним выражением их композиционной структуры. Рубрикация соответствует разбивке составляющие текста на составные части. Она включает графический распределение частей, использование заголовков, подзаголовков, пунктов, подпунктов, абзацев, нумерацией и пунктуации.

Основную часть текста стандарта принято делить на разделы, подразделы, пункты, подпункты. Обязательным является наличие заголовков разделов и подразделов. Заголовки (подзаголовки) является важным средством рубрикации

и, следовательно, элементом структурной формализации. Они в достаточно сжатой, краткой и лаконичной форме отражают тематику и основную идею выделенной части документа, выступая тем самым как важнейшие единицы сообщения, передающие определенную информацию.

Пунктуация является не только элементом рубрикации и структурной формализации, но и несет определенное семантико-синтаксическое и функционально-коммуникативное значение, которое следует учитывать при реализации процедуры информационного поиска в текстах стандартов.

Текст стандарта – это письменное уведомление, объективированное в виде письменного документа, состоящего из ряда высказываний, объединенных различными типами лексической, грамматической и логической связи, имеющего определенный моральный характер, прагматическую установку и соответственно литературно обработанное [215].

Тексты стандартов как класс документов отличаются выраженной и внешне оформленной композиционной структурой, а одним из средств оформления этой структуры является рубрикация, т.е. разбиение текста на составные графически-распределенные части, а также употреблением обобщенно-отвлеченных лексических единиц при изложении самих требований [216].

Основную часть текста стандарта принято делить на разделы, подразделы, пункты, подпункты. Обязательным является наличие заголовков разделов и подразделов. Заголовки (подзаголовки) является важным средством рубрикации и, следовательно, элементом структурной формализации. Структурные элементы стандарта делят на следующие элементы: титульный лист, предисловие, содержание, введение, название, область применения, нормативные ссылки, термины и определения понятий, обозначения и сокращения, требования к объекту стандартизации, приложения и библиографические данные [14].

Титульный лист является первой страницей; предисловие размещают на второй странице. Содержание включает в себя порядковые номера и названия

разделов, приложений с обозначением их заголовков. Введение приводят при необходимости обосновать причины разработки стандарта. Структурный элемент область применения приводят для обозначения область его применения и, при необходимости, уточнения объекта стандартизации. Структурный элемент нормативны ссылки содержит перечень стандартов, на которые в тексте стандарта приведены ссылки. Элемент термины и обозначения содержит определения, необходимые для уточнения и определения значений терминов, использованных в стандарте. В приложениях могут быть материалы, которые дополняют положения стандартов. Приложениями, например, могут быть графические материалы, таблицы большого формата, расчеты, описание приборов, алгоритмов, программ. Приложения могут быть обязательными и информационными. Информационные приложения могут быть рекомендательного и справочного характера. В стандарте также указывают библиографические данные [217].

Структурные элементы, за исключением элементов «Титульный лист», «Предисловие», «Наименование», «Требования к объекту стандартизации», приводят при необходимости, в зависимости от особенностей объекта стандартизации. Названия структурных элементов в текстах стандартов короткие, точно характеризируют объект стандартизации. Названия стандартов, как правило, не содержат сокращений, римских цифр, математических знаков, греческих букв.

Наличие повторяющихся не только по тексту одного стандарта, а целых коллекций текстов стандартов заголовков и подзаголовков делает необходимым представление структуры текста стандарта в виде иерархии, что, в свою очередь, дает возможность при информационном поиске учесть место и взаимосвязь заголовка/подзаголовка в тексте стандарта, а также дает однозначное представление о содержании этого фрагмента текста.

Четкое распределение структурных элементов текстов нормативной базы, их деление на разделы, подразделы, пункты, подпункты порождает следующую особенность: значения нижних элементов стандартов можно установить

однозначно только с учетом значения верхних элементов. В качестве примера приведен пример из стандарта МЭК 60880 Атомные электростанции. Системы контроля и управления, важные для безопасности. Программное обеспечение компьютерных систем, выполняющих функции категории А (Рис. 2.4):

МЭК 60880 АТОМНЫЕ ЭЛЕКТРОСТАНЦИИ. Системы контроля и управления, важные для безопасности. Программное обеспечение компьютерных систем, выполняющих функции категории А

10. Программные аспекты валидации системы

10.3 Программные аспекты отчета о валидации системы

10.3.5 В данном отчете должна быть дана оценка соответствия системы всем требованиям.

Рис. 2.4 – Фрагмент представления иерархической структуры текста стандарта ГОСТ Р МЕК 60880

Как показано на Рис. 2.4 отличительной особенностью текстов стандартов можно назвать краткость и лаконичность заголовков, в состав которых входят лексические единицы, обозначающие обобщенные понятия. При этом у человека данный факт не вызывает трудностей в силу того, что он погружен в

иерархический контекст самого текста. Так, для уточнения смысла заголовка 10.3.5 нужно обратиться к названию подраздела 10.3, а значение заголовка подраздела 10.3 и раздела 10 уточняются через заголовок раздела 10 и названия самого стандарта, соответственно.

В результате анализа композиционной структуры и лексических особенностей текстов стандартов, выявлено, что они имеют ярко-выраженную иерархическую структуру, содержат определенный набор элементов, за каждым из которых закреплено свое место в тексте документа, взаимоувязанную систему заголовков разделов и подразделов.

Текст стандарта целесообразно разделить на три части: предварительная часть, рекомендации и требования, и информативная часть. К предварительной части относится вступление, назначение, нормативные ссылки, термины и определения, общие сведения. Часть рекомендаций и требований содержит разделы стандарта, в которых непосредственно описываются требования. Информативная часть содержит приложения и библиографические ссылки [217].

В нотациях Бекуса-Наура композиционную структуру текстов стандартов можно задать следующим образом:

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

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:

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

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. 2.5 представлена полученная структурная схема элементов текста стандарта.

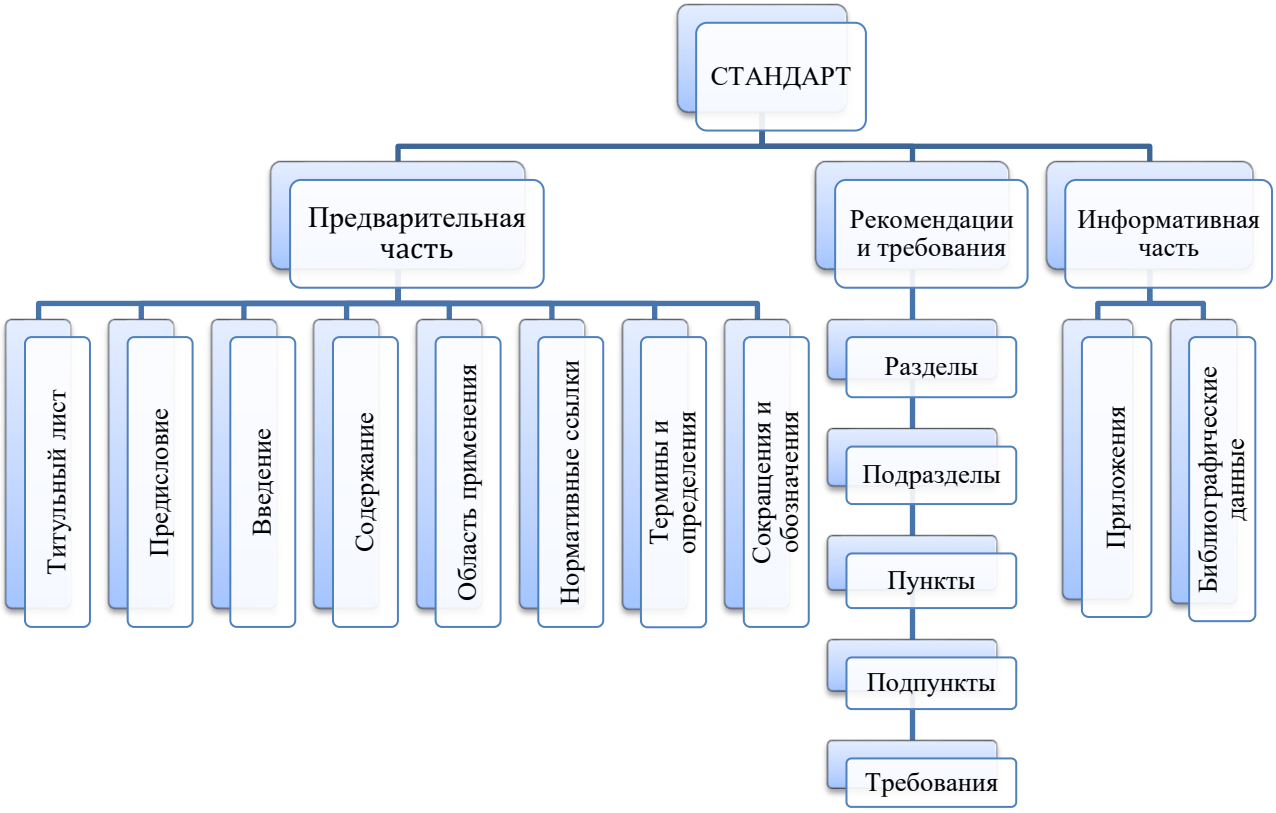


Рис. 2.5 – Структурные элементы текста стандарта

На основе проведенного анализа композиционной структуры текстов стандартов модель текста стандарта *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 l_1 – стандарт, l_2 – основная часть, l_3 – раздел, l_4 - подраздел, l_5 – пункт, l_6 – подпункт, l_7 – требование,

l_8 – предложение.

Такой подход порождает дальнейшую возможность анализировать с помощью математических методов как вероятностную, так и логическую структуру всего исследуемого текста стандарта в целом. Описанная выше модель дает возможность в процессе компьютерного анализа текста стандарта определить тип структурного элемента, степень вложенности, за счет подачи стандарта в виде конечного множества его составных частей.

Обработка поискового запроса *отчет от тестирования ПО АЭС* с помощью разработанной модели сначала провести поиск в названиях документов, затем в названиях разделов и дальше в самих требованиях, например:

Стандарт	***Атомные электростанции**. Системы контроля и управления, важные для безопасности. **Программное обеспечение** компьютерных систем, выполняющих функции категории А*
Раздел	***Отчёт о тестированиях** ПО*
Требования	*8.2.3.1.3.1 В отчёте о тестированиях ПО должны быть представлены результаты верификации, описанные в спецификации тестирований ПО и устанавливающие, работает или нет ПО в соответствии со спецификацией проекта ПО.* *8.2.3.1.3.2 В данном документе должны быть отмечены все расхождения между проектом и реализацией, обнаруженные в процессе тестирований.* *8.2.3.1.3.3 Отчёт о тестированиях ПО должен включать следующие пункты, как на уровне модуля, так и на уровне основного проекта…*

Таким образом, при обработке текстов стандартов в параллельном корпусе научно-технических текстов необходима разработка специальных методов поиска при обработке информационных запросов к параллельному корпусу, которые будут учитывать лексические и композиционные особенности текстов. При этом, информационный поиск целесообразно проводить в два этапа. На первом этапе в силу того, что названия структурных элементов содержат наиболее значимую информацию о содержании всего текста, необходимо осуществлять отбор самих стандартов и/или их разделов, а на втором этапе

осуществлять построчный поиск лексических единиц из поискового запроса.

2.4. Особенности обработки композиционной структуры научно-технических текстов в параллельном корпусе

В настоящее время тексты, отобранные для наполнения корпуса чаще всего, хранятся в пдф формате и соответственно размечены с использованием XML тегов. Описание на языке XML представляет собой операторы, написанные с соблюдением определенного синтаксиса. При этом при создании XML-документа вместо использования ограниченного набора определенных элементов можно создавать собственные элементы и присваивать им любые имена, позволяет использовать XML для описания практически любого документа, а также в перспективе расширить спектр возможных видов разметок научно-технических текстов в параллельном корпусе.

<тема >

<название>Режимы резонансной прозрачности в условиях синхронизма длинных и коротких волн</название>

<дата публикации>13 февраля 2004г. </дата публикации>

<авторы>

<автор>С.В. Сазонов</автор>

<автор>Н.В. Устинов</автор>

</авторы>

</тема >

<аннотация>Исследованы особенности нелинейного ...

</аннотация>

Помимо стандартных тегов предусмотрены специальные теги, отражающие особенности научно-технических текстов. Перечень основных тегов представлен в таблице 2.3.

Таблица 2.3. Тэги в научно-технических текстах

	Тэг	Уровень разметки
1	<inventory>	коллекции текстов по направлению подготовки
2	<article>	научно-технические статьи
3	<textbook>	учебные пособия
4	<standard>	нормативная документация
5	<title>	название
6	<author>	автор
7	<organization>	аффилиация
8	<e-mail>	электронная почта
9	<abstract>	аннотация
10	<key words>	ключевые слова
11	<introduction>	введение
12	<main body>	основная часть текста
13	<passage>	абзац
14	<sentence>	предложение
15	<term>	термин
16	<phrase>	сочетание слов
17	<example>	пример
18	<citation>	цитата

Выравнивание параллельных текстов состоит в установлении соответствия фрагментов текста на языке источника фрагментам текста на языке перевода. Таким образом, можно устанавливать соответствия между структурными элементами текстов на разных языках, как показано на Рис.2.6.

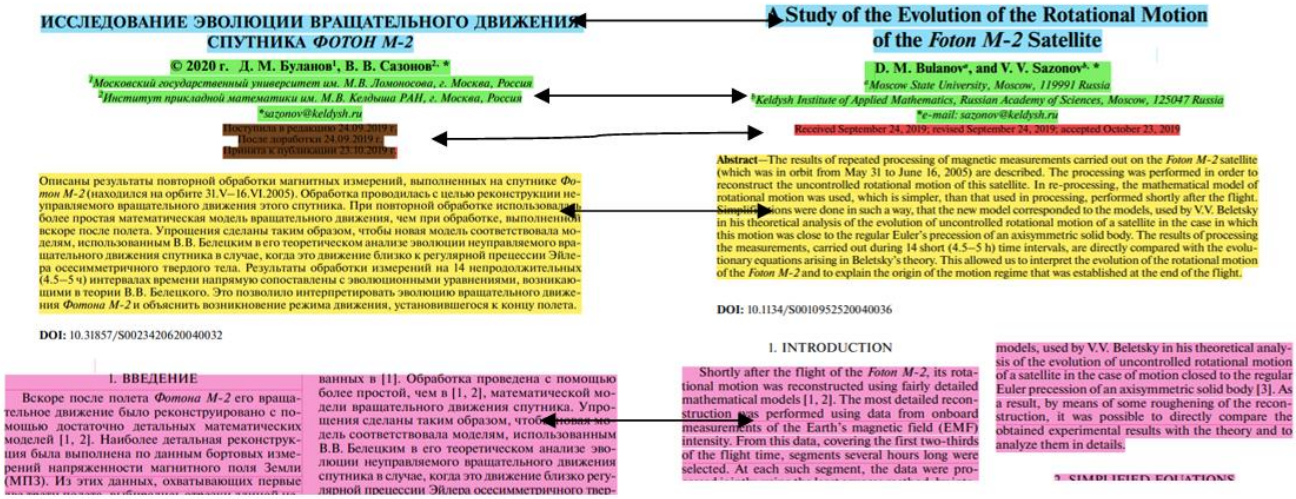


Рис. 2.6 – Выравнивание структурных элементов текста

На основе анализа, проведенного в главе 1, научно-технические тексты в параллельном корпусе необходимо выравнивать по не только по структурным элементам, абзацам, предложениям, но и словосочетаниям разных типов.

Проблемы асимметрии параллельных текстов вызваны ассиметричным членением предложения и асимметрией на уровне перевода, что может быть вызвано рядом факторов. При переводе следующих друг за другом предложений смысловая часть одного предложения на языке оригинала может быть отражена во втором предложении на языке перевода. На практике также встречаются случаи, когда фрагмент не имеет перевода или перевод содержит неточности или ошибки, а также могут быть использованы переводческие трансформации или изменена линейная последовательность фрагментов предложений или целых предложений [218].

На уровне предложений часто возможны объединения простых предложений в сложное или разбиение сложного предложения в несколько простых. Тогда в процессе выравнивания текста одному предложению в одном языке будет соответствовать два предложения в другом языке. Объединение или разбиение на 3 и более предложений в научно-технических текстах практически не встречается.

Выравнивания на уровне слов возникают следующие ситуации:

1. есть однозначные соответствия слово-слово в обоих языках;

2. есть соответствия, но имеют разную формальную структуру, например, слову соответствует словосочетание в другом языке;

3. нет соответствия в другом языке, например, при переводческой трансформации – опущение;

4. отсутствие слова в языке оригинала и его наличие в тексте перевода;

5. слово отражено другой частью речи или было передано другими языковыми средствами.

Структурная разметка научно-технических текстов позволяет повысить эффективность обработки научно-технических текстов, с одной стороны, за счет их обработки как совокупности взаимосвязанных элементов, а с другой стороны,

обрабатывать отдельные элементы определенной совокупности текстов.

Так, при реализации терминологической разметки структурная разметка позволяет исключить структурные элементы текста, в которых отсутствуют терминологические единицы: например: фамилии и имена авторов, аффилиация, адрес электронной почты, благодарности, список использованных источников и др. Однако, данные элементы текста перегружают систему извлечения терминов словосочетаниями, которые по свой структуре схожи со структурой терминологических единиц.

В таблицах 2.4 и 2.5 представлены результаты анализа эффективности извлечения терминологических единиц из параллельных русско- и англоязычных текстов научно-технических статей со структурной разметкой терминов и без нее.

Таблица 2.4. Анализ эффективности разметки русскоязычных терминов с предварительной структурной разметкой

		терминов-кандидатов			
		всего	основная часть	не обраб.	
				терминов	%
1	CR-2019-1-1	687	462	225	**32,8**
2	CR-2019-1-2	649	606	43	**6,6**
3	CR-2019-1-3	351	290	61	**17,4**
4	CR-2019-1-4	808	559	249	**30,8**
5	CR-2019-1-5	634	403	231	**36,4**
6	CR-2019-1-6	499	384	115	**23,0**
7	CR-2019-1-7	466	335	131	**28,1**
8	CR-2019-1-8	871	674	197	**22,6**
9	CR-2019-1-9	621	551	70	**11,3**
10	CR-2019-2-1	664	466	198	**29,8**
11	TF-2019-1-1	369	266	103	**27,9**
12	TF-2019-1-2	593	471	122	**20,6**
13	TF-2019-1-3	507	358	149	**29,4**
14	TF-2019-1-4	651	402	249	**38,2**
15	TF-2019-1-5	597	381	216	**36,2**
16	TF-2019-1-6	622	447	175	**28,1**
17	TF-2019-1-7	697	511	186	**26,7**
18	TF-2019-1-8	545	375	170	**31,2**
19	TF-2019-1-9	836	569	267	**31,9**

20	TF-2019-1-10	1145	819	326	**28,5**
21	FP-2019-1-1	863	711	152	**17,6**
22	FP-2019-1-2	508	346	162	**31,9**
23	FP-2019-1-3	576	450	126	**21,9**
24	FP-2019-1-4	712	448	264	**37,1**
25	FP-2019-1-5	1047	787	260	**24,8**
26	FP-2019-1-6	568	421	147	**25,9**
27	FP-2019-1-7	971	668	303	**31,2**
28	FP-2019-1-8	412	315	97	**23,5**
29	FP-2019-1-9	622	468	154	**24,8**
30	FP-2019-1-10	604	465	139	**23,0**
	Среднее	**653**	**480**	**173**	**26,6**

Таблица 2.5. Анализ эффективности разметки англоязычных терминов с предварительной структурной разметкой

		терминов-кандидатов			
		всего	основная часть	не обраб.	
				терминов	%
1	CR-2019-1-1	779	505	274	**35,2**
2	CR-2019-1-2	702	623	79	**11,3**
3	CR-2019-1-3	448	321	127	**28,3**
4	CR-2019-1-4	981	634	347	**35,4**
5	CR-2019-1-5	835	493	342	**41,0**
6	CR-2019-1-6	565	392	173	**30,6**
7	CR-2019-1-7	550	379	171	**31,1**
8	CR-2019-1-8	983	765	218	**22,2**
9	CR-2019-1-9	690	571	119	**17,2**
10	CR-2019-2-1	780	488	292	**37,4**
11	TF-2019-1-1	407	294	113	**27,8**
12	TF-2019-1-2	663	542	121	**18,3**
13	TF-2019-1-3	573	398	175	**30,5**
14	TF-2019-1-4	733	431	302	**41,2**
15	TF-2019-1-5	637	418	219	**34,4**
16	TF-2019-1-6	684	483	201	**29,4**
17	TF-2019-1-7	703	524	179	**25,5**
18	TF-2019-1-8	712	478	234	**32,9**
19	TF-2019-1-9	885	597	288	**32,5**
20	TF-2019-1-10	1233	837	396	**32,1**
21	FP-2019-1-1	973	799	174	**17,9**
22	FP-2019-1-2	638	413	225	**35,3**
23	FP-2019-1-3	596	455	141	**23,7**
24	FP-2019-1-4	816	445	371	**45,5**

25	FP-2019-1-5	1123	799	324	**28,9**
26	FP-2019-1-6	642	435	207	**32,2**
27	FP-2019-1-7	1056	649	407	**38,5**
28	FP-2019-1-8	458	312	146	**31,9**
29	FP-2019-1-9	728	485	243	**33,4**
30	FP-2019-1-10	711	510	201	**28,3**
	Среднее	**742**	**515**	**227**	**30,3**

Таким образом, структурная разметка является важным элементов обработки научно-технических текстов и позволяет сократить на 25-30% объем обрабатываемых слов при терминологической разметке параллельного корпуса. Также, структурная разметка позволяет выделить определённые элементы для анализа, например, заголовки научно-технических статей.

2.5. Выводы по главе

Выявлено, что научно-технические тексты как источники наполнения параллельного корпуса обладают рядом особенностей, которые прежде всего выражены в их композиционной структуре, наборе обязательных и факультативных элементов, а также их закреплении в строго определённой последовательности в тексте.

Получены следующие выводы:

1. Выявлено, что композиционная структура научно-технической статьи состоит из трех частей: реферативного раздела, включающего УДК, название статьи, информацию об авторах, место работы и контактную информацию об авторах, аннотацию и ключевые слова; корпус научно-технической статьи, включающий введение, материалы и методы, результаты, обсуждение результатов и заключение; а также информационный раздел, куда входят примечания и ссылки на источники. Структурные элементы УДК, примечания приводятся в текстах статей факультативно. Охарактеризовано приблизительное содержание каждого структурного элемента.

2. Показано, что структурные элементы научно-учебного текста делятся на обязательные, а именно название, автор(ы), оглавление, введение, основной

текст и ссылки на источники, и факультативные: предисловие, вопросы, задания и упражнения, примечания и приложения. Охарактеризовано приблизительное содержание каждого структурного элемента.

3. Проанализированы лингвистические особенности текстов стандартов, выделены обязательные и факультативные структурные элементы текста стандарта. Показано, что наиболее значимыми особенностями текстов стандартов являются строгая иерархическая композиция и взаимосвязь элементов стандарта, а также особенности наименования заголовков, которые носят максимально обобщенный характер. Подчеркнуто, что обработку текстов стандартов целесообразно проводить в два этапа: на первом этапе отбирать стандарты или их фрагменты, а на втором уже проводить сам поиск.

4. Показано, что модель композиционной структуры научно-технического – это граф, вершинами и ребрами которого являются только полноценные единицы – разделы, пункты, подпункты, то есть наиболее значимые структурные элементы. Наличие структурной разметки текста научно-технической статьи при создании корпуса научно-технических текстов значительно расширит исследовательский потенциал корпуса, что в свою очередь позволит при разработке систем обработки естественного языка учитывать композиционные особенности научно-технических текстов, в целом, и их отдельных структурных компонентов, в частности.

6. Выделены уровни выравнивания научно-технических текстов в параллельном корпусе. Показана необходимость выравнивания межуровневых элементов, к которым относят многокомпонентные и литеральные термины, некоторые виды переводческих трансформаций. Проанализированы способы выравнивания неэквивалентных переводческих соответствий, а также способы их отражения в параллельном корпусе научно-технических текстов.

Основные результаты к разделу опубликованы в работах [3, 14-16, 29, 71, 145, 197, 202, 216].

3. МОДЕЛИ И МЕТОДЫ РАЗМЕТКИ И ВЫРАВНИВАНИЯ АНГЛО- И РУССКОЯЗЫЧНЫОЙ СПЕЦИАЛЬНОЙ ЛЕКСИКИ В ПАРАЛЛЕЛЬНОМ КОРПУСЕ

3.1. Структурные модели русско- и англоязычных многокомпонентных терминов

Длина и структура терминологической единицы являются одними из острейших проблем терминоведения. Термин может быть однокомпонентным и состоять из ключевого слова или представлять собой терминологическую группу, в состав которой входит ключевое слово или ядро группы, одно или несколько левых определений и одно или несколько правых, или предложных определений, которые уточняют или модифицируют смысл терминологической единицы [219]. При анализе терминологии также часто возникает вопрос о количестве компонентов в составе термина: например, в термине *отчет о поступлении и расходовании средств* их четыре или шесть, то есть необходимо ли учитывать служебные слова как отдельные элементы термина. Не менее дискуссионным и сложным является вопрос о максимальном количестве компонентов [220]. На практике встречаются термины, которые состоят из двенадцати-тринадцати компонентов, что представляет сложности при их употреблении в практической деятельности, поэтому по прошествии некоторого времени они чаще всего уменьшаются до 2-5 компонентных терминов.

В структуре термина можно выделить ядерный элемент, левые и правые определения [221]. Ядерный элемент – главный элемент в структуре многокомпонентного термина, который вступает в словоизменительную парадигму в предложении. Левое и правое определение 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.

При реализации традиционных методов извлечения терминов на первом этапе проводится лемматизация, суть который состоит в приведении всех слов в начальную форму. Такой подход приводит к тому, что правые определения выделяются как отдельные термины. На рис. 3.1 показана модель наследования грамматических характеристик элементов термина при его нормализации, где стрелками показаны грамматические характеристики, которые должны совпадать терминологической единице при корректном извлечении терминов с правыми определениями.

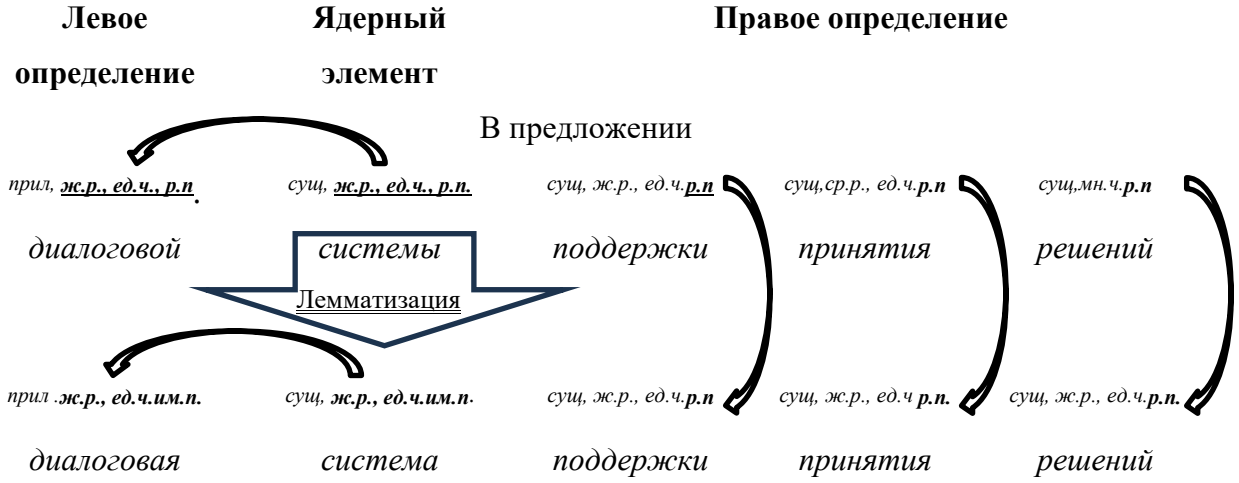


Рис. 3.1. Модель наследования грамматических характеристик при нормализации многокомпонентного термина

Учитывая вышесказанное, при создании моделей и методов автоматического извлечения многокомпонентных терминов необходимо ориентироваться на возможности машинной обработки научно-технических текстов, например, при учете количества компонентов учитывать предлоги и союзы как отдельные элементы термина, а при создании системы разметки терминов реализовать возможность увеличения количества компонентов термина. А также учитывать структуры терминов, в состав которых могут входить символы разных алфавитов, цифры, химические формулы и т.д.

Основой для реализации терминологической разметки в параллельном корпусе выступает морфологическая разметка, которая представляет собой

процесс выделения слова в тексте (корпусе) как соответствующего определенной части речи на основе как его определения, так и контекста. Процесс автоматической разметки частей речи сложный, потому что некоторые слова могут представлять собой более одной части речи в разных ситуациях. В естественных языках большой процент словоформ неоднозначен.

Помимо разметки частей речи (имя существительное, глагол, имя прилагательное, предлог, местоимение, наречие, союз, междометие, причастие, деепричастие), морфологическая разметка различает формы числа, рода, падежей. Глаголы размечаются по времени, виду, лицу, наклонению.

Морфологическая информация, приписываемая произвольному слову в тексте, состоит из четырех «полей», или групп помет:

• лексема, которой принадлежит словоформа (указывается «словарная запись» данной лексемы и ее принадлежность к той или иной части речи).

• множество грамматических признаков данной лексемы, или словоклассифицирующие характеристики (например, род для существительного, переходность для глагола).

• множество грамматических признаков данной словоформы, или словоизменительные характеристики (например, падеж для существительного, число для глагола).

• информация о нестандартности грамматической формы, орфографических искажениях и т. п.

Частеречная принадлежность элементов многокомпонентных терминов также имеет свои отличительные особенности: в состав многокомпонентного термина могут входить следующие части речи: имя существительное, имя прилагательное, имя числительное, наречение, предлог, причастие. При этом такие части речи как глагол, местоимение, междометие, а также любые знаки препинания не участвуют в образовании многокомпонентных терминов [2].

Стоит отметить, что возможности морфологических анализаторов для русского языка достигли достаточно высокого уровня развития, однако на практике встречаются некоторое ошибки, например:

1) неверно определены падеж имен существительных:

Когда^{СОЮЗ} *потребитель*$^{СУЩ,од,мр\ ед,им}$ *хочет*^{ГЛ,несов,перех,ед,3л,наст,изъяв} *сохранить*^{ИНФ,сов,перех} *анонимность*$^{СУЩ,неод,жр\ ед,им}$ (именительный падеж вместо винительного);

2) причастия и деепричастия часто распознаются как имена прилагательные:

Модель^{СУЩ,неод,жред,им} *построенная*^{ПРИЧ,сов,перех,прош,страд,жр,ед,им} *с*^{ПР} *использованием*^{СУЩ,неод,ср,ед,тв} *весов* ^{СУЩ,неод,мн,рд} *факторов*^{СУЩ,неод,мн,рд,} *потребует*^{ГЛ,сов,перех,ед,3л,буд,изъяв} *дополнительных*$^{ПРИЛ\ мн,рд}$ *форм*^{СУЩ,неод,жр,мн,рд,} *связанных*$^{ПРИЛ,кач\ мн,рд}$ *с*^{ПР} *учётом*$^{СУЩ,неод,мр\ ед,тв}$ *воздействия*^{СУЩ,неод,ср,ед,рд} *каждого*$^{ПРИЛ,мест-п,ср,ед,рд}$ *фактора*^{СУЩ,неод,мр,ед,рд,} *но*^{СОЮЗ} *даст*^{ГЛ,сов,перех,ед,3л,буд,изъяв} *более*^{Н} *точную*$^{ПРИЛ,кач\ жр,ед,вн}$ *картину*$^{СУЩ,неод,жр,ед,вн.}$ (причастие «*связанных*» распознано как имя прилагательное)

3) неверно определяется падеж имен прилагательных, причастий и деепричастий:

Целевой$^{ПРИЛ\ жр,ед,дт}$ аудиторией^{СУЩ,неод,жр,ед,тв} в^{ПР} этом^{МС,ср,ед,пр} случае^{СУЩ,неод,мр,ед,пр} выступает^{ГЛ,несов,неперех,ед,3л,наст,изъяв} большая^{ПРИЛ,кач,жр,ед,им} часть^{СУЩ,неод,жр,ед,им} населения$^{СУЩ,неод,ср\ ед,рд}$ страны$^{СУЩ,неод,жр\ ед,рд.}$ (в словосочетании «целевой аудиторией» не совпадают падежи);

4) не распознаны сокращения:

Слушатели$^{СУЩ,од,мр\ мн,им}$ *радиоканала*$^{СУЩ,неод,мр\ ед,рд}$, *участники*$^{СУЩ,од,мр\ мн,им}$ *интернет*$^{СУЩ,неод,мр\ ед,им}$*-сообществ*$^{СУЩ,неод,ср\ мн,рд}$ *и*^{СОЮЗ} *др* ^{НЕИЗВ} (сокращение «*др*.» не распознается программой);

5) неверно определена частеречная принадлежность местоимений:

Такая$^{ПРИЛ,мест-п\ жр,ед,им}$ *оценка*$^{СУЩ,неод,жр\ ед,им}$ *должна*$^{КР\_ПРИЛ\ жр,ед}$ *появиться*^{ИНФ,сов,неперех} *в*^{ПР} *результате*$^{СУЩ,неод,мр\ ед,пр}$ *проведения*$^{СУЩ,неод,ср\ ед,рд}$ *соответствующего*$^{ПРИЛ\ мр,ед,рд}$ *маркетингового*$^{ПРИЛ\ мр,ед,рд}$ *либо*^{СОЮЗ} *иного*$^{ПРИЛ,мест-п\ ср,ед,рд}$ *исследования*$^{СУЩ,неод,ср\ ед,рд}$ (местоимение «*такая*» определяется как прилагательное);

6) составные части речи рассматриваются как отдельные единицы:

*Обновление*СУЩ,неод,ср ед,им *информации*СУЩ,неод,жр ед,рд *в*ПР *печатных*ПРИЛ,кач мн,рд *СМИ*СУЩ,неод,хр,pl,0 мн,им *не*ЧАСТ *может*ГЛ,несов,неперех ед,3л,наст,изъяв *успеть*ИНФ,сов,неперех *ни*ЧАСТ *за*ПР *телевидением*СУЩ,неод,ср ед,тв, *ни*ЧАСТ *тем*СОЮЗ *более*Н *за*ПР *Интернетом*СУЩ,неод,мр ед,тв; (составной союз «*тем более*» распознан как две единицы).

7) грамматические омонимы распознаются некорректно:

Дуговая ПРИЛ жр,ед,им *сварка* СУЩ,неод,жр ед,им *неплавящимся* ДЕЕПРИЧ мр,ед,тв *электродом* СУЩ,неод,мр ед,тв *в* ПР *защитной* ПРИЛ,кач жр,ед,тв *атмосфере* СУЩ,неод,жр ед,пр *инертного* ПРИЛ,кач мр,ед,рд *газа* СУЩ,неод,жр,sg,гео ед,им __ ЗПР *метод* СУЩ,неод,мр ед,им *дуговой* ПРИЛ мр,ед,им *сварки* СУЩ,неод,жр ед,рд, ЗПР *который* ПРИЛ,мест-п,субст?,Анаф мр,ед,им *используется* ГЛ,несов,неперех ед,3л,наст,изъяв *для* ПР *сварки* СУЩ,неод,жр ед,рд *алюминия* СУЩ,неод,мр ед,рд, ЗПР *магния* СУЩ,неод,мр ед,рд *и* СОЮЗ *их* МС,3л,Анаф мн,рд *сплавов* СУЩ,неод,мр мн,рд, ЗПР *нержавеющей* ПРИЛ жр,ед,рд *стали* ГЛ,сов,неперех мн,прош,изъяв, ЗПР *никеля* СУЩ,неод,мр ед,рд, ЗПР *меди* СУЩ,неод,жр ед,рд , ЗПР *бронзы* СУЩ,неод,жр ед,рд, ЗПР *титана* СУЩ,неод,мр ед,рд, ЗПР *циркония*СУЩ,неод,мр ед,рд *и* СОЮЗ *других* ПРИЛ,мест-п,субст? мн,рд *неферромагнитных* ПРИЛ мн,рд *металлов* СУЩ,неод,мр мн,рд. ЗПР (имя существительное «стали» распознано как глагол).

Данные особенности работы морфологичеких анализаторов должны быть учтены при разработке моделей и методов разметки многокомпонентных терминов в параллельном корпусе научно-технических текстов.

Таким образом, наиболее сложным явлением в процессе автоматической разметки терминов в корпусе научно-технических текстов представляют собой многокомпонентные термины – терминологические словосочетания, образованные лексическим и синтаксическим способами, то есть словосочетания, образованные по определенным моделям. При создании системы автоматической разметки корпуса научно-технических текстов прежде всего необходимо установить все возможные структурные модели англо- и русскоязычных многокомпонентных терминов, а также возможности и потенциальные ошибки морфологических анализаторов.

Для выявления именно устойчивых терминологических сочетаний и

рассмотрения уже устоявшихся структурных моделей с последующей возможностью их разметки в корпусах научно-технических текстов была выбрана терминосистема предметной области «Сварка». Терминология данной области зародилась еще в древности, но свое основное развитие получила в XX в., в результате чего она является достаточно хорошо исследованной. Кроме того, почти на всем протяжении XX в. проводились работы по упорядочению и унификации сварочной терминологии. Первой попыткой ее обобщения, стал «Бюллетень комиссии технической терминологии» под редакцией академика Чаплыгина С.А. и Лотте Д.С., изданный в 1937 г. [222]. Позднее Д.С. Лотте вводит понятие правильноориентирующие *термины*, т.е. те, мотивировка которых не вступает в противоречие с характером обозначаемого ими понятия [223].

В 60-е и 70-е гг. начался следующий этап в работе по унификации сварочной терминологии. Появляется «Англо-русский словарь по сварочному производству» Золотых В.Т. [224] и «Словарь-справочник по сварке» Кулика Т.В. [225] а в 1984 г. была проведена унификация имевшихся на то время терминов, результатом чего стал ГОСТ 2601-84 «Сварка металлов. Термины и определения основных понятий» [226]. В этом и последующих стандартах закреплялись наиболее приемлемые по семантической наполненности и структуре термины, и отмечались недопустимые для употребления термины. С учетом вышесказанного можно считать, что терминологи предметной области «Сварка» является устоявшейся и может быть использована при выделении структурных моделей терминологических единиц.

В таблице 3.1 приведены структурные модели русскоязычных терминологических единиц на примере предметной области «Сварка», их описание и примеры. Терминологические элементы классифицированы по количеству элементов, входящих в состав терминологического словосочетания. В приведенной таблице служебные слова также относятся к элементам терминологических словосочетаний и считаются элементами термина.

Таблица 3.1 Структурные модели терминологических единиц предметной области «Сварка»

	Кол-во компонентов	Модель	Примеры
1	1	Имя существительное	*сварка*
2	2	Имя прилагательное + имя существительное	*вибродуговая сварка, вибродуговая сварка*
3	2	Причастие + имя существительное	*защищенная сварка, механизированная сварка*
4	2	Имя существительное + имя существительное в творительном падеже	*сварка взрывом, сварка трением, сварка лазером*
5	2	Имя существительное + имя существительное в родительном падеже	*фронт волн, волна колебаний*
6	3	Имя прилагательное + имя прилагательное + имя существительное (ядерный элемент)	*автоматическая точечная сварка, конденсаторная ударная сварка*
7	3	Имя прилагательное + имя существительное (ядерный элемент) + имя существительное в творительном падеже	*газовая сварка плавлением, холодная сварка давлением*
8	3	Имя прилагательное + имя существительное (ядерный элемент) + наречие	*кузнечная сварка ввилку, кузнечная сварка внахлест*
9	3	Имя прилагательное + имя существительное (ядерный элемент) + имя существительное в винительном падеже	*высокочастотная сварка пластмасс*
10	3	Имя прилагательное + имя существительное (ядерный элемент) + имя существительное в родительном падеже	*направленное движение плазмы*
11	4	Имя прилагательное + имя существительное (ядерный элемент) + имя существительное в предложном падеже	*диффузионная сварка в вакууме, кузнечная сварка в штампе*
12	4	Имя прилагательное + имя существительное (ядерный элемент) + имя прилагательное + имя существительное в творительном падеже	*аргонодуговая сварка вольфрамовым электродом,*
13	5	Имя прилагательное + имя существительное (ядерный элемент) + имя прилагательное + имя существительное в творительном падеже (с предлогом «с»)	*дуговая сварка с магнитным флюсом, газоэлектрическая сварка с магнитным флюсом*
14	5	Имя прилагательное + имя существительное (ядерный элемент) + имя прилагательное + имя существительное в предложном падеже	*газопрессовая сварка в пластическом состоянии*
15	5	Имя прилагательное + имя существительное (ядерный элемент) + имя существительное в предложном падеже + имя существительное в родительном падеже	*дуговая сварка в среде гелия*
16	5	Имя прилагательное + имя существительное (ядерный элемент) + имя существительное в творительном падеже (с предлогом) + имя	*контактная сварка с накоплением энергии*

		существительное в родительном падеже	
17	5	Имя прилагательное + имя прилагательное + существительное (ядерный элемент) + существительное в творительном падеже (с предлогом)	*автоматическая дуговая сварка под флюсом*
18	5	Имя прилагательное + имя существительное (ядерный элемент) + имя существительное в творительном падеже + имя существительное в творительном падеже (с предлогом)	*стыковая сварка оплавлением с осадкой*
19	5	Имя прилагательное + имя прилагательное + имя существительное (ядерный элемент) + (прилагательное + имя существительное в творительном падеже)	*автоматическая дуговая сварка покрытым электродом*
20	5	Имя прилагательное + имя существительное (ядерный элемент) + (имя прилагательное + имя прилагательное + имя существительное в творительном падеже)	*дуговая сварка газообразующей электродной проволокой*
21	5	Имя прилагательное + имя существительное (ядерный элемент) + (имя существительное в творительном падеже + имя существительное в родительном падеже + имя существительное в родительном падеже)	*дуговая сварка методом опирания электрода*
22	6	Имя прилагательное + имя существительное (ядерный элемент) + (имя существительное в творительном падеже (с предлогом) + имя прилагательное + имя существительное в творительном падеже)	*газопрессовая сварка с нагревом ацетиленокислородным пламенем*
23	6	Имя прилагательное + имя существительное (ядерный элемент) + (имя существительное в творительном падеже (с предлогом) + имя прилагательное + имя существительное в родительном падеже)	*газовая сварка с расплавлением свариваемых кромок*
24	6	Имя прилагательное + имя существительное (ядерный элемент) + имя прилагательное + предлог + имя существительное в творительном падеже + имя существительное в родительном падеже	*дуговая сварка с поперечными колебаниями электрода, точечная сварка с интенсивным охлаждением электродов*

В ходе исследования выявлено 24 модели терминологических словосочетаний. В литературных источниках встречаются термины, состоящие из 7-13 компонентов, однако, в силу того что они, с одной стороны, крайне редко встречаются в текстах и могут иметь множество структурных моделей они не включены в структурные модели данного исследования, а с другой стороны, со временем их структура трансформируется в более компактные формы.

В таблице 3.2 приведены структурные модели англоязычных

многокомпонентных терминов, их описание и иллюстративные примеры.

Таблица 3.2. Структурные модели англоязычных терминов для разметки научно-технических текстов в параллельном корпусе

	Кол-во компонентов	Модель	Примеры
1	1	Имя существительное	*welding*
2	2	Имя существительное + имя существительное (ядерный элемент)	*arc welding, blacksmith welding, braze welding*
3	2	Имя прилагательное + имя существительное (ядерный элемент)	*chemical welding, cold welding, dielectric welding*
4	2	Причастие + имя существительное (ядерный элемент)	*powdered welding, flux-cored welding, pressure-controlled welding*
5	3	Имя существительное (ядерный элемент) + имя существительное с предлогом *in*	*welding in air, welding in space*
6	3	Имя существительное (ядерный элемент) + имя существительное с предлогом *with*	*welding-on with pressure*
7	3	Имя прилагательное + имя существительное + имя существительное (ядерный элемент)	*atomic hydrogen welding, cold pressure welding*
8	3	Причастие + существительное + существительное (ядерный элемент)	*closed joint welding, hammered resistance welding*
9	3	Имя прилагательное + имя прилагательное + имя существительное (ядерный элемент)	*electromagnetic percussive welding, electrostatic percussive welding*
10	3	Имя существительное + имя существительное + имя существительное (ядерный элемент)	*carbon arc welding, laser hybrid welding*
11	4	Имя существительное + имя существительное (ядерный элемент) + имя существительное с предлогом in	*diffusion welding in vacuum*
12	4	Имя существительное (ядерный элемент) + имя прилагательное + имя существительное с предлогом under	*welding under controlled atmosphere*
13	4	Имя существительное (ядерный элемент) + имя прилагательное + имя существительное с предлогом with	*welding with independent arc*
14	4	Причастие+ имя существительное + имя существительное + имя существительное (ядерный элемент)	*controlled rate arc welding*
15	4	Причастие+ имя существительное + имя существительное + имя существительное (ядерный элемент)	*stored energy resistance welding*
16	4	Имя существительное + имя прилагательное + имя существительное + имя существительное (ядерный элемент)	*robot remote laser welding*

17	4	Имя существительное + имя существительное + имя существительное + имя существительное (ядерный элемент)	*gas tungsten arc welding*
18	4	Имя существительное + имя прилагательное + имя существительное + имя существительное (ядерный элемент)	*metal active gas welding*
19	5	Имя прилагательное + имя существительное + имя существительное + имя существительное + имя существительное (ядерный элемент)	*consumable electrode inert arc welding*
20	5	Имя существительное + причастие + имя существительное + имя существительное + имя существительное (ядерный элемент)	*gas shielded metal arc welding*
21	6	Причастие + имя прилагательное + существительное + имя существительное + имя существительное + имя существительное (ядерный элемент)	*shielded inert gas metal arc welding*

В ходе исследования выделена 21 структурная модель англоязычных научно-технических терминов, которые служат основой для метода разметки терминологических единиц в параллельном корпусе научно-технических текстов.

Выделение большого числа моделей многокомпонентных терминов позволяется учитывать некоторые ошибки морфологических анализаторов, например, в моделях многокомпонентных терминов для имен прилагательных значимым тегом является только часть речи, а такую категорию как падеж часто морфологические анализаторы определяют неверно. Неверное определение частичной принадлежности причастий, деепричастий и местоимений разрешается за счет того, что в моделях многокомпонентных терминологических словосочетаний указаны как модели со всеми указанными частями речи, что означает, что термин все равно попадет в термины-кандидаты и будет рассмотрен системой с учетом других данных.

2.2. Структурные модели англо- и русскоязычных номенклатурных наименований

Развивающиеся терминологии компьютерных наук, авиации и космонавтики, нанотехнологий и других предметных областей используют

новые способы терминообразования, таким образом расширяя структурные модели терминов. К таким моделям словообразования можно отнести номенклатурные наименования, имеющие отличительную структуру от других пластов специальной лексики [182].

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, ракета-носитель «Союз-2.1Б» с разгонным блоком «Фрегат».*

Номенклатурные наименования обозначают единичные вещи, предметное значение в них преобладает над понятийным – это знаки промышленных товаров, наименования механизмов и машин, видов животных, сортов растений, медикаментов и т.п., например, *пистолет «Браунинг», самолет «Боинг 747», насадки Kärcher 310, витамин В* и т.п.

Каждый номенклатурный список включает в себя совокупность однородных предметов, обладающих общими существенными признаками и различающихся второстепенными. Так, самые разнообразные специальные единицы в области космонавтики: *спутник Бион М-1, миссия Луна-25, эксперимент “Матрешка-Р”, космический корабль Космос-321* являются одновременно и номенклатурными наименованиями, и терминами соответствующей области знания. В их составе синтаксически главные единицы: *Бион М-1, Луна-25, “Матрешка-Р”, Космос-321* являются номенами (номенклатурными маркерами, языковыми «этикетками») и выделяют из родового понятия данное частное понятие, содержание которого фиксируется в специальных описаниях, толкованиях, дефинициях и т.п.

Существует два основных способа создания номенклатурных знаков – с помощью исконной лексики и из заимствованных элементов. Во многих научных номенклатурных перечнях имеет место интернациональная греко-латинская морфемика, своеобразный искусственный, формально единообразный язык для общения и полного взаимопонимания специалистов разных стран [227].

В таблицах 3.3 и 3.4 приведены структурные модели англо- и русскоязычных номенклатурных наименований на примере предметной области «Космонавтика», их описание и примеры.

Таблица 3.3. Структурные модели англоязычных номенклатурных наименований для разметки научно-технических текстов в параллельном корпусе

	Кол-во компонентов	Модель	Примеры
1	2	аббревиатура + слово	*SPH method*
2	2	аббревиатура + БЧП	*SL Soyuz-2.1a*
3	2	термин + слово	*Spacecraft Pogo, Spacecraft Explorer, Spacecraft Magsat*
4		слово+ БЧП	*Soyuz-2, MAC-1*
5	2	БЧП+термин	*SGP4 orbital motion model, AIST-2D Small Spacecraft*
6	2	слово + термин	*Vostochny Cosmodrome*
7	2	слово + БЧП	*Soyuz-2.1a*
8	2	аббревиатура + термин	*CHAMP mission*
9	3	термин +слово+ БЧП	*Domestic Spacecraft Cosmos 26*
10	3	аббревиатура + слово + БЧП	*BS Bion M-1*

Номенклатурные наименования классифицированы по количеству элементов, входящих в состав терминологического словосочетания, а также языковой принадлежности элементов номена. 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.

Таблица 3.4. Структурные модели русскоязычных номенклатурных наименований для разметки научно-технических текстов в параллельном корпусе

	Кол-во компонентов	Модель	Примеры
1	2	аббревиатура + слово	*КА Аполлон*
2	2	аббревиатура + БЧП	*КК Orion* *КА Pogo*
3	2	аббревиатура + БЧП	*АП НР-23* *АИ АЗТ – 11* *ЦЧ Ч3-35*
4	2	термин + слово	*Микроробот Микроид*
5	2	термин + слово на латинице	*Алгоритм PredGuid* *Возвращаемый аппарат Orion*
6	2	Термин + БЧП	*Трасса JXN-MIK* *Трасса GQD-MIK* *Трасса NAA-MIK*
7	3	аббревиатура, слово, БЧП	*МКА Аист-2Д* *РН Союз-2* *АИ Цейсс-600* *АС Луна-24*
8	3	аббревиатура, термин, БЧП	*СДВ радиостанция JXN* *СДВ радиостанция GQD* *СДВ радиостанция NAA*
9	3	аббревиатура, слово, БЧП	*РН Протон-М* *КК Союз-19* *КА Аполлон-11* *АС Луна-24*
10	3	аббревиатура, термин, слово	*АДЗ Земли Аврора*
11	3	термин + слово + БЧП	*Спутник Фотон М-4* *Спутник Бион М-1* *Спутник Бион М-2*

Сложившаяся ситуация говорит о том, что русский язык оказывается более емким и лаконичным в аспекте образования специальной лексики, в то время как английский язык прибегает к использованию большего числа специальных лексических единиц для передачи одного и того же объема информации. Также это свидетельствует о неравномерном развитии специальной лексики космонавтики в России и англоязычных странах.

На данном этапе очень важна систематизация и группирование синтаксически главных единиц терминов по определённым критериям, для последующего добавления их в систему данных. Наличие базы данных

структурных моделей номенклатурных наименований позволит систематизированным поисковым системам быстрее извлекать многокомпонентные номенклатурные наименования из научно-технических текстов на разных языках.

Такая классификация терминов позволит решать ряд прикладных задач, например, бывает так, что номенклатурные наименования могут совпадать с аббревиатурами, в таком случае можно использовать классификацию номенклатурных названий, тем самым избежать неточного извлечения номенклатуры.

3.3. Методы разметки и выравнивания специальной лексики на основе структурных моделей англо- и русскоязычных терминов

Предлагаемый метод к автоматическому извлечению русскоязычных многокомпонентных терминов на основе структурных моделей англо- и русскоязычных терминологических словосочетаний состоит из пяти основных этапов, представленных на рис. 3.2.

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 предложения, то есть приписываем каждому слову его морфологические характеристики: часть речи, род, число, падеж, например:

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. В состав терминологических словосочетаний не входят глаголы, союзы, местоимения, частицы, знаки препинания, а также некоторые сочетания частей речи как наречие + предлог и прочее.

Рис 3.2. Этапы метода извлечения русскоязычных многокомпонентных 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 «стоп-слов»,

которые указаны в специальной зоне словаря, и удаляем их. Под стоп-словами понимаем слова, которые образуют широко используемые коллокации с терминами, но в совокупности, не являющиеся терминами по сути, например, *современная* химия, *рассматриваемый* метод синтеза о-гликозидов.

4. Полученные цепочки слов соотносим с шаблонами терминологических словосочетаний, имеющихся в базе структурных моделей терминов.

наука $_{СУЩ,неод,жр\ ед,им}$ – имя существительное - принадлежит к моделям терминов

химия $_{неод,жр\ ед,им}$ *природных* $_{ПРИЛ,кач\ мн,рд}$ *соединений* $_{СУЩ,неод,ср\ мн,рд}$ – имя существительное* + имя прилагательное + имя существительное - принадлежит к моделям терминов

органической $_{ПРИЛ\ жр,ед,ТВ}$ *химией* $_{СУЩ,неод,жр\ ед,ТВ}$. - имя прилагательное + имя существительное - принадлежит к моделям терминов

5. Полученные термины-кандидаты проверяем по словарю корпуса. Если термин-кандидат есть в словаре, то извлекаем его как термин. Если полученный термин кандидат отсутствует в словаре, то отправляем терминологу для обработки данного термина-кандидата вручную.

Исследования терминологических словосочетаний в испанском, английском, немецком и других языках свидетельствуют о наличии структур многокомпонентных терминов, образованных по определенным моделям, отражающим особенности языка, на котором они образованы. Для использования описанного метода извлечения многокомпонентных терминологических словосочетаний из научно-технических текстов на разных языках необходимо только наличие базы данных структурных моделей терминологических словосочетаний обрабатываемого языка.

В качестве примера для демонстрации метода по извлечению англоязычных терминов-кандидатов, взят небольшой отрывок из предметной области «Авиация и космонавтика», на примере следующего фрагмента текста:

Transfers in the central Newtonian field are considered under the assumption

that low thrust that is constant in magnitude is zeroed when spacecraft with solar batteries enters the Earth's shadow.

Отбор терминов-кандидатов также как и для русского языка состоит из нескольких этапов. На первом этапе производится морфологический разбор текста, то есть каждому слову приписываются его морфологические характеристики:

Transfers$^{n,\ v}$ *in*$^{prep,\ adv,\ v,\ n,\ adj}$ *the*$^{art,\ adv}$ *central*^{adj} *Newtonian*$^{adj,\ n}$ *field*$^{n,\ v}$ *are*$^{v,\ n}$ *considered*^{v} *under*$^{prep,\ adv,\ adj}$ *the*$^{art,\ adv}$ *assumption*^{n} *that*$^{conj,\ det,\ pron,\ adv,\ n}$ *low*$^{adj,\ n,\ adv,\ v}$ *thrust*$^{n,\ v}$ *that*$^{conj,\ det,\ pron,\ adv,\ n}$ *is*$^{v,\ n}$ *constant*$^{adj,\ n}$ *in*$^{prep,\ v,\ adv,\ n,\ adj}$ *magnitude*^{n} *is*$^{v,\ n}$ *zeroed*^{v} *when*$^{adv,\ conj,\ pron,\ n,\ interj}$ *spacecraft*^{n} *with*$^{prep,\ adv,\ n}$ *solar*$^{adj,\ n}$ *batteries*^{n} *enters*$^{n,\ v}$ *the*$^{art,\ adv}$ *Earth's*$^{n,\ v}$ *shadow*$^{n,\ v,\ adj}$.

После морфологического анализа необходимо убрать часть слов, которые не входят в терминологическую систему (глаголы, знаки препинания и т. д.), чтобы выявить допустимые терминологические словосочетания:

Transfers$^{n,\ v}$ *in*$^{prep,\ adv,\ v,\ n,\ adj}$ ~~*the*~~$^{~~art,\ adv~~}$ *central*^{adj} *Newtonian*$^{adj,\ n}$ *field*$^{n,\ v}$ ~~*are*~~$^{~~v,\ n~~}$ ~~*considered*~~$^{~~v~~}$*-under*$^{prep,\ adv,\ adj}$ ~~*the*~~$^{~~art,\ adv~~}$ *assumption*^{n} ~~*that*~~$^{~~conj,\ det,\ pron,\ adv,\ n~~}$ *low*$^{adj,\ n,\ adv,\ v}$ *thrust*$^{n,\ v}$ ~~*that*~~$^{~~conj,\ det,\ pron,\ adv,\ n~~}$ ~~*is*~~$^{~~v,\ n~~}$ *constant*$^{adj,\ n}$ *in*$^{prep,\ v,\ adv,\ n,\ adj}$ *magnitude*^{n} ~~*is*~~$^{~~v,\ n~~}$ ~~*zeroed*~~$^{~~v~~}$ ~~*when*~~$^{~~adv,\ conj,\ pron,\ n,\ interj~~}$ *spacecraft*^{n} *with*$^{prep,\ adv,\ n}$ *solar*$^{adj,\ n}$ *batteries*^{n} ~~*enters*~~$^{~~n,\ v~~}$ ~~*the*~~$^{~~art,\ adv~~}$ *Earth's*$^{n,\ v}$ *shadow*$^{n,\ v,\ adj}$.

На третьем этапе необходимо сравнить цепочки слов допустимых терминологических словосочетаний с терминологическими моделями, то есть извлечь последовательности слов, которые соответствуют морфосинтаксическим шаблонам однословных терминов и терминологических словосочетаний. Таким образом в рассматриваемом примере получен следующий список терминов-кандидатов:

1. *Transfers in (the) central Newtonian field* (Имя существительное* + предлог + имя прилагательное + имя прилагательное + имя существительное);

2. *Central Newtonian field* (Имя прилагательное + имя прилагательное + имя существительное*);

3. *Low thrust* (Имя прилагательное + имя существительное*);

4. *Constant in magnitude* (Имя существительное* + предлог + имя существительное);

5. *Spacecraft with solar batteries* (Имя существительное* + предлог + имя прилагательное + имя прилагательное + имя существительное);

6. *Solar batteries* (Имя прилагательное + имя существительное*);

7. *Earth's shadow* (Имя существительное + имя существительное*);

На третьем этапе выполняется проверка ряда лингвистических условий, например, терминологическое словосочетание как в английском, так и русском языках не может начинаться с предлога или состоять из одного элемента кроме существительного. В первом случае предлог исключается из словосочетания, а во втором слово удаляется из списка терминов-кандидатов. В английском языке также опускаются артикли.

На четвёртом этапе, происходит оценка выбранных слов и словосочетаний. Сначала каждое словосочетание проверяется по терминологическому словарю. Если такое словосочетание отсутствует в словаре, то оставшиеся словосочетания проходят проверку на стоп-слова. Под стоп-словами в разрабатываемой системе понимаются слова, которые сочетаются с терминами таким образом, что рассматриваемая синтаксическая конструкция совпадает со структурной моделью многокомпонентного термина, например, имя прилагательное + имя существительное + имя существительное: *artificial information technology* (термин) и *new information technology* (общеупотребительное слово + термин). В таком случае необходимо сформировать список общеупотребительных слов с оценочной или внепредметной семантикой (*new, modern, developed, analyzed и т.п.*), чтобы не учитывать их при извлечении терминов.

К наиболее часто используемым моделям терминов в русском языке можно отнести двух и трехкомпонентные словосочетания, образованные по следующим моделям: существительное + существительное в родительном падеже, прилагательное + существительное, прилагательное + прилагательное + существительное, прилагательное + существительное + существительное в родительном падеже и др. Для английского языка наиболее продуктивными являются модели: имя существительное + имя существительное*, имя прилагательное + имя существительное*, имя существительное* + предлог + имя существительное; имя существительное + имя существительное + имя существительное*.

3.4. Метод разметки англо- и русскоязычных номенклатурных наименований в научно-технических текстах

Метод разметки номенклатурных наименований, так же, как и методы разметки многокомпонентных терминов основан на структурных моделях номенклатурных наименований в английском и русском языках.

В связи с тем, что в состав номенклатурного наименования входят термин и номен, разметку номенклатурных наименований следует проводить после разметки многокомпонентных терминов. При таком подходе, во-первых, термин, состоящий из двух и более элементов, будет обозначен как одна единица, а во-вторых, он будет являться ядром номенклатурного наименования и, в свою очередь, индикатором, что рядом может стоять номен.

Основные этапы метода разметки номенклатурных наименований в параллельном корпусе научно-технических текстов представлены на рис. 3.3.

Рассмотрим пример разметки номенклатурных наименований в русскоязычных научно-технических текстах.

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.

Рис 3.3. Этапы метода разметки номенклатурных наименований в научно-технических текстах

Проводим терминологическую разметку с использованием метода, предложенного в предыдущем пункте, и получаем размеченный текст, представленный ниже:

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 к результатам морфологической разметки добавляем результаты терминологической разметки многокомпонентных терминов. Ядерный элемент многокомпонентного термина обозначен *.

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 предлога.

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 + буквы/числа.

Рассмотрим пример разметки номенклатурных наименований в англоязычных научно-технических текстах, на примере следующего предложения.

Thus, at wavelengths longer than 4 cm, the emission of radio sources within the entire solar disk is detected in RATAN-600 scans.

Проводим морфологический анализ и получаем размеченный текст, представленный ниже:

Thus adv*., at* prep *wavelengths* n *longer* adj *than* conj *4* prep*. cm, the* art. *Emission* n *of* prep *radio* n *sources* n *within* prep *the entire* adj *solar* n *disk* n *is* v *detected* v *in* prep *RATAN-600 scans* n*.*

На следующем этапе к результатам морфологической разметки добавляем результаты терминологической разметки многокомпонентных терминов. Ядерный элемент многокомпонентного термина обозначен *.

Thus adv*., at* prep ***[wavelengths*** n] *longer* adj *than* conj *4* prep*. cm, the* art. ***[emission*** n ***of*** prep ***radio*** n ***sources*** n**]* *within* prep *the entire* adj ***[solar*** n ***disk*** n**]* *is* v *detected* v *in* prep *RATAN-600*? ***[scans*** n**].*

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

На следующем этапе полученный кандидат-номенклатурное наименование проверяем по моделям номенклатурных наименований: *RATAN-600[?] [scans[n]*]* совпадает с моделью «*БПЧ+термин*».

В связи с тем, что одним из требований, предъявляемых к параллельному корпусу, является филологическая корректность, то получившееся номенклатурное наименование проверяем по терминологической базе данных корпуса. Если оно отсутствует в словаре, то поступает на обработку к терминологу, который или добавит его в словарь или в список неверно распознанных номенов для улучшения качества работы системы.

Для оценки эффективности предложенного метода использовалась экспертная оценка, проведенная филологами (Таблица 3.5). Всего проанализировано 20 текстов научно-технических статей по космонавтике, опубликованных в журнале «Космические исследования» в 2018-2019 гг. Оценка эффективности метода извлечения терминов проведена путем сравнения с методом [171] извлечения терминов на основе синтаксических шаблонов.

Таблица 3.5. Оценка качества метода извлечения терминов из научно-технических текстов, в %

Количество компонен-тов термина	Кол-во уникаль-ных тер-минов	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

Прочерк в таблице 3.5 означает отсутствие синтаксических шаблонов для

терминов такой структуры. Учтены только уникальные термины независимо от их частотности вхождения в тексты научно-технических статей.

Повышение эффективности извлечения терминов из научно-технических текстов происходит за счет добавления моделей терминов с правыми определениями, а также использования дополнительных грамматических характеристик многокомпонентных терминов, заложенных в обновленные синтаксические шаблоны. Стоит отметить, что при оценке качества извлечения терминов из текстов наиболее спорными стали одно- и двухкомпонентные термины, у которых есть как терминологическое значение, так и общеупотребительное, например лексическая единица «*игрок*» является термином из теории игр и общеупотребительным словом. При этом, такое явление не наблюдалось при извлечении терминов из трех и более компонентов.

Таким образом, результатом реализации методов разметки многокомпонентных терминов и номенклатурных наименований в параллельном корпусе научно-технических текстов станет «слияние» многокомпонентных терминологических единиц и номенклатурных наименований в одну лексическую единицу, что позволит при дальнейшей обработке текстов учитывать значительно меньший объем лингвистических единиц.

3.5. Метод выравнивания многокомпонентных терминов и номенклатурных наименований в параллельных научно-технических текстах

Выравнивание терминологических единиц разной формальной структуры в параллельных текстах затруднено рядом факторов, а именно:

- разная длина и формальная структура терминологических единиц в русском и английском языках;

- разница в синтаксической структуре англо- и русскоязычных предложений;

- использование переводческих трансформаций.

В таблице 3.6 представлены примеры структурных переводческих трансформаций англоязычных терминов, образованных по модели noun+noun, а также структурные переводческие трансформации их эквивалентов в русском языке. Знаком * отмечены ядерные элементы терминов. Ядерный элемент является «главным» элементом терминологической единицы, который вступает с словоизменительную парадигму при сочетании с другими единицами в предложении [228]. Остальные элементы термина, кроме имен прилагательных и причастий, стоящих перед ядерным элементом, остаются неизменными. В терминологических единицах ядерные элементы имеют тенденцию совпадать в переводных эквивалентах, т.е. ядерный элемент в языке оригинала будет переводным эквивалентом в языке перевода, кроме тех случаев, где в языке оригинала одно слово, а на языке перевода два и более.

Таблица 3.6. Структурные переводческие трансформации модели «имя существительное + имя существительное»

Примеры англоязычных терминов, образованных по модели Noun+noun*	Виды структурных переводческих трансформаций эквивалентных русскоязычных терминов	
	Модель	Пример
stud welding*	Сущ*+сущ в родительном падеже	Приварка* шпилек
resistance welding*	Прил+прил+сущ*	электрическая контактная сварка*
upset welding*	Прил+сущ*+сущ в творительном падеже	стыковая сварка* сопротивлением
pressure welding*	Сущ*+сущ в творительном падеже	сварка* давлением
gravity welding*	Сущ*+прил+сущ в творительном падеже	сварка* наклонным электродом
dielectric welding*	Прил+сущ*+сущ в родительном падеже	высокочастотная сварка* пластмасс

Стоит также отметить, что формальная структура терминов в русском языке отличается не только по структуре, но и количеству компонентов. Таким образом, при решении задачи выравнивания многокомпонентных терминов

могут быть использованы разные варианты выравнивания, примеры которых представленные в таблице 3.7.

Таблица 3.7. Вариантное соответствие количества единиц в англоязычном термине и его переводном эквиваленте

	Соответствие	Пример
1	1 к 1	*welding - сварка*
2	1 к 2	*welding – сварочные работы*
3	2 к 1	*automatic welding - автосварка*
4	2 к 2	*chemical welding – химическая сварка*
5	2 к 3	*frame welding – контурная сварка пластмасс*
6	2 к 4	*die welding - кузнечная сварка в штампе*
7	2 к 5	*печная сварка - pressure welding with furnace heating*
8	2 к 13	*squirt welding - полуавтоматическая дуговая сварка под флюсом с подачей флюса из бункера, укрепленного на держателе*
9	3 к 6	*gas braze welding - газовая сварка с применением твёрдого припоя*
10	4 к 2	*condenser energy-storage welding - конденсаторная сварка*
11	4 к 4	*flux cored arc welding - дуговая сварка порошковой проволокой*

Другим аспектом, который необходимо учитывать при создании метода выравнивания многокомпонентных терминов, являются различия в актуальном членении предложения в русском и английском языках. Предложение состоит из двух компонентов: темы, т.е. того, что является в данной ситуации известным или, по крайней мере, может быть легко понято и из чего исходит говорящий, и ремы, то есть то, что говорящий сообщает об исходной точке высказывания» [229]. Соотношение компонентов актуального членения предложения типизировано, представлено в виде моделей, линейно-динамических структур, порядка слов и т.д. Роль темы и ремы могут исполнять любые члены предложения [230]. В русском языке актуальное членение предложения выражено чаще всего в определенном порядке слов во взаимодействии с интонацией, а в английском языке способы выражения более разнообразны, например, тема-тематические отношения могут быть выражены через

употребление определённых и неопределенных артиклей, инверсии в предложении. При этом прямой порядок слов присущ всем синтаксическим конструкциям английского языка.

В качестве примера (Рис. 3.4) рассмотрим предложение «*Группу документов, по которой осуществляется поиск, мы будем называть коллекцией*». Тема в данном случае выражена первой частью предложения *группу документов, по которой осуществляется поиск*, а рема - *будем называть коллекцией.* В переводе это предложение будет иметь следующий вид: «*We will refer to the group of documents over which we perform retrieval as a collection*», а тема и рема выражены через употребление артиклей.

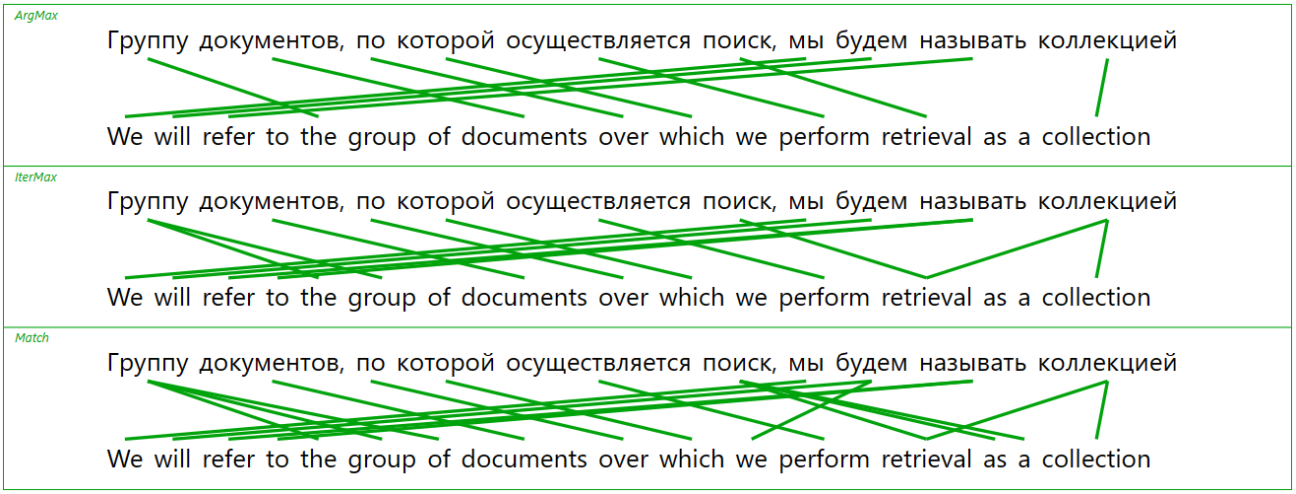


Рис. 3.4. Различия в порядке слов в русском и английском языках

Третьим фактором, который необходимо учитывать при выравнивании терминов, это наличие переводческих трансформаций в параллельных текстах. Под переводческими трансформациями понимают лексические, грамматические и лексико-грамматические преобразования, с помощью которых можно осуществить переход от единиц оригинала к единицам перевода в указанном смысле [231]. Широкое использование переводческих трансформаций при переводе научно-технических текстов в паре английского и русского языков приводит к тому, что термин может быть добавлен или опущен, его значение передано другими лексико-грамматическими средствами, например: в

предложении «***Интенсивно исследуется*** *новый класс оптических волокон — фотонно-кристаллические волокна, которые расширяют возможности* ***управления*** *дисперсионными характеристиками и нелинейными свойствами волокон»* и его переводном эквиваленте «***A lot of research is being carried out*** *regarding a new class of optical fibres, photonic crystal fibres, which extend the ability to* ***manage*** *the dispersion characteristics and nonlinear properties of fibres»* русскоязычный глагол *исследуется* переведен англоязычным термином *research*, а русскоязычный термин *управление* – англоязычным глаголом *manage*. Таким образом в процессе выравнивания текстов может наблюдаться неравномерное количество терминологических единиц в текстах оригинала и перевода.

При использовании пакета SimAlign для выравнивания многокомпонентных терминов возможно с использованием нескольких сценариев:

1. Праллельные тексты можно сначала выровнять по словам, а затем на основе синтаксических шаблонов англо- и русскоязычных терминов извлечь выровненные переводные эквиваленты.

2. Сформировать «мешок терминов» на основе синтаксических шаблонов англо- и русскоязычных терминов, а затем выровнять списки терминов.

Первый сценарий широко используется на практике при выравнивании терминов из параллельных текстов на английском, немецком и французском языках с использованием переводных словарей [232], английском и словенском языках на основе машинного обучения [233].

Использование подходов на основе переводных словарей для близкородственных языков показывает высокие результаты за счет грамматической и лексической близости языков, принадлежащих к одной языковой семье. В свою очередь подходы на основе машинного обучения требуют наличия параллельных корпусов, однако таких ресурсы для языковой пары английский-русский в настоящее практически отсутствуют.

В настоящее время также разработаны пакеты, которые выравниванию слова в параллельных предложениях. Для каждого элемента термина на одном языке в

другом языке есть множество переводных эквивалентов, что приводит к тому, что нейронная сеть выравнивает предложения, не являющиеся переводными эквивалентами, как представлено на рис. 3.5 для предложений:

«*Рассмотрено компоновочное решение — комбинированная конструкция центробежного сепаратора для подготовки природного газа с долей попутного нефтяного газа»* и *«The paper presents numerical computation results for the separation simulation, as well as data obtained during actual separator operation for different heat and pressure values»*.

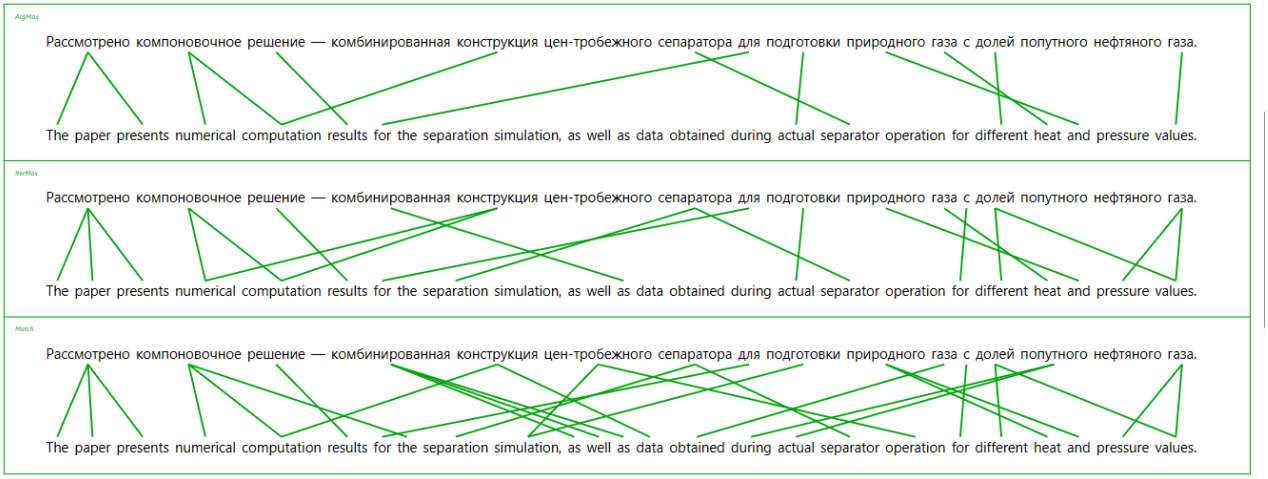


Рис. 3.5. Пример выравнивания двух предложений по одной тематике, не являющихся переводными эквивалентами

При этом похожая ситуация часто возникает при выравнивании параллельных предложений на русском и английском языках, как показано на рис. 3.6 для предложений:

«*Например, обычные текстовые данные имеют скрытую структуру, характерную для естественных языков*» и «*This is definitely true of all text data if you count the latent linguistic structure of human languages*».

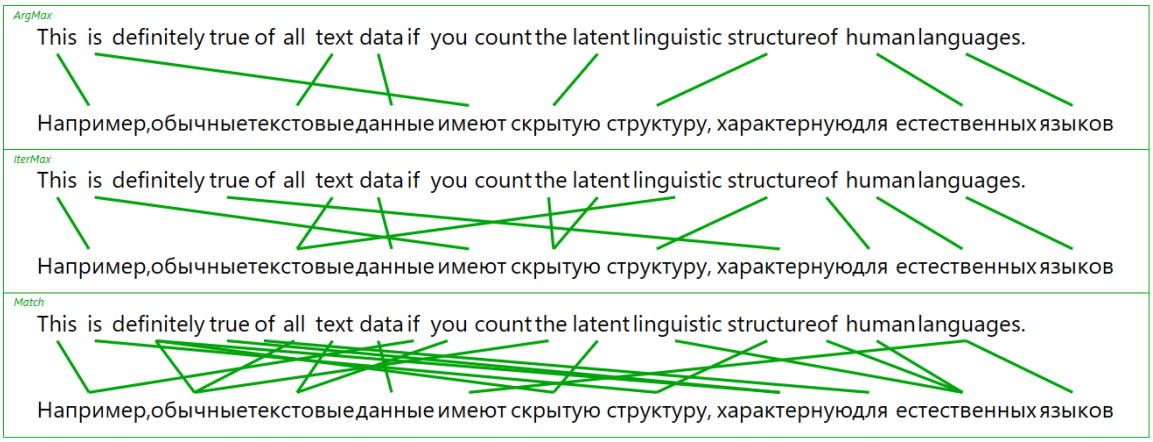


Рис. 3.6. Пример выравнивания параллельных предложений

Более того, на практике встречаются тексты, в которых один и тот же терминоэлемент может встречаться в предложении два и более раза на английском языке, а в русском может быть переведен как существительное и прилагательное (Рис. 3.7) или вообще не установить связи между парами повторяющихся слов.

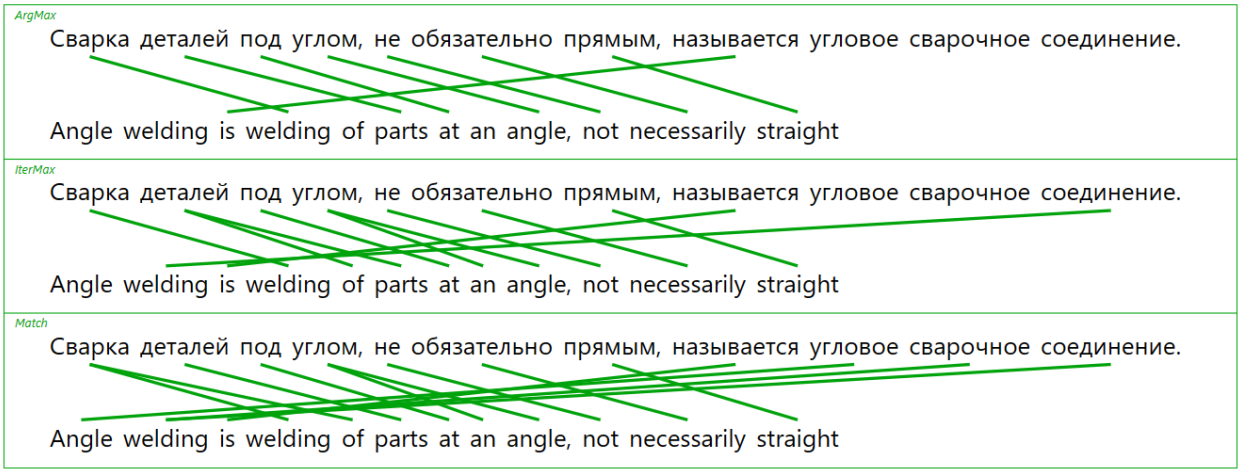


Рис. 3.7. Пример некорректного выравнивания термина «сварка/welding»

На рис. 3.8 в предложениях «*В оптических* ***сетях*** *доступа и локальных* ***сетях*** *связи используются также более простые в эксплуатации многомодовые полимерные оптические волокна*» и «*Optical access* ***networks*** *and local communication* ***networks*** *also use multimode polymer optical fibres that are easier to operate*», слова *сети / networks* встречаются по 2 раза, а по результатам выравнивания получилась лишь одна пара.

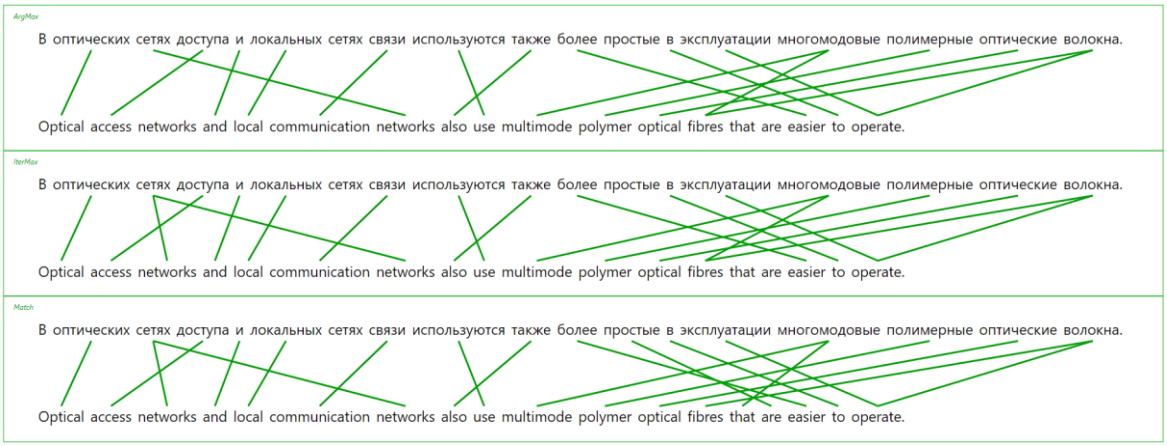


Рис. 3.8. Пример некорректного выравнивания слова «сети / networks»

Таким образом, использование нейронных сетей для выравнивания слов, а затем разметки выровненных терминов в параллельных текстах показывает высокую эффективность для пар языков с фиксированным порядком слов: английский- немецкий, английский-французский и т.д. [234].

В основе второго сценария лежит идея использования «мешка терминов», по аналогии с «мешком слов» [235], когда выравнивание будет происходить между двумя списками терминов, извлеченных из параллельных текстов.

Для извлечения многокомпонентных терминов из параллельных научно-технических текстов на русском и английском языках можно использовать подходы на основе синтаксических шаблонов для русского и английского языков [2]. На вход нейронной сети поступают полученные списки терминов-кандидатов, а задача нейронной сети – выровнять в этих списках слова. Если в двух строках есть совпадения – то есть слова выровнены, а строки, эквивалентные многокомпонентному термину, также будут переводными эквивалентами друг друга.

В связи с тем, что количество терминов-кандидатов, выделенных системой извлечения многокомпонентных терминов, для английского языка разное, и для последних в тексте лексических единиц разница может достигать 100, текст целесообразно предварительно выравнять тексты по абзацам. Абзацы в параллельных текстах всегда строго соблюдены, то есть их количество в обоих текстах всегда одинаковое. Такой подход позволит нивелировать большую

разницу в количестве терминов-кандидатов, что особенно важно для второй половины текста.

В процессе перевода также возможны различные перестановки терминов из-за различий в синтаксической структуре между русским и английским языками, но при реализации разработанного метода разметки многокомпонентных терминов в рамках решения задачи выравнивания многокомпонентных целесообразно ввести допущения на несоответствия порядка следования в списках терминов-кандидатов. Для языка оригинала – этот список будет строго фиксированным по порядку встречаемости слова в тексте, а для текста на языке перевода- выравнивается в соответствии с результатами работы пакета SimAlign.

Для этого нужно сначала выполнить выравнивание параллельных текстов по абзацам. Выравнивание по предложениям может не привести к желаемому результату в силу того, что предложения при переводе довольно часто объединяют в одно или разбивают на несколько. Абзацы же при переводе подобных изменений не претерпевают.

Рассмотрим предложенный подход на примере научно-технической статьи, первый и последний абзацы которой представлены в таблице 3.8.

Таблица 3.8. Фрагмент параллельных научно-технических текстов

1 абзац	
Как отмечено ранее (часть 1 [1]), совмещенная шкала абсолютных потенциалов может иметь двойственный характер. Она допускает совместное использование не только величин абсолютных (а) поверхностных потенциалов E_S^a, относящихся к существованию промежуточных частиц – адатомов и вакансий, но и абсолютных потенциалов E_i^a любых обратимых электродных реакций. Здесь особо будет рассмотрено использование абсолютного потенциала водородного электрода и ПНЗ металла, на котором выделяется водород. Переход от шкалы абсолютных поверхностных потенциалов (ASP) к	*As was noted previously (Part 1 [1]), the combined scale of absolute potentials can have a dual character. This assumes the sharing of not only the absolute (a) surface potentials E_S^a, related to the existence of intermediate particles, adatoms and vacancies, but also the absolute potentials of any reversible electrode reactions E_i^a. Here, the use of the absolute potential of the PZC, on which hydrogen is evoluted, will be specifically considered. The transition from the scale of absolute surface potentials (ASPs) to the hydrogen scale (SHS) is necessary for the practical application of scale ASPs. This transition is based on an analysis of the mechanism of hydrogen*

водородной шкале (SHS) необходим для практического применения шкалы ASP. Этот переход произведен на основе анализа механизма выделения водорода, в котором выделяют две основные стадии:	*evolution, in which there are two main stages:*
……	
Последний абзац	
В процессах выделения водорода на металлах и пассивации металлов наблюдается переход от поликристаллической структуры поверхности ГЦК металлов к монокристаллической (111), который объясняется минимумом поверхностной энергии грани (111) по сравнению с энергией других граней поликристалла, а также быстрой диффузией атомов, связанной с высокой концентрацией атомных вакансий в поверхностном слое металла (см. часть 1).	*In the process of hydrogen evolution on metalsand passivation of metals, there is a transition from the polycrystalline structure of the surface of the fcc metals to a single crystal (111), which is explained by the minimum surface energy of the (111) facet compared to the energy of other facets of the polycrystal, as well as by the rapid diffusion of atoms associated with a high concentration atomic vacancy in the surface layer of the metal (see part 1).*

Результат извлечения терминов-кандидатов из вышеприведенного фрагмента текста с использованием системы извлечения терминов, разработанной на основе предложенных выше моделей и методов, приведен в таблице 3.9 в порядке их следования в текстах.

Таблица 3.9. Списки терминов-кандидатов на русском и английском языках

Абзац 1	
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. величина	5. absolute
6. поверхностный потенциал	6. surface potentials ESa
7. ESa	7. existence of intermediate particles
8. существование промежуточных частиц	8. adatoms
9. адатом	9. vacancies
10. вакансия	10. absolute potentials
11. абсолютный потенциал	11. reversible electrode reactions
12. eia	12. use of absolute potential
13. обратимая электродная реакция	13. PZC

14. использование абсолютного потенциала	14. hydrogen
15. водородный электрод	15. transition
…..	
Последний абзац	
1. процесс выделения	1. process of hydrogen evolution
2. водород	2. metals
3. металл	3. passivation of metals
4. пассивация металлов	4. transition
5. переход	5. polycrystalline structure of surface
6. поликристаллическая структура поверхности	6. fcc metals
7. минимум поверхностной энергии	7. single crystal
8. грань	8. minimum surface energy
9. сравнение	9. energy of facets
10. энергия граней	10. polycrystal
11. поликристалл	11. rapid diffusion of atoms
12. быстрая диффузия атомов	12. high concentration atomic vacancies

На рис. 3.9 представлен результат работы пакета SimAlign по выравниванию терминов.

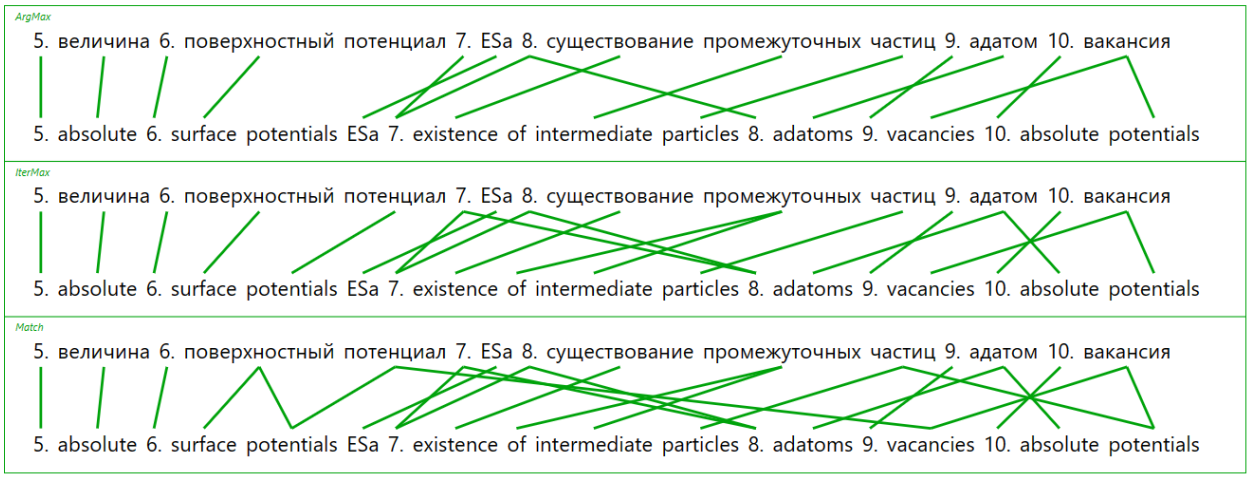


Рис. 3.9. Выравнивание элементов терминов в пакете SimAlign

В таблице 3.10 цветом выделены ядерные элементы терминов в английском и русском языках. Совпадение цвета означает, что термины

являются переводными эквивалентами и должны быть выровнены в корпусе.

Таблица 3.10. Результат работы пакета SimAlign по выравниванию терминов

Абзац 1	
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. величина	5. absolute
6. поверхностный потенциал	6. surface potentials ESa
7. ESa	7. existence of intermediate particles
8. существование промежуточных частиц	8. adatoms
9. адатом	9. vacancies
10. вакансия	10. absolute potentials
11. абсолютный потенциал	11. reversible electrode reactions
12. eia	12. use of absolute potential
13. обратимая электродная реакция	13. PZC
14. использование абсолютного потенциала	14. hydrogen
15. водородный электрод	15. transition
…..	
Последний абзац	
1. процесс выделения	1. process of hydrogen evolution
2. водород	2. metals
3. металл	3. passivation of metals
4. пассивация металлов	4. transition
5. переход	5. polycrystalline structure of surface
6. поликристаллическая структура поверхности	6. fcc metals
7. минимум поверхностной энергии	7. single crystal
8. грань	8. minimum surface energy
9. сравнение	9. energy of facets
10. энергия граней	10. polycrystal
11. поликристалл	11. rapid diffusion of atoms
12. быстрая диффузия атомов	12. high concentration atomic vacancies

Таблица 3.11. Списки выровненных терминов

Абзац 1	
1. часть	1. part
2. шкала абсолютных потенциалов	2. scale of absolute potentials
3. двойственный характер	3. dual character
4. совместное использование	4. sharing
6. поверхностный потенциал	6. surface potentials ESa
8. существование промежуточных частиц	8. existence of intermediate particles
9. адатом	9. adatoms
10. вакансия	10. vacancies
11. абсолютный потенциал	absolute potentials of reversible electrode
12. использование абсолютного потенциала	12. use of absolute potential
……	
Последний абзац	
1. процесс выделения	1. process of hydrogen evolution
2. металл	2. metals
3. пассивация металлов	3. passivation of metals
4. переход	4. transition
5. поликристаллическая структура поверхности	5. polycrystalline structure of surface
7. минимум поверхностной энергии	7. minimum surface energy
9. энергия граней	9. energy of facets
10. поликристалл	10. polycrystal
11. быстрая диффузия атомов	11. rapid diffusion of atoms

В таблице 3.12 представлены результаты выравнивания терминов в параллельных текстах на английском и русском языках для предложенных выше сценариев.

Таблица 3.12. Оценка эффективности сценариев извлечения 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**

Стоит отметить, что точность результатов, представленных в таблице 5 напрямую зависит от результатов извлечения терминов из английского и русского языков: корректно выровненными считались только пары терминов, в которых термины были выделены правильно, то есть пример *welding parts* и *угловая сварка частей*, хотя формально и совпадают с шаблонами терминов в обоих языках, но у англоязычного термина упущен терминоэлемент *angle*. При этом, результаты второго подхода оказались выше для рассматриваемой языковой пары в связи с тем, что учитывают некоторые особенности работы пакета и различия в самих языках.

3.6. Выводы по главе

Установлено, что многокомпонентные англо- и русскоязычные термины разных предметных областей в рамках одного языка имеют одни и те же способы словообразования, а сами структурные модели имеют схожие принципы их образования, что позволяет использовать их при терминологической разметке многокомпонентных терминов в параллельном корпусе.

Получены следующие результаты:

1. Выделено 24 модели русскоязычных и 21 модель для англоязычных многокомпонентных терминов на основе терминосистемы предметной области «Сварка»/ «Welding types». При формировании перечня структурных моделей англо- и русскоязычных многокомпонентных терминов учтены ошибки в определении частеричной принадлежности причастий, деепричастий и некоторых местоимений, а также падежей имен прилагательных, возникающих при осуществлении автоматической морфологической разметки. Предложен способ определения ядерного элемента многокомпонентных терминов.

2. Выделены структурные модели англо- и русскоязычных номенклатурных наименований, которые учитывают не только самые разнообразные варианты структур номенклатурных наименований, но и вариации как в их написании, так и возможностях обработки морфологическими анализаторами. За основу при выделении моделей номенклатурных

наименований взяты номенклатурные наименования предметной области «Космонавтика».

3. Предложен метод извлечения англо- и русскоязычных многокомпонентных терминов в корпусе научно-технических текстов на основе структурных моделей терминологических единиц, который состоит из следующих этапов: анализ предложения по частям речи, исключение частей речи, не входящих в состав терминологических словосочетаний, проверка терминов-кандидатов на стоп-слова, добавление новых терминов и их контекстов в словарь параллельного корпуса

4. Разработан метод извлечения англо- и русскоязычных номенклатурных наименований из параллельных научно-технических текстов. Показано, что основой метода являются морфологическая и терминологическая разметки. Обосновано, что поиск номена как части номенклатурного наименования следует реализовывать после разметки терминов.

5. Предложен метод выравнивания англо- и русскоязычных многокомпонентных терминов в параллельных научно-технических текстах на основе пакета SimAlign. Проанализировано два сценария выравнивания терминов: выравнивание слов + разметка терминов и разметка терминов + выравнивание элементов термина. Более высокий показатель эффективности второго сценария – 84% объясняется разницей в формальной структуре терминов в английском и русском языках.

Основные результаты к разделу опубликованы в работах [1, 2, 4-8, 221, 236, 237, 238].

4. МЕТОДЫ РАЗМЕТКИ РУССКОЯЗЫЧНЫХ МАШИННО-СГЕНЕРИРОВАННЫХ И МАШИНННО-ПЕРЕВЕДЕННХ ТЕКСТОВ

4.1. Актуальное членение предложения как маркер машинных текстов

Термин актуальное членение введен чешским лингвистом В. Матезиусом, который считал, что предложение с точки зрения цели высказывания состоит из двух компонентов: темы, т.е. того, что является в данной ситуации известным или, по крайней мере, может быть легко понято и из чего исходит говорящий, и ремы, то есть то, что говорящий сообщает об исходной точке высказывания» [229]. Соотношение компонентов актуального членения предложения типизировано, представлено в виде моделей, линейно-динамических структур, порядка слов и т.д. Роль темы и ремы могут исполнять любые члены предложения [230].

Актуальное членение предложения как языковая универсалия представлена в каждом языке, но различны средства ее языкового оформления в разных языках: в русском языке чаще всего это порядок слов во взаимодействии с интонацией, а в английском языке способы выражения более разнообразны, например, тема-тематические отношения могут быть выражены через употребление определённых и неопределенных артиклей. При этом прямой порядок слов присущ всем синтаксическим конструкциям английского языка.

В качестве примера рассмотрим предложение «*Группу документов, по которой осуществляется поиск, мы будем называть коллекцией*». Тема в данном случае выражена первой частью предложения *группу документов, по которой осуществляется поиск*, а рема - *будем называть коллекцией.* В переводе это предложение будет иметь следующий вид: «*We will refer to the group of documents over which we perform retrieval as a collection»,* а тема и рема выражены через употребление артиклей. Таким образом, можно предположить, что актуальное членение предложения в русском языке, основанное на семантике и фоновых знаниях автора текста, может служить показателем того, что текст написан или переведен человеком, а не сгенерирован машинным способом.

Для подтверждения гипотезы о том, что актуальное членения предложения может являться основой для выявления машинных текстов проведен сначала сравнительный анализ машинно-сгенерированных, а затем машинно-переведенных текстов с текстами, написанными и переведенными человеком.

Материалом исследования послужили 400 предложений, из которых 200 сгенерированы при помощи больших языковых моделей, а другие 200 предложений отобраны из научно-технических текстов, написанных в 2009-2010 гг., когда средства автоматической генерации текстов еще не получили широкого повсеместного распространения. Тематика сгенерированных текстов охватывает основные понятия информационного поиска. Для анализа актуального членения предложения использован машинный переводчик DeepL, который переводит каждое русскоязычное предложение на английский язык, а затем на основе пакета SimAlign, особенности работы которого описаны в [239], реализовано выравнивание слов каждого предложения.

Современные машинные переводчики способны обнаруживать в русскоязычных предложения главные и второстепенные члены предложения и трансформировать русскоязычное предложение в соответствии с особенностями порядка слов в английском языке.

На рис. 4.1 представлена схема, иллюстрирующая выравнивание слов в русскоязычном предложении и его англоязычном переводном эквиваленте, переведенном с помощью машинного переводчика DeepL, а затем полученная пара предложений выровнена в пакете SimAlign. Пример приведен для предложения «*Группу документов, по которой осуществляется поиск, мы будем называть коллекцией*» и его перевода «*We will refer to the group of documents over which we perform retrieval as a collection*». Зеленые линии указывают на выровненные пары слов, а пересечения этих линий на изменение порядка слов при переводе. В пакете SimAlign есть три режима работы, которые отличаются некоторыми настройками. Для решаемой задачи наиболее подхобдящей является настройка ArgMax, так как в ней наибольшее число однозначных выравниваний слов, то есть 1 к 1.

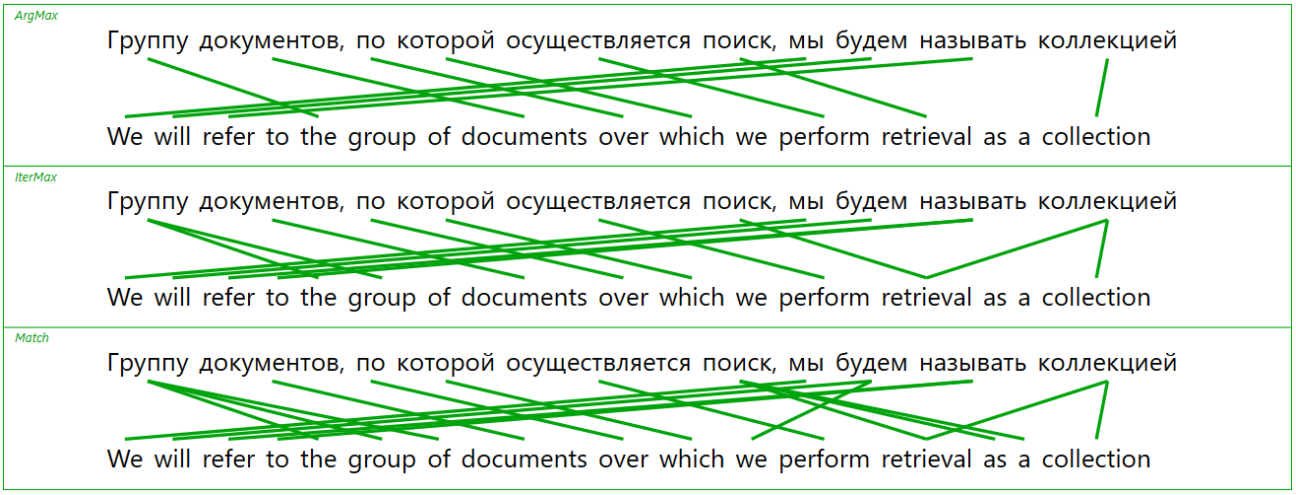


Рис. 4.1. Выравнивание слов в пакете SimAlign

Схема выравнивания по словам для машинно-сгенерированного предложения «*Информационным поиском называют процесс нахождения информации в большом объеме данных с использованием специализированных инструментов и алгоритмов*» представлена на рис. 4.2. Порядок слов русско- и англоязычном предложениях полностью совпал.

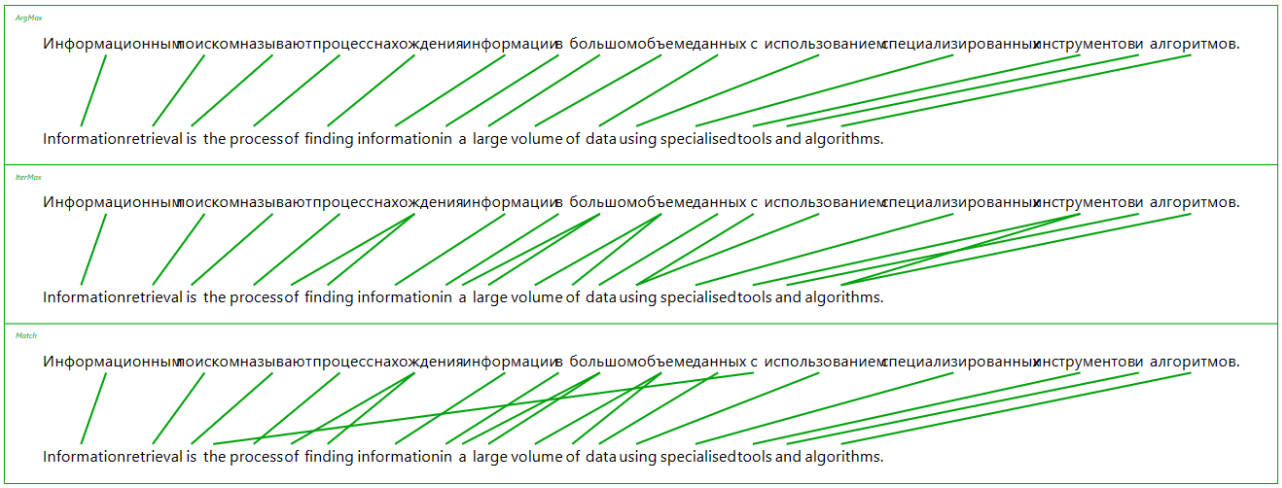


Рис. 4.2. Пример выравнивания слов в машинно-сгенерированном тексте

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.

4.3 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. А на рис. 4.1, например, пара слов *we* (1) и *мы* (7) имеет расхождение 6.

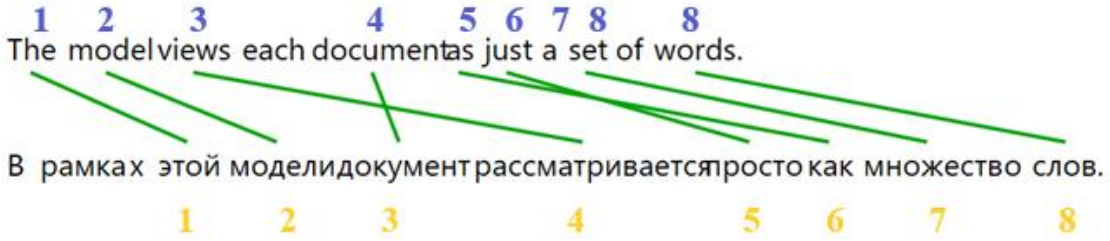


Рис. 4.3. Установление соответствия порядка слов в русском и английском языках

В таблице 4.1 представлен сравнительный анализ изменения порядка слов в машинно-сгенерированных и написанных человеком русскоязычных текстах.

Таблица 4.1. 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-сгенерированный 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 таблице 4.1 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. 1 – это 6, а для рис. 2 – 0, то есть полное совпадение порядка слов.

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 слова – 54% случаев. Отсюда следует, что чат-боты 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.

Материалом исследования послужили 400 предложений, из которых 200 переведены с использование машинного переводчика, а другие 200 предложений и их переводные эквиваленты отобраны из научно-технических текстов, переведенных в 2009-2010 гг. Тематика параллельных текстов охватывает основные понятия информационного поиска. Для анализа актуального членения предложения использован машинный переводчик, который переводит каждое русскоязычное предложение на английский язык, а затем на основе пакета SimAlign проводится выравнинивание слов в параллельных предложениях. Результаты сравнительного анализа приведены в таблице 4.2.

Таблица 4.2. Сравнительный анализ изменения порядка слов в машинно-переведенных 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

Таким образом, наиболее существенным маркером текстов, переведенных человеком, является использование так называемых переводческих

трансформаций. В то же время при машинном переводе этого же предложения переводческие трансформации отсутствуют. Стоит отметить, что использования разного вида переводческих трансформаций не в меньшей степени свойственно и при выполнении ручного перевода с русского на английский язык.

4.2. Метод выявления русскоязычных машинно-сгенерированных текстов на основе особенностей актуального членения предложения

Метод выявления русскоязычных машинного-сгенерированных текстов на основе анализа особенностей актуального членения предложения, описанных в главе 4.1, представлен на рис. 4.4.

Рис. 4.4. Метод выявления машинно-сгенерированных фрагментов русскоязычных текстов

На первом этапе необходимо текст разбить на предложения, а затем к каждому предложению добавить его англоязычный вариант, полученный с

помощью машинного переводчика. На следующем этапе в пакете SimAlign выровнять лексические единицы в русско- и англоязычном предложении, а затем каждому русскоязычному предложению выставить максимальное значение расхождения в порядке слов в предложении и его переводном эквиваленте.

На основе результатов, полученных в таблице 4.1, при реализации метода выявления машинно-сгенерированных текстов необходимо учитывать несколько особенностей. Во-первых, на практике встречаются случаи, когда порядок слов в русском и английском языках совпадает, но частотность такого явления для русского языка значительно меньше, при этом высок процент изменения порядка слов на 4 и более позиций. Во-вторых, встречаются незначительные расхождения в порядке слов, что обусловлено в подавляющем большинстве случаев расхождением в формальной структуре терминов в русском и английском языках: *источник информации – information resource, результаты поиска - search results*. В-третьих, значительные расхождения в порядке слов при выявлении машинно-сгенерированных текстов должны иметь более значимые весовые коэффициенты, однако наличие только 1 такой пары слов в предложении может быть следствием некорректной работы пакета SimAlign. Так, например, в одном из рассмотренных машинно-сгенерированных предложений «*Ранжирование в информационном поиске - это процесс упорядочивания результатов поиска на основе их релевантности запросу пользователя (Ranking in information retrieval is the process of ordering search results based on their relevance to the user query)* у слова *ранжирование* есть две связи *ranking* и *retrieval*, хотя такое двойное выравнивание в программе свойственно лишь для обозначения артиклей и имен существительных или предлогов и имен существительных (Рис. 4.5).

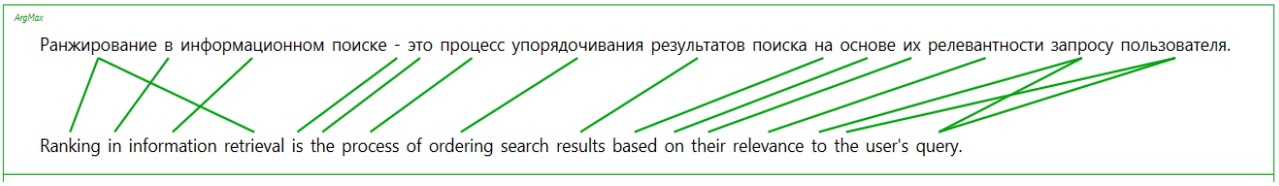


Рис. 4.5. Пример некорректной работы пакета SimAlign

На рис. 4.6. и 4.7 показаны результаты анализа расхождений порядка слов у написанного человеком и машинно-сгенерированного текстов соответственно. В качестве примера написанного человеком текста взята научно-техническая статья по информатике, опубликованная в 2010 году. По аналогичной статье тематике сгенерирован текст с помощью инструмента Wordybot.ai.

Рис. 4.6. Расхождения порядка слов для написанного человеком текста

Рис. 4.7. Расхождения порядка слов в тексте, сгенерированном Wordybot.ai

Сравнительный анализ расхождений порядка слов в текстах, написанных человеком и сгенерированных нейронной сетью представлены на рис. 4.8. и 4.9.

Рис. 4.8. Различия в порядке слов машинных и написанных человеком текстах

Рис. 4.9. Различия в порядке слов машинных и написанных человеком текстах в нормализованном виде

Средние значения расхождения порядка слов в машинно-сгенерированных текстах составляют 4,3-4,6. При этом, анализ пиковых значений расхождения синтаксических структур (Рис. 4.10) показал, что они часто являются результатом ошибок в выравнивании слов нейронной сетью, использованной в пакете SimAlign.

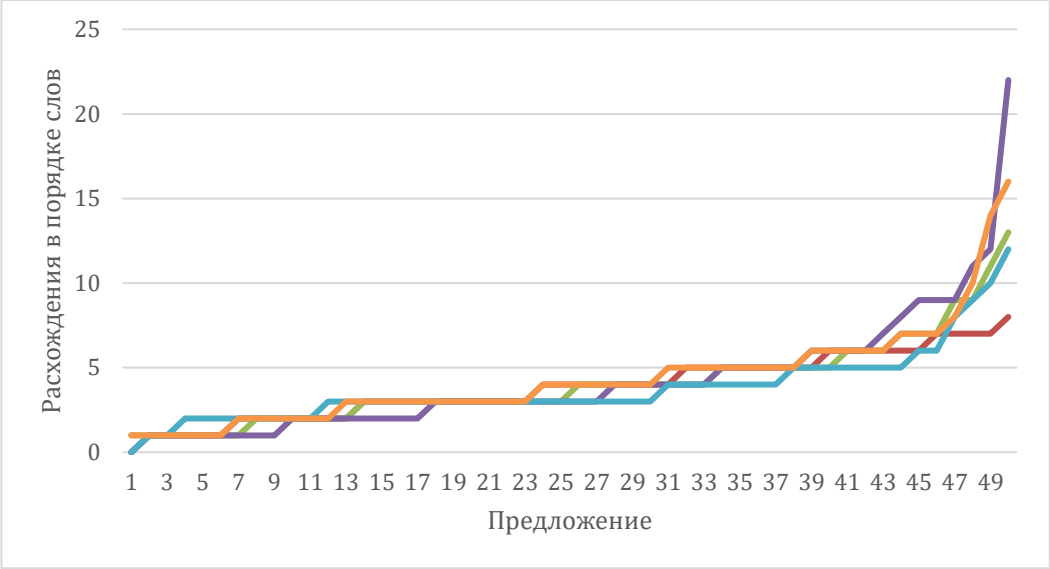


Рис. 4.10. Различия в порядке слов в машинно-сгенерированных текстах

Таким же способом были проанализированы и различия в синтаксических структурах для научно-технических статей, относящихся к разным научным направлениям: математика, космонавтика, языкознание и информатика (Рис. 4.11).

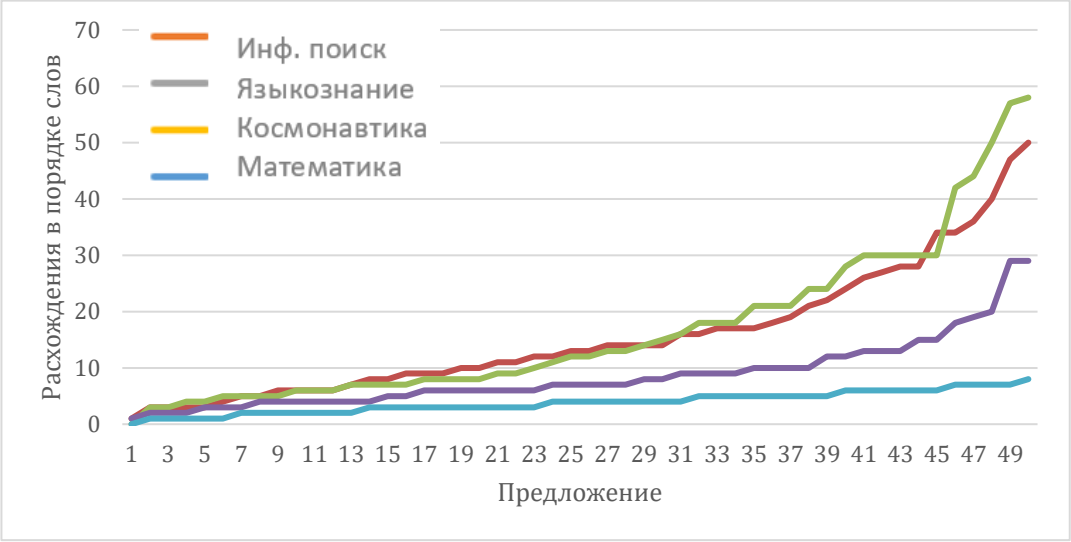


Рис. 4.11. Различия в порядке слов в написанных человеком текстах разных научных направлений

Результаты показали, что наличие существенных различий между текстами разных научных направлений, которые могут привести к появлению

ложноположительных результатов при выявлении машинно-сгенерированных текстов. Так, у научно-технических текстов по математике различия в синтаксических структурах близки к машинно-сгенерированным из-за широкого употребления формул и высокой клишированности синтаксических конструкций. Статистическая обработка научно-технических текстов в аспекте выявления машинных текстов и их фрагментов описана в главе 4.5.

4.3. Метод выявления переводческих трансформаций как маркеров ручного перевода научно-технических текстов

Машинно-переведенные тексты с английского языка на русский также сохраняют прямой порядок (Таблица 4.2). Кроме того, к различиям между машинным и ручным переводом относятся так называемые переводческие трансформации, которые используют лингвисты при выполнении ручного перевода.

Переводческие трансформации – технические приёмы преобразования элементов исходного текста с целью достижения эквивалентности перевода, то есть сохранения равенства содержательной, семантической, стилистической и функционально-коммуникативной информации в оригинале и переводе [240]. Единой классификации переводческих трансформаций не существует – в своих работах лингвисты используют разные подходы. В.Н. Комиссаров [241] выделяет следующие группы переводческих трансформаций: лексические, грамматические и лексико-грамматические. При этом, в одном предложении может быть использовано несколько трансформаций одновременно. На рис. 4.12 показаны 2 переводческие трансформации, грамматическая - «*морфонологическое расщепление = Morpheme split into*» и лексическая – добавление слова *distinct* в английском языке. Использование переводческих трансформаций можно увидеть при визуализации результатов выравнивания фрагментов англо-и русскоязычного предложения, а определить к какой группе принадлежит трансформация можно на основе морфологических характеристик невыровненных или неравномерно выровненных (соотношение 1 к 2 и более)

лексических единиц.

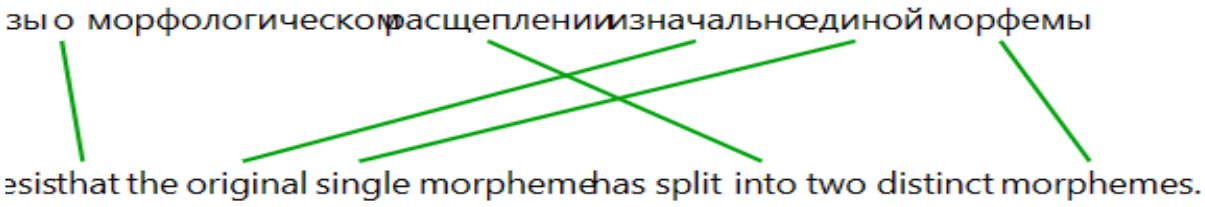


Рис. 4.12. Выявление непереведенных лингвистических единиц в параллельных текстах

Метод выявления переводческих трансформаций в параллельных текстах представлен на рис. 4.13.

Рис. 4.13. Основные этапы метода выявления переводческих трансформаций в параллельных текстах

Наличие лексической трансформации в переводе можно определить по наличию невыровненных лексических единиц. На рис. 4.14 приведен пример лексической трансформации путем добавления в русскоязычном предложении слова «*вопрос*», а на рис. 4.15 опущен причастный оборот «*записанных крюковой нотацией*». В англоязычной версии эквиваленты этих слов отсутствуют.

Рис. 4.14. Пример лексической трансформации (1)

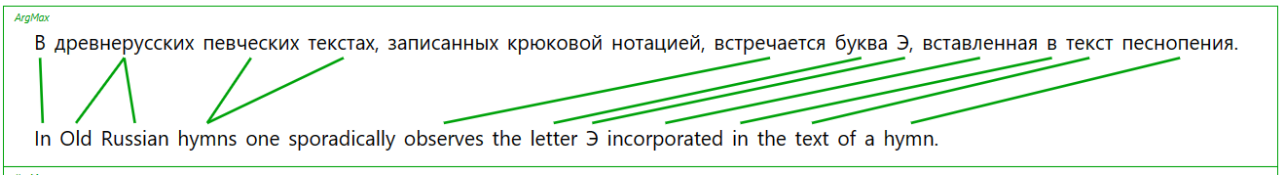


Рис. 4.15. Пример лексической трансформации (2)

Грамматические трансформации чаще всего проявляются в актуальном членении предложения и грамматических заменах, проявляющихся в заменах частей речи их морфологических и синтаксических особенностях. Пример грамматической трансформации с изменением порядка слов в предложении показан на Рис. 4.16

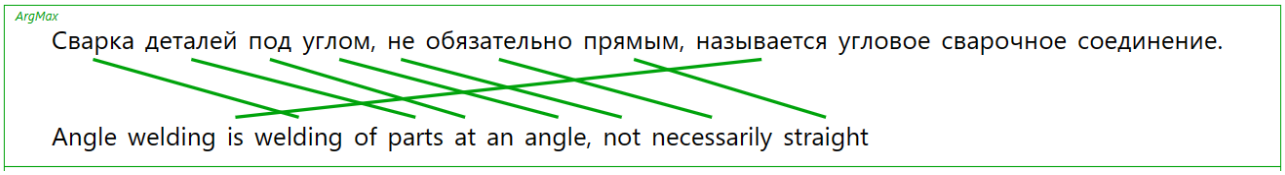


Рис. 4.16. Пример грамматической переводческой трансформации

Суть комплексных или лексико-грамматических трансформаций состоит в комплексном преобразовании предложения на языке перевода, объединяя лексические и грамматические трансформации. В примере на Рис. 4.17 значение имени числительного «*a lot of*» передано именем прилагательным «*интенсивно*», а подлежащие выраженное именем существительным «*research*» в английском языке переведено сказуемым, выраженным глаголом, «*исследуется*» в русском языке, что привело к опущению глагола «*is being carried out*» при переводе на русский язык.

Разметка переводческих трансформаций в параллельном корпусе научно-технических текстов позволит проводить целый спектр исследований в области переводоведения, а также позволит выявить закономерности их преобразования для создания методов их реализации в средствах машинного перевода.

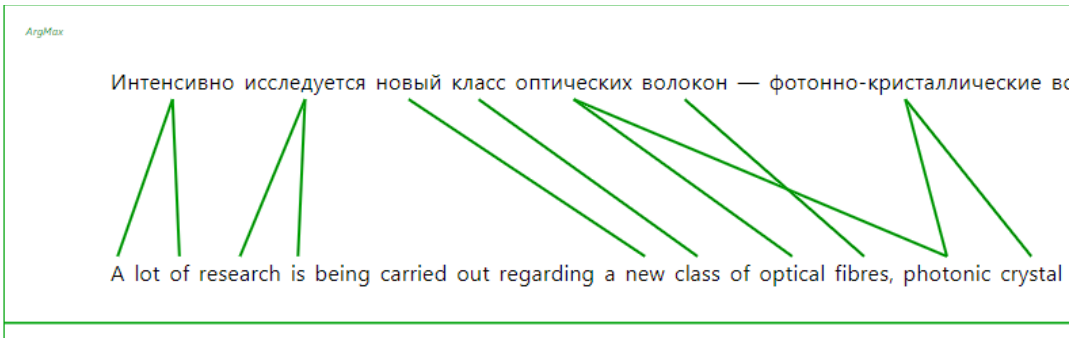


Рис. 4.17. Комплексная переводческая трансформация

4.4. Метод выявления русскоязычных машинно-переведенных текстов на основе особенностей актуального членения предложения

В основе метода выявления машинно-переведенных русскоязычных текстов также как и в методе распознавания машинно-сгенерированных текстов положены особенности актуального членения русскоязычного предложения. Предложенный метод состоит из нескольких этапов, представленных на Рис. 4.18.

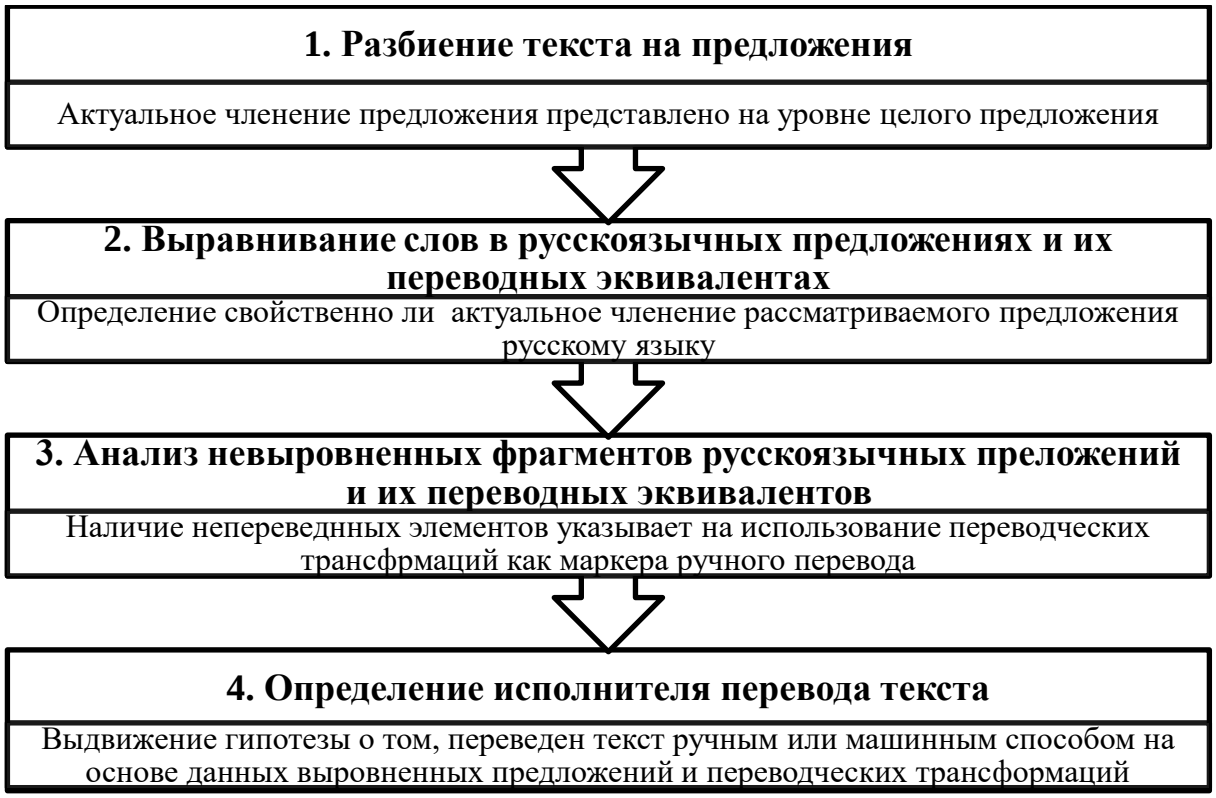


Рис. 4.18. Метод выявления машинно-переведенных фрагментов русскоязычных текстов

На рис. 4.19. и 4.20 показаны результат анализа расхождений синтаксических структур научно-технического текста, переведенного человеком и машинным переводчиком соответственно. В качестве примера приведен оригинал текста учебного пособия «Introduction to information retrieval» (2009), его переводная версия «Введение в информационный поиск» Маннинг Кристофер Д., Рагхаван П., Шютце Х., опубликованная в 2011 году, а также перевод оригинального текста, выполненного машинным переводчиком DeepL.

Рис. 4.19. Расхождения синтаксических структур в тексте, переведенном человеком

Рис. 4.20. Расхождения синтаксических структур в машинно-переведенном тексте

На Рис. 4.21 и 4.22 представлено сравнение расхождений синтаксических структур при ручном и машинном переводе. Средние значения расхождения синтаксических структур в машинно-переведенных текстах составляют 4,7-5,3. А средние оценки расхождений синтаксических структур в переведенных человеком текстах зависят от опыта и квалификации самого переводчика, использования им машинных переводчиков для ускорения перевода, а также значительное влияние оказывают стилистические особенности текста оригинала.

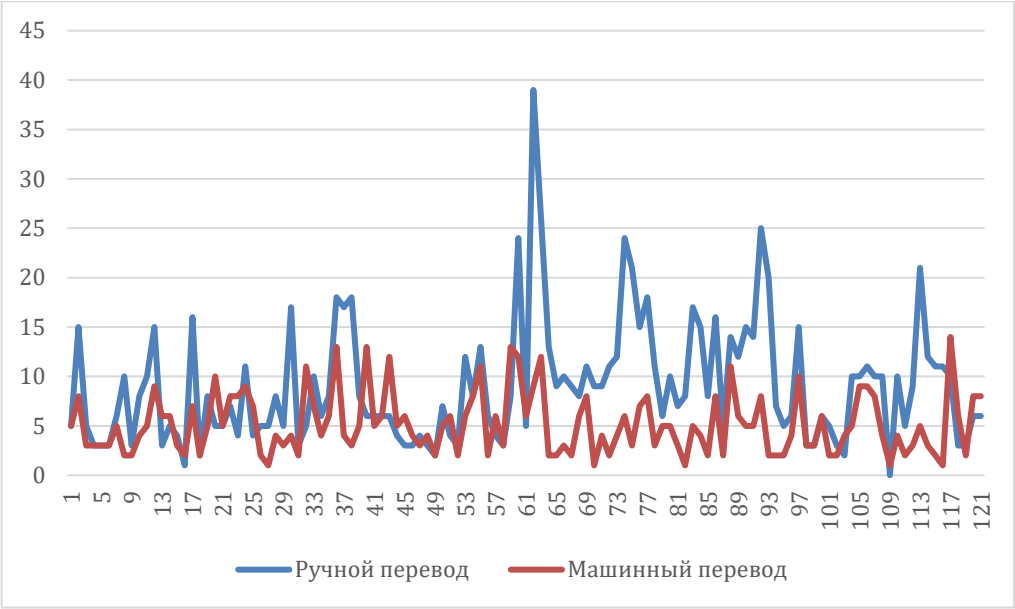


Рис. 4.21. Сравнение расхождений синтаксических структур при ручном и машинном переводе

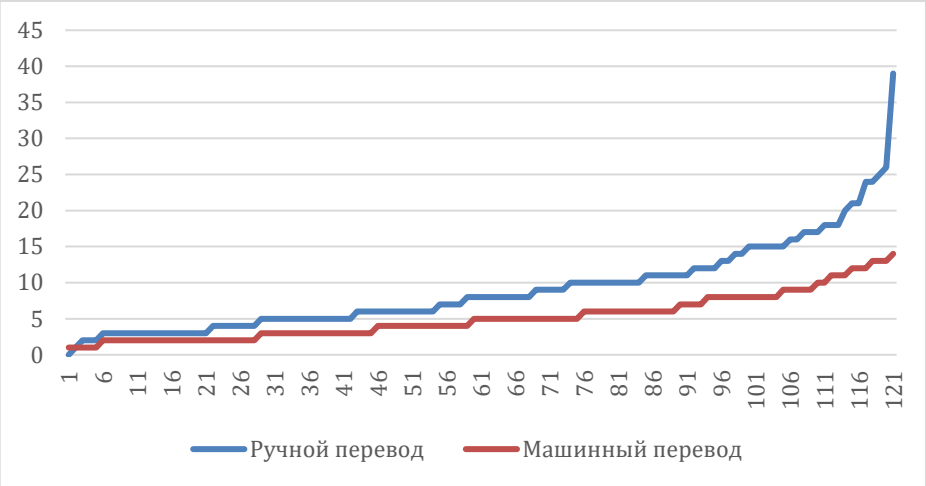


Рис. 4.22. Сравнение расхождений синтаксических структур при ручном и машинном переводе в нормализованном виде

4.5. Статистическая обработка научно-технических текстов в аспекте выявления машинных текстов и их фрагментов

В связи с тем, что разрабатываемый корпус может быть использован как набор обучающих данных для нейронных сетей, то необходимо предусмотреть возможность выявления машинных текстов.

На основе результатов анализа различий в синтаксических структурах для научно-технических статей, относящихся к разным научным направлениям, показанные на Рис. 4.11 авторство текста целесообразно проверять на основе конкурирующих гипотез, то есть оценивать близость текста к машинно-сгенерированным текстам или текстам, написанным человеком и принадлежащим к одному научному направлению [242].

Каждому предложению анализируемого текста присвоим балл равный его значению максимального расхождения в порядке слов между предложениями на английском и русском языках. Получены суммарные оценки расхождений для 7 фрагментов текстов, написанных человеком. Каждый фрагмент состоит из 12-15 предложений, и средние значения расхождений в каждом из фрагментов равны: 25; 23; 28; 26; 30; 22; 19.

Вариационный ряд [243] выборки $n = 7$

$$x_{(1)} = 19;\ x_{(2)} = 22;\ x_{(3)} = 23;\ x_{(4)} = 25;\ x_{(5)} = 26;\ x_{(6)} = 28;\ x_{(7)} = 30.$$

Эмпирическая функция распределения $F_n(x) = \frac{r(x)}{n}$ [244];

$$\hat{F}_n(x) = \begin{cases} 0 & x \le 19 \\ \frac{1}{7} & 19 \le x \le 22 \\ \frac{2}{7} & 22 \le x \le 23 \\ \frac{3}{7} & 23 \le x \le 25 \\ \frac{4}{7} & 25 \le x \le 27 \\ \frac{5}{7} & 27 \le x \le 28 \\ \frac{6}{7} & 28 \le x \le 30 \\ 1 & x > 30.5 \end{cases}$$

Найдем характеристики выборки.

Размах выборки:

$$R_n = \mathrm{x}_{(7)} - \mathrm{x}_{(1)} = 30 - 19 = 11$$

Среднее выборочное:

$$\overline{\mathrm{x}} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} \mathrm{x}_{\mathrm{i}}\ ; \qquad \overline{\mathrm{x}} = \frac{1}{7}(19 + 22 + 23 + 25 + 26 + 28 + 30) = 24.7$$

Дисперсия:

$$\mathrm{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{\mathrm{i}} - \bar{x})^2\ ; \qquad \mathrm{S}^2 = \frac{1}{7}(5.7^2 + 2.7^2 + 1.7^2 + 0.3^2 + 1.3^2 + 3.3^2 + 5.3^2) = 11.9$$

Среднеквадратическое отклонение:

$$\mathrm{S} = \sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{n}(x_{\mathrm{i}} - \bar{x})^2} = \sqrt{S^2} \qquad \mathrm{S} = \sqrt{11.7} = 3.45$$

Затем методом моментов по выборке $\mathrm{x}_1, \mathrm{x}_2 \ldots \mathrm{x}_{\mathrm{n}}$ находим точечные оценки [245] параметров для плотности распределения для функции: (указать справочник откуда взята функция по полученным числам):

$$\mathrm{f(x)} = \frac{\lambda\sqrt{\lambda x}}{\Gamma(3/2)} \cdot e^{-\lambda x},\ x > 0$$

$$\mu_1 = \mu_1(\lambda) = \int_{-\infty}^{+\infty} x f(x,\lambda)dx = \int_0^{\infty} \frac{\lambda^{\frac{3}{2}} x^{\frac{3}{2}}}{\Gamma(3/2)} e^{-\lambda x}dx = \left| t = \lambda x \right| = \frac{1}{\lambda}\int_0^{\infty} \frac{t^{\frac{3}{2}}}{\Gamma(3/2)} e^{-t}dt = \frac{3/2 \cdot \Gamma(3/2+1)}{\Gamma(3/2)\cdot\lambda} = \frac{3}{2\lambda}$$

$$\mu_2 = \mu_2(\lambda,\alpha) = \int_{-\infty}^{+\infty} x f(x,\lambda)dx = \int_0^{\infty} \frac{\lambda^{\alpha} x^{\alpha+1}}{\Gamma(\alpha)} e^{-\lambda x}dx = \left| t = \lambda x \right| = \frac{1}{\lambda^2}\int_0^{\infty} \frac{t^{\alpha+1}}{\Gamma(\alpha)} e^{-t}dt = \frac{1}{\lambda^2}\frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} = \frac{\alpha(\alpha+1)}{\lambda^2} = \frac{15}{4\lambda^2}$$

$$D(\lambda,\alpha) = \mu_2(\lambda,\alpha) - \mu_1{}^2(\lambda,\alpha) = \frac{15}{4\lambda^2} - \frac{9}{4\lambda^2} = \frac{6}{4\lambda^2} = \frac{3}{2\lambda^2}$$

Среднее выборочное значение, полученное выше, $\bar{x} = 24.7$.

Тогда имеем

$$\hat{\lambda} = \frac{\overline{x}}{S^2};\ \hat{\alpha} = \frac{\overline{x}^2}{S^2}$$

$$\hat{\lambda} = \frac{24.7}{11} = 2.25; \quad \hat{\alpha} = \frac{24.7^2}{11} = 55.5$$

Методом максимального правдоподобия [246] по выборке $\mathrm{x}_1, \mathrm{x}_2 \ldots \mathrm{x}_{\mathrm{n}}$ находим точную оценку параметра для плотности распределения Вейбулла:

$$f(x)=\alpha\cdot\Theta\cdot x^{\alpha-1}e^{-\Theta x^{\alpha}},\ x>0,\alpha=2$$

$$x_1=2,\quad x_2=5,\quad x_3=6,\quad x_4=8,\quad x_5=10$$

$$x_{(1)}=19;\ x_{(2)}=22;\ x_{(3)}=23;\ x_{(4)}=25;\ x_{(5)}=26;\ x_{(6)}=28;\ x_{(7)}=30.$$

Запишем функцию правдоподобия [247]:

$$L(x_1,...,x_5,\Theta)=f(x_1,\Theta)f(x_2,\Theta)\cdot...\cdot f(x_n,\Theta)=2\Theta\cdot2\cdot e^{-4\Theta}\cdot2\Theta\cdot5\cdot e^{-25\Theta}\cdot2\Theta\cdot6\cdot e^{-36\Theta}\cdot2\Theta\cdot8\cdot e^{-64\Theta}=$$

$$\cdot2\Theta\cdot10\cdot e^{-100\Theta}=4\Theta e^{-4\Theta}\cdot10\Theta e^{-25\Theta}\cdot12\Theta e^{-36\Theta}\cdot16\Theta e^{-64\Theta}\cdot20\Theta e^{-100\Theta}=153600\Theta^5e^{-229\Theta}$$

Для удобства прологарифмируем функцию:

$$\ln L(x_1,...,x_5,\Theta)=\ln(153600\Theta^5e^{-229\Theta})=\ln153600+5\ln\Theta-229\Theta=5\ln\Theta-229\Theta+11.94$$

Для нахождения максимума функции правдоподобия [243] найдем её производную по параметру Θ и найдем уравнение правдоподобия:

$$\frac{\partial}{\partial\Theta}\ln L(x_1,...,x_5,\Theta)=0$$

$$\frac{\partial}{\partial\Theta}(5\ln\Theta-229\Theta+11.94)=5\frac{1}{\Theta}-229=0$$

Отсюда

$$\frac{5}{\Theta}=229\Rightarrow\hat{\Theta}=\frac{5}{229}=0.02$$

а) Исходная центральная статистика

$$T=\left(\frac{\bar{x}-\mu}{\sigma}\right)\sqrt{n}$$

Система уравнений

$$\left.\begin{cases}\dfrac{\sqrt{n}(\bar{x}-\underline{\mu})}{\sigma}=t_2=U_{1-\varepsilon}\\ \dfrac{\sqrt{n}(\bar{x}-\overline{\mu})}{\sigma}=t_1=U_{\varepsilon}\end{cases}\right|U_{\varepsilon}=-U_{1-\varepsilon}$$

$$\varepsilon = \frac{1-\gamma}{2} = \frac{1-0.9}{2} = 0.05 \Rightarrow U_{0.95} = 1.645$$

Границы интервала:

$$\left.\begin{aligned} \underline{\mu} &= \bar{x} - U_{0.95}\left(\frac{\sigma}{\sqrt{n}}\right) \\ \bar{\mu} &= \bar{x} + U_{0.95}\left(\frac{\sigma}{\sqrt{n}}\right) \end{aligned}\right|$$ Расчеты здесь и далее проведем для выборки из 1-ой задачи $\bar{x} = 25.2;\ n = 7;\ \sigma = 5.5$

$$\underline{\mu} = 25.2 - 1.645\left(\frac{5.5}{\sqrt{7}}\right) = 21.8$$

$$\bar{\mu} = 25.2 + 1.645\left(\frac{5.5}{\sqrt{7}}\right) = 28.6$$

Доверительный интервал $21.8 \le \mu \le 28.6$

б) Исходная центральная статистика:

$$T = \left(\frac{\bar{x} - \mu}{S}\right)\sqrt{n-1}$$

Система уравнений

$$\left.\begin{aligned} \sqrt{n-1}\,\frac{\bar{x} - \underline{\mu}}{S} &= t_{1-\varepsilon}(n-1) \\ \sqrt{n-1}\,\frac{\bar{x} - \bar{\mu}}{S} &= t_{\varepsilon}(n-1) \end{aligned}\right| \quad t_{\varepsilon}(n-1) = -t_{1-\varepsilon}(n-1)$$

$$\varepsilon = \frac{1-\gamma}{2} = \frac{1-0.9}{2} = 0.05$$

$$\left.\begin{aligned} \underline{\mu} &= \bar{x} - t_{1-\varepsilon}(n-1)\frac{S}{\sqrt{n-1}} \\ \bar{\mu} &= \bar{x} + t_{1-\varepsilon}(n-1)\frac{S}{\sqrt{n-1}} \end{aligned}\right| \quad t_{1-0.05}(7-1) = t_{0.95}(6) = 1.94 \qquad \bar{x} = 25.2;\ n = 7;\ S = 3.42$$

$$\underline{\mu} = 25.2 - 1.94\left(\frac{3.42}{\sqrt{6}}\right) = 22.49$$

$$\bar{\mu} = 25.2 + 1.94\left(\frac{3.42}{\sqrt{6}}\right) = 27.9$$

Доверительный интервал $22.49 \le \mu \le 27.9$

в) Исходная центральная статистика

$$T = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2}$$

Система уравнений

$$\begin{cases} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\underline{\sigma}^2} = \chi^2_{1-\varepsilon}(n) \\ \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\overline{\sigma}^2} = \chi^2_{\varepsilon}(n) \end{cases}$$

$$\varepsilon = \frac{1-\gamma}{2} = \frac{1-0.95}{2} = 0.025$$

$$\underline{\sigma} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\chi^2_{1-\varepsilon}(n)}}; \quad \overline{\sigma} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\chi^2_{\varepsilon}(n)}}$$

$$\chi^2_{1-\varepsilon}(n) = \chi^2_{0.975}(7) = 16.01 \qquad \chi^2_{\varepsilon}(n) = \chi^2_{0.025}(7) = 1.69$$

$$\underline{\sigma} = \sqrt{\frac{(19.7-25)^2 + (22.8-25)^2 + (23.1-25)^2 + (25.3-25)^2 + (26.9-25)^2 + (28.4-25)^2 (30.5-25)^2}{16.01}} = 2.26$$

$$\overline{\sigma} = \sqrt{\frac{(19.7-25)^2 + (22.8-25)^2 + (23.1-25)^2 + (25.3-25)^2 + (26.9-25)^2 + (28.4-25)^2 (30.5-25)^2}{1.69}} = 6.97$$

Доверительный интервал $2.26 \le \sigma \le 6.97$

г) Исходная центральная статистика

$$T = \frac{nS^2}{\sigma^2}$$

Система уравнений

$$\begin{cases} \frac{nS^2}{\underline{\sigma}^2} = \chi^2_{1-\varepsilon}(n-1) \\ \frac{nS^2}{\overline{\sigma}^2} = \chi^2_{\varepsilon}(n-1) \end{cases}$$

$$\varepsilon = \frac{1-\gamma}{2} = \frac{1-0.95}{2} = 0.025$$

$$\chi^2_{1-\varepsilon}(n-1) = \chi^2_{0.975}(7-1) = \chi^2_{0.975}(6) = 14.45$$

$$\chi^2_{\varepsilon}(n-1) = \chi^2_{0.025}(7-1) = \chi^2_{0.025}(6) = 1.24$$

$$\underline{\sigma} = \frac{S\sqrt{n}}{\sqrt{\chi^2_{1-\varepsilon}(n-1)}} \qquad \overline{\sigma} = \frac{S\sqrt{n}}{\sqrt{\chi^2_{\varepsilon}(n-1)}}$$

$$\underline{\sigma} = \frac{3.42\sqrt{7}}{\sqrt{14.45}} = 2.38 \qquad \overline{\sigma} = \frac{3.42\sqrt{7}}{\sqrt{1.24}} = 8.12$$

Доверительный интервал $2.38 \le \sigma \le 8.12$

Приблизительный доверительный интервал $23.54 \le \mu \le 26.857$

Построить доверительный интервал с коэффициентом доверия $\gamma = 0.97$ для параметра вероятности «успеха» p в схеме Бернулли [242] при условии, что в серии из 20 испытании наблюдалось 7 «успехов».

Рассмотрим случай при $m = 7; n = 20$

Центральная статистика

$$T = \frac{m - np}{\sqrt{np(1-p)}}$$

Неравенство

$$-U_{1-\varepsilon} \le \frac{m - np}{\sqrt{np(1-p)}} \le U_{1-\varepsilon}$$

$$\varepsilon = \frac{1-\gamma}{2} = \frac{1-0.97}{2} = 0.015;$$

$$U_{1-\varepsilon} = U_{1-0.015} = U_{0.985} = 2.17; \quad \frac{m}{n} = \frac{7}{20} = 0.35$$

$$\frac{m}{n} - \frac{U_{1-\varepsilon}}{\sqrt{n}}\sqrt{p(1-p)} \le p \le \frac{m}{n} + \frac{U_{1-\varepsilon}}{\sqrt{n}}\sqrt{p(1-p)}$$

Границы

$$\underline{p} = \frac{m}{n} - \frac{U_{1-\varepsilon}}{\sqrt{n}}\sqrt{\frac{m}{n}\left(1-\frac{m}{n}\right)}; \quad \overline{p} = \frac{m}{n} + \frac{U_{1-\varepsilon}}{\sqrt{n}}\sqrt{\frac{m}{n}\left(1-\frac{m}{n}\right)}$$

$$\underline{p} = 0.35 - \frac{2.17}{\sqrt{20}}\sqrt{0.35(1-0.35)} = 0.35 - 0.485 \cdot 0.477 = 0.118$$

$$\overline{p} = 0.35 + \frac{2.17}{\sqrt{20}}\sqrt{0.35(1-0.35)} = 0.35 + 0.485 \cdot 0.477 = 0.581$$

Доверительный интервал $0.188 \le p \le 0.581$

На следующем этапе необходимо проверить гипотезу, извлечена ли некоторая выборка текстов из нормальной генеральной совокупности. В таблице 4.3 дан вариационный ряд из научно-технического текста по космонавтике. Проверку можно провести с помощью χ^2 - критерия Пирсона [247] с количеством интервалов *k*, равным *4* и *6*.

Таблица 4.3. Вариационный ряд из научно-технического текста по космонавтике

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					

n = 49;

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

Для числа интервалов k = 4:

Длина интервала

$$d = \frac{R}{k} = \frac{13.64}{4} = 3.41$$

Интервалы в общем виде:

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

После подстановки:

$$\Delta_1 = (-\infty; 13.446);\ \Delta_2 = (13.446; 16.856);\ \Delta_3 = (16.856; 20.266);\ \Delta_4 = (20.266; +\infty)$$

Количество n_i элементов выборки, попавших в каждый интервал:

$$n_1 = 7 \qquad n_2 = 22 \qquad n_3 = 16 \qquad n_4 = 4$$

Поскольку $n_4 < 5$ объединяем 3$^{\text{й}}$ и 4$^{\text{й}}$ интервалы, тогда:

$$\Delta_1' = (-\infty; 13.446); \quad \Delta_2' = (13.446; 16.856); \quad \Delta_3' = (16.856; +\infty)$$

$$n_1' = 7 \qquad n_2' = 22 \qquad n_3' = 20$$

Вероятность попадания в каждый интервал:

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

$h_0 = -\infty; \quad h_1 = 13.446; \quad h_2 = 16.856, \quad h_3 = +\infty$

найдем среднее выборочное:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{49}\sum_{i=1}^{49} x_i = 16.05$$

найдем СКО:

$$S = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{48}\sum_{i=1}^{49}(x_i - \bar{x})^2} = 3.14$$

$$P_1 = \Phi\left(\frac{13.446 - 16.05}{3.14}\right) - \Phi\left(\frac{-\infty - 16.05}{3.14}\right) = \Phi(-0.83) - \Phi(-\infty) = 0.467$$

$$P_2 = \Phi\left(\frac{16.856 - 16.05}{3.14}\right) - \Phi\left(\frac{13.446 - 16.05}{3.14}\right) = \Phi(0.256) - \Phi(-0.83) = 0.134$$

$$P_3 = \Phi\left(\frac{\infty - 16.05}{3.14}\right) - \Phi\left(\frac{16.856 - 16.05}{3.14}\right) = 1 - \Phi(0.256) = 0.399$$

Величины $C_i = np_i$

$C_1 = 49 \cdot 0.467 = 22.883; \quad C_2 = 49 \cdot 0.134 = 6.566; \quad C_3 = 49 \cdot 0.399 = 19.551$

Значение статистики:

$$\chi^2 = \sum_{i=1}^{3}\frac{(n_i - n_{pi})^2}{n_{pi}} = \frac{(7 - 22.883)^2}{22.883} + \frac{(22 - 6.566)^2}{6.566} + \frac{(20 - 19.551)^2}{19.551} = 11.02 + 36.28 + 0.01 = 47.31$$

Число степеней свободы: $\nu = k - l - 1; \quad k' = 3; \quad l = 2 \Rightarrow \nu = 0$

По таблице 3 из методических указаний для распределения χ^2 видим , что для $\nu = 1$ и $\alpha = 0.001$

$$\chi^2_{0.999}(1) = 10.83$$

Поскольку $\chi^2 > \chi^2_{0.999}(1)$, гипотезу о том, что данная выборка извлечена из нормальной генеральной совокупности нужно отклонить.

б) Для числа интервалов $k = 6$:

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 5й и 6й 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 для 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 $\alpha = 0.01$.

Затем можно проверить, может ли та же самая выборка быть извлечена из логнормальной генеральной совокупности – сложная непараметрическая гипотеза. Проверку можно провести с помощью критерия Колмогорова [243].

Для применения критерия Колмогорова необходимо произвести расчет:

$$c_i = \max\left(\left|\frac{i-1}{n} - F_0(x_i)\right|, \left|\frac{i}{n} - F_0(x_i)\right|\right)$$

Для расчета $F_0(x_i) = \Phi\left(\ln\left(\frac{x_i - \bar{x}}{S}\right)\right)$ - функции Лапласа [244] для логнормального распределения воспользуемся программой MathCAD и функцией plnorm (x, μ, σ).

Здесь в качестве μ берется его оценка $\bar{x} = 16.05$ и вместо σ также берется его оценка $S = 3.14$. Т.е. plnorm(x,16.05,3.14).

Произведем расчет c_1:

$$F_0(x_1) = \Phi\left(\ln\left(\frac{10.036 + 16.05}{3.14}\right)\right) = \Phi(-1.915) = 1 - 0.973 = 0.027$$

$$c_1 = \max\left(\left|\frac{1-1}{49} - 0.027\right|, \left|\frac{1}{49} - 0.027\right|\right) = \max\left(0.027, 6.59 \cdot 10^{-3}\right) = 0.027$$

Расчет остальных c_i производим аналогично. Результаты расчета приведены в таблицах 4.4 и 4.5:

Таблица 4.4. Результаты расчета функции Лапласа для логнормального распределения

i	$F_0(x_i)$	i	$F_0(x_i)$	i	$F_0(x_i)$	i	$F_0(x_i)$	i	$F_0(x_i)$
1	0.027	11	0.267	21	0.227	31	0.645	41	0.801
2	0.032	12	0.269	22	0.240	32	0.685	42	0.825
3	0.05	13	0.273	23	0.350	33	0.716	43	0.846
4	0.063	14	0.275	24	0.411	34	0.721	44	0.855
5	0.077	15	0.297	25	0.495	35	0.725	45	0.877
6	0.132	16	0.304	26	0.516	36	0.746	46	0.953
7	0.136	17	0.308	27	0.528	37	0.753	47	0.975
8	0.240	18	0.316	28	0.557	38	0.777	48	0.987
9	0.253	19	0.261	29	0.593	39	0.787	49	0.991
10	0.260	20	0.296	30	0.606	40	0.791	50	

Таблица 4.5. Результаты расчета функции Лапласа для логнормального распределения

i	c_i	i	c_i	i	c_i	i	c_i	i	c_i
1	0.027	11	0.063	21	0.201	31	0.033	41	0.036
2	0.012	12	0.045	22	0.209	32	0.052	42	0.032
3	0.011	13	0.028	23	0.119	33	0.063	43	0.031
4	0.018	14	0.011	24	0.079	34	0.047	44	0.043
5	0.025	15	0.011	25	0.015	35	0.031	45	0.041
6	0.03	16	0.022	26	0.014	36	0.032	46	0.034
7	0.013	17	0.039	27	0.023	37	0.018	47	0.036
8	0.097	18	0.051	28	0.014	38	0.022	48	0.028
9	0.089	19	0.127	29	0.022	39	0.011	49	0.011
10	0.076	20	0.112	30	0.014	40	0.025	50	

Из таблицы видно, что: $\max c_i = 0.209$

$$T = \max c_i \cdot \sqrt{n} = 0.209 \cdot \sqrt{49} = 1.463$$

Из таблицы для критерия согласия Колмогорова критическое значение $t_{0.01} = 1.63$. Поскольку $T < t_{0.01}$ гипотезу о том, что выборка была извлечена из логнормальной генеральной совокупности следует принять на уровне значимости $\alpha = 0.01$

Проверить, может ли та же самая выборка быть извлечена из экспоненциальной совокупности. Проверку провести с помощью критерия ω^2. Прежде чем приступить к проверке гипотезы по критерию ω^2, необходимо рассмотреть функцию

$$f(x, \lambda) = 1 - e^{-\lambda x}, \quad x > 0$$

По методу максимального правдоподобия точечная оценка [243] параметра λ:

$$\hat{\lambda} = \frac{1}{\bar{x}}; \quad \hat{\lambda} = \frac{1}{16.05} = 0.062$$

Таблица 4.6. Результаты расчетов

i	$F_0(x_i)$	i	$F_0(x_i)$	i	$F_0(x_i)$	i	$F_0(x_i)$	i	$F_0(x_i)$
1	0.4632	11	0.5792	21	0.5923	31	0.6561	41	0.6842
2	0.4685	12	0.5808	22	0.5942	32	0.6635	42	0.6864
3	0.4911	13	0.5812	23	0.6015	33	0.6692	43	0.6968
4	0.5022	14	0.5827	24	0.6284	34	0.6702	44	0.6992
5	0.5121	15	0.5832	25	0.6293	35	0.6709	45	0.7052
6	0.5406	16	0.5841	26	0.6334	36	0.6711	46	0.7336
7	0.5423	17	0.5847	27	0.6354	37	0.675	47	0.7483
8	0.5723	18	0.5896	28	0.6402	38	0.6765	48	0.7616
9	0.5759	19	0.59	29	0.6471	39	0.6814	49	0.7696
10	0.5759	20	0.5915	30	0.6492	40	0.6835	50	

Для $x_{(1)} = 10.036$ $F_0(x_1) = 0.4632$

Тогда

$$F_0(x_1) - \frac{2 \cdot 1 - 1}{2 \cdot 49} = 0.4632 - 0.0102 = 0.4530$$

Расчет остальных слагаемых проведем в программе MathCAD

$$m\omega^2 = 0.1877$$

по таблице для экспоненциального закона распределения критическое значение при $\alpha = 0.01 \quad \omega = 0.2299$. Поскольку $m\omega^2 < \omega_{0.01}$, то гипотезу о том, что выборка научно-технических текстов по космонавтике извлечена из экспоненциальной совокупности можно принять на уровне значимости $\alpha = 0.01$.

Предложенные в главе методы выявления машинно-сгенерированных машинно-переведенных текстов оценивают весь текст целиком, однако, в корпусе необходимо также реализовать возможность выявления отдельных машинно-сгенерированных или машинно-переведенных фрагментов и / или их комбинаций. На практике встречаются случаи, когда пользователь генерирует не весь текст, а его некоторые фрагменты, вставляя фрагменты в написанный человеком текст или переведенный текст из источника на другом языке. К такого рода средствам повышения оригинальности текста или сокращении временных затрат на его создание могут прибегать при написании научных работ недобросовестные авторы, что в свою очередь может оказывать воздействие на результаты поисковой выдачи в корпусе. В связи с этим необходимо предусмотреть возможность исключения таких фрагментов научно-технических текстов из конкорданса.

С этой целью был смоделирован ряд ситуаций, когда в написанный человеком текст добавлены фрагменты машинно-сгенерированных текстов, и наоборот, в машинно-сгенерированный текст вкраплены написанные человеком фрагменты. На Рис. 4.23 представлены расхождения в синтаксических структурах смоделированного научно-технического текста. Машинно-сгенерированные фрагменты добавлены в интервале с 9 по 19 и 60-69 предложения.

В связи с тем, что прямой порядок слов встречается в русском языке (таблица 4.1), целесообразно группировать предложения в интервалы. При этом,

чем больше интервал, тем больше вероятность, что машинно-сгенерированный фрагмент, попавший в этот интервал, будет распознан как написанный человеком, за счет больших значений в расхождениях синтаксических структур.

Рис. 4.23. Написанный человеком текст с вкрапленным сгенерированным текстом

Для выявления фрагментов машинных текстов предложено разбивать текст на интервалы по 3, 5, 7 предложений, а затем относить каждый интервал к машинным или написанным человеком текстах (Рис. 4.24-4.26).

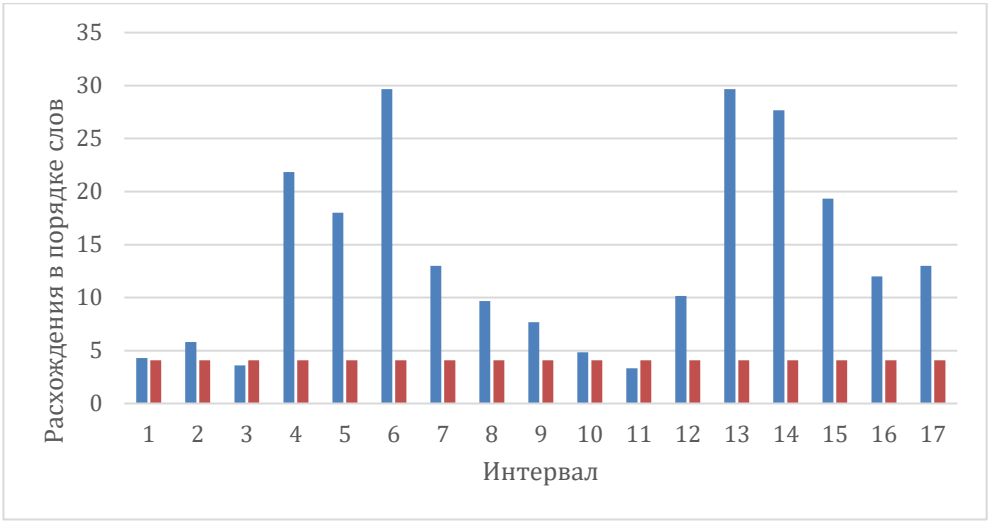


Рис. 4.24. Выявление машинно-сгенерированных фрагментов (в интервале 7 предложений)

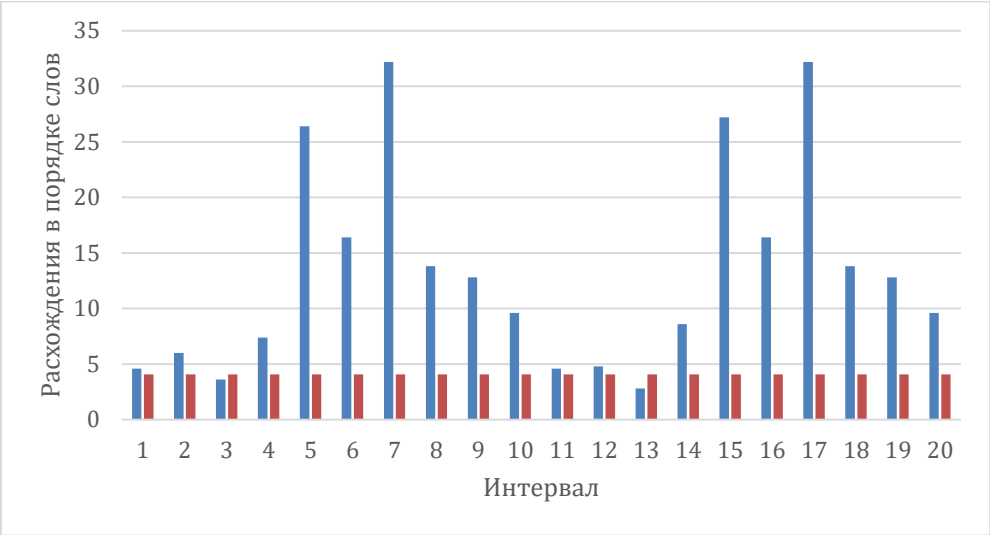


Рис. 4.25. Выявление машинно-сгенерированных фрагментов (в интервале 5 предложений)

Рис. 4.26. Выявление машинно-сгенерированных фрагментов (в интервале 3 предложения)

В связи с тем, что в каждый интервал входит разное количество предложений предложено расчет эффективности проводить по количеству правильно распознанных предложений. То есть при вкраплении фрагментов машинных текстов объемом 10 предложений интервалы могут быть распространены разными способами. Текст из 10 предложений может быть

разбит на 2 интервала по 5 предложений. Если оставшиеся 2 предложения имеют высокие показатели расхождений синтаксических структур, то эти интервалы могут быть распознаны как написанные человеком, то есть некорректные. В связи с этим, предложено сглаживать пиковые значения. Результаты оценки эффективности метода выявления машинных текстов в параллельном корпусе представлены в таблице 4.6.

Таблица 4.6. Оценка эффективности метода выявления фрагментов МТ в научно-технических текстах, в %

	Количество предложений в интервале	Полнота	Точность	F-мера
1	7	70	66	68
2	5	51	50	50
3	3	71	68	70
4	7 (сглаживание)	70	67	69
5	5 (сглаживание)	75	73	74
6	3 (сглаживание)	**83**	**88**	**85**

4.6. Выводы по главе

Предложены методы разметки машинных русскоязычных научно-технических текстов и их фрагментов на основе особенностей актуального членения предложений в русском языке.

Получены следующие результаты:

1. Обосновано, что особенности актуального членения предложения в русском языке являются основой для выявления машинных текстов: машинно-сгенерированных и машинно-переведенных. 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 предложено использовать машинный переводчик, который переводит каждое русскоязычное предложение на английский язык, а затем на основе пакета SimAlign, реализовывать выравнивание слов каждого предложения.

2. Разработан метод выявления русскоязычных машинно-сгенерированных текстов на основе особенностей актуального членения предложения. Представлен сравнительный анализ изменения порядка слов в машинно-сгенерированных и написанных человеком русскоязычных текстах.

3. Проанализированы различия в порядке слов машинно-переведенных текстов и текстов, переведенных человеком. Выявлено, что маркерами машинно-переведённых текстов являются прямой порядок слов в русского язычных текстах и отсутствие каких-либо переводческих трансформаций. Показано, что использование пакета SimAlign позволяет увидеть наличие не выровненных фрагментов текстов, которые при дальнейшем морфологическом анализе можно относить к тому или иному виду переводческих трансформаций.

4. Предложен метод выявления машинно-переведенных русскоязычных текстов или их фрагментов, в основу которого положены данные о синтаксической структуре предложения и наличия в нем переводческих трансформаций.

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.

Основные результаты к разделу опубликованы в работах [3, 28, 31, 71, 236, 247].

5. СИСТЕМА УПРАВЛЕНИЯ КОРПУСНЫМИ ДАННЫМИ ПАРАЛЛЕЛЬНОГО КОРПУСА НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ

5.1. Концепция системы управления корпусными данными параллельного корпуса

Обзор и анализ современных корпусов, проведенный в главе 1, выявил тот факт, что чаще всего корпуса ориентированы на использование лингвистами при проведении широкого спектра исследований языка. При этом, большие размеченные массивы текстовых данных также вызывают значительный интерес у программистов, аналитиков данных и других специалистов. Однако использование существующих корпусов и корпусных менеджеров значительно усложнено отсутствием доступа к размеченным данным, и даже при его наличии нет возможности «посмотреть» на результаты разметки текстов. Таким образом, при проектировании системы корпусных данных необходимо учитывать интересы и потребности широкого круга исследователей, а также обеспечивать возможность:

1. Автоматической разметки и выравнивания текстовых элементов разных уровней для ускорения процессов создания и пополнения корпуса.

2. Просматривать результаты разметки и выравнивания текстовых элементов – то есть по запросу выдавать информацию о том, какие метки были прописаны текстовым единицам автоматическими средствами системы управления корпусными данными.

3. Ручной коррекции результатов автоматической разметки при обязательном сохранении как некорректных, так и исправленных разметок.

4. Создания терминологического словаря или базы данных на основе размеченных текстов, а также его последующего использования при обработке научно-технических текстов в корпусе.

5. Скачать / обрабатывать в корпусе размеченные тексты или их фрагменты, которые соответствуют специальным параметрам, например написаны человеком, а не сгенерированы нейронной сетью.

6. Скачать в виде пригодном для машинной обработки как один, так и несколько видов разметки для решения практических задач по обработке естественного языка.

Обработка научно-технических текстов при создании параллельного корпуса подразумевает не только процедуру разметки текстов и организации эффективного поиска, но и последующую обработку результатов поиска в коллекции таких текстов. На рис. 5.1 представлена обобщенная схема взаимодействия элементов системы управления корпусными данными, в состав которой входит система разметки научно-технических текстов, база данных корпуса, корпусный менеджер и словарь корпуса.

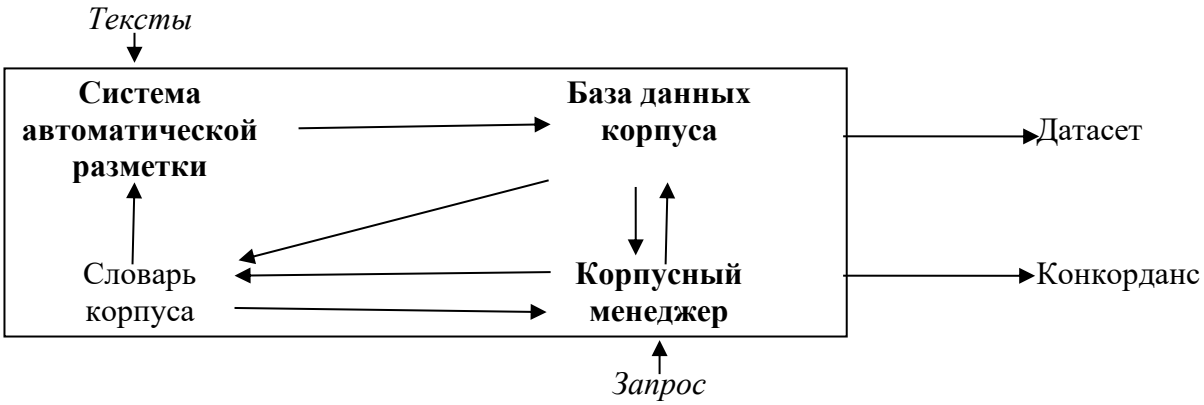


Рис. 5.1. Обобщенная схема взаимодействия элементов системы управления корпусными данными

Система разметки научно-технических текстов состоит из комплекса подсистем, реализующих разные виды разметки, а именно метатекстовую, морфологическую, терминологическую, стилистическую, семантико-синтаксическую. Разработанная система функционирует в двух режимах: режим разметки и навигации по базе данных размеченных текстов корпуса, что значительно расширяет спектр прикладных задач, основанных на размеченных данных. Безусловно, что филологическая корректность является одним из требований, предъявляемых корпусу, однако она требует привлечения экспертов-лингвистов, которые в ручном режиме будут выверять размеченные данные.

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.

В базе данных корпуса хранятся текстовые документы с размеченными научно-техническими текстами. Разные виды разметок реализованы путем добавления новых тэгов в существующие документы.

Словарь корпуса содержит все терминологические единицы, содержащиеся в корпусе, а также контексты их употребления в научно-технических текстах в параллельном корпусе. Шаблон словарной статьи при добавлении новой терминологической единицы преставлен в таблице 5.1.

Таблица 5.1. Шаблон словарной статьи

Терминологическая единица		
	Русский	Английский
ID словарной статьи		
ID понятия		
Дефиниции		
C = <C_1, C_2, C_3, C_4>		
D = <D_1, D_2>		
E = {E_1, E_2, E_3, …}		
F = {F_1, F_2, F_3, …}		
G = {G_1, G_2, G_3, …}		
H = <H_1, H_2>		
Семантические падежи		
Глаголы, с которыми употребляется		
Свободные словосочетания		
Контексты		

C = <C1, C2, C3, C4> – набор описываемых типов концептуальных объектов, в котором различаются четыре типа:

•C1 сущность: материальные и нематериальные объекты, способы их рассмотрения;

•свойства C2: количественные, качественные, реляционные (отношения);

•действия C3: операции, процессы, состояния;

•значения (размеры) C4: время, положение, пространство.

D = <D1, D2> – пара наборов свойств понятия, где D1–набор качественных свойств; D2–набор количественных свойств.

E = {E1, E2, E3, …} – совокупность понятий, описывающих методы/функции, свойственные данному понятию, и отражающих прагматику, связанную с данным понятием.

F = {F1, F2, F3, …} – множество синонимов понятий, или, другими словами, множество понятий, имеющих количественные отношения (отношение тождества) с данным понятием.

G = {G1, G2, G3, …} – множество коррелятов, или, другими словами, множество понятий, имеющих отношение оппозиции к данному понятию.

H = <H1, H2> – пара множеств понятий, имеющих качественные отношения с данным понятием, где:

•H1 – набор понятий, образующих отношение обобщения с данными <H11, H12>, H11– родовые понятия, H12 – набор видовых понятий;

•H2 – набор понятий, составляющих отношение агрегации с данными <H21, H22>, H21 – понятие, которое является «целым» по отношению к описываемому, H22 – набор понятий, которые являются частью описываемого.

Корпусный менеджер представляет собой систему поиска в параллельном корпусе, которая обрабатывает запрос пользователя, и выводит все контексты употребления лексических единиц из запроса, а также позволяет обрабатывать сам конкорданс.

Снятие многозначности поискового запроса при поиске в корпусе. На практике пользователю обычно требуется найти не какой-то конкретный,

заранее известный документ, а некие сведения (факты), знание которых необходимо для решения поставленной задачи (или же для удовлетворения любопытства). Возникающая в данном случае проблема заключается в том, что пользователь имеет очень ограниченное представление о той информации, которую ему нужно найти. Самым надежным способом составления поискового предписания представляется включение в поисковый образ запроса ключевых слов (или словосочетаний), которые, по мнению пользователя, непременно должны входить в текст документа, содержащего нужные сведения. Однако здесь возникает следующая дилемма: если включить в поисковый запрос небольшое количество «наиболее вероятных» слов, то его результатом будут сотни (а то и тысячи) документов, далеко не все из которых будут содержать ответ именно на поставленный вопрос. Если же включить в запрос много «предполагаемых» ключевых слов (или даже целую фразу), то существует риск получить на выходе пустое множество документов, поскольку авторы документов требуемой тематики могли описывать интересующий пользователя предмет фразами, несколько отличающимися от заданной в запросе [248].

Практически каждое слово естественного языка, кроме строго однозначных терминов, особенно принадлежащее общеупотребительной лексике, многозначно. Понятие «многозначность» довольно неопределенно и зависит от контекста употребления слова. Можно различать экстралингвистический контекст, характеризующий общую обстановку межъязыковой коммуникации. Лингвистический контекст также достаточно общее понятие. Можно различать контекст всего текста в целом, отдельной его части, от страницы до абзаца, и, наконец, ближайший контекст в виде непосредственно соседствующих с основным словом лексических единиц [249].

Под контекстом принято понимать семантико-грамматическое и коммуникативное единство определенного текстового элемента с текстовым и ситуативным окружением как индикатором значения и функционального веса этого элемента. Различают микро- и макроконтекст. Микроконтекст – ближайшее окружение текста. Так, предложение получает смысл в контексте

абзаца, абзац в контексте главы и т.д. Макроконтекст – это вся система знаний, связанная с предметной областью, то есть знания об особенностях и свойствах, явно не указанных в тексте. Другими словами, любое знание обретает смысл в контексте некоторого метазнания [27].

Основными задачами, стоящими перед поисковыми системами, остаются выдача пользователям конкретных ответов на поставленные вопросы и определение сайтов с наилучшей выдачей ответов. Построение оптимальной последовательности применения тех или иных инструментов на каждом шаге поиска и предопределяет его эффективность.

С общим ростом количества информации и развития общества в целом, многие слова постоянно приобретают все новые и новые значения. Таким образом, ведя поиск по слову «Эос», пользователь получает поисковую выдачу из источников, относящихся как минимум к пяти разным областям знаний: медицина, мифология, образование, программное обеспечение, криптовалюта. Ответом на запрос является набор всевозможных значений данного сочетания букв, огромное количество информации из разных источников, каждый из которых нужно изучить, на что тратятся и время, и силы. Стоит отметить, что пользователь погружен в контекст, т.е. некоторую предметную область. Так, при чтении книги о легендах и мифах Древней Греции, искомое значение для слова «Эос», будет из предметной области мифологии.

Предлагаемый метод к разрешению лексической многозначности можно описать следующим образом. На вход поисковой системы поступает поисковой запрос пользователя. Поисковая система обращается к поисковому тезаурусу для поиска запроса. Если лексическая единица из поискового запроса является многозначной или имеет омонимы, то поисковая система предложит пользователю перечень предметных областей, в которых была найдена лексическая единица из поискового запроса. Зачастую пользователь заранее ищет результат из определенной предметной области. Когда предметная область определена, поисковая система определяет ближайшие элементы в структуре

онтологии, и при ранжировании поисковой выдачи будет ориентироваться на их наличие или отсутствие.

В качестве примера предположим, что пользователь ввел запрос «эос». На первом этапе проводится процедура поиска лексической единицы из поискового запроса «эос», что показано на Рис. 5.3.

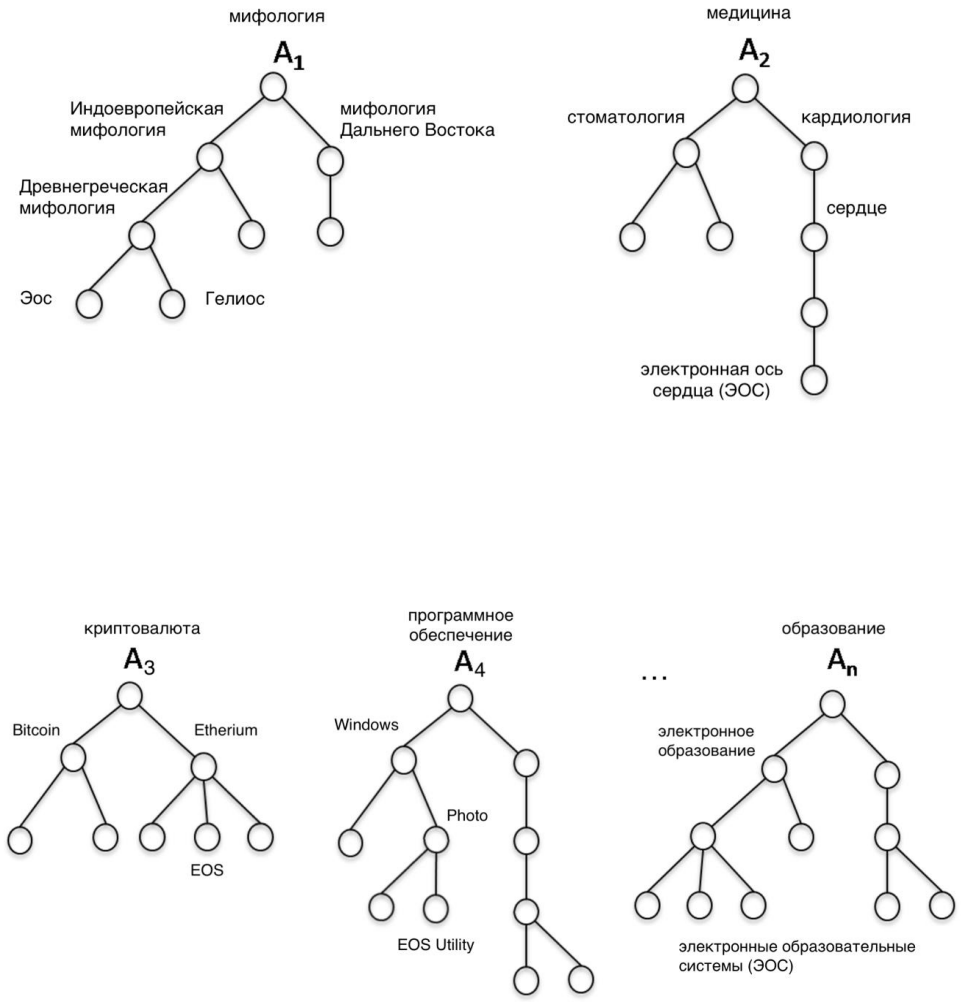


Рис. 5.3. Предметные области тезауруса

Результаты поиска представлены на Рис. 5.4. Как видно из примера, лексическая единица «эос» встречается в предметных областях по мифологии, медицине, крипотовалютах, программном обеспечении, образовании.

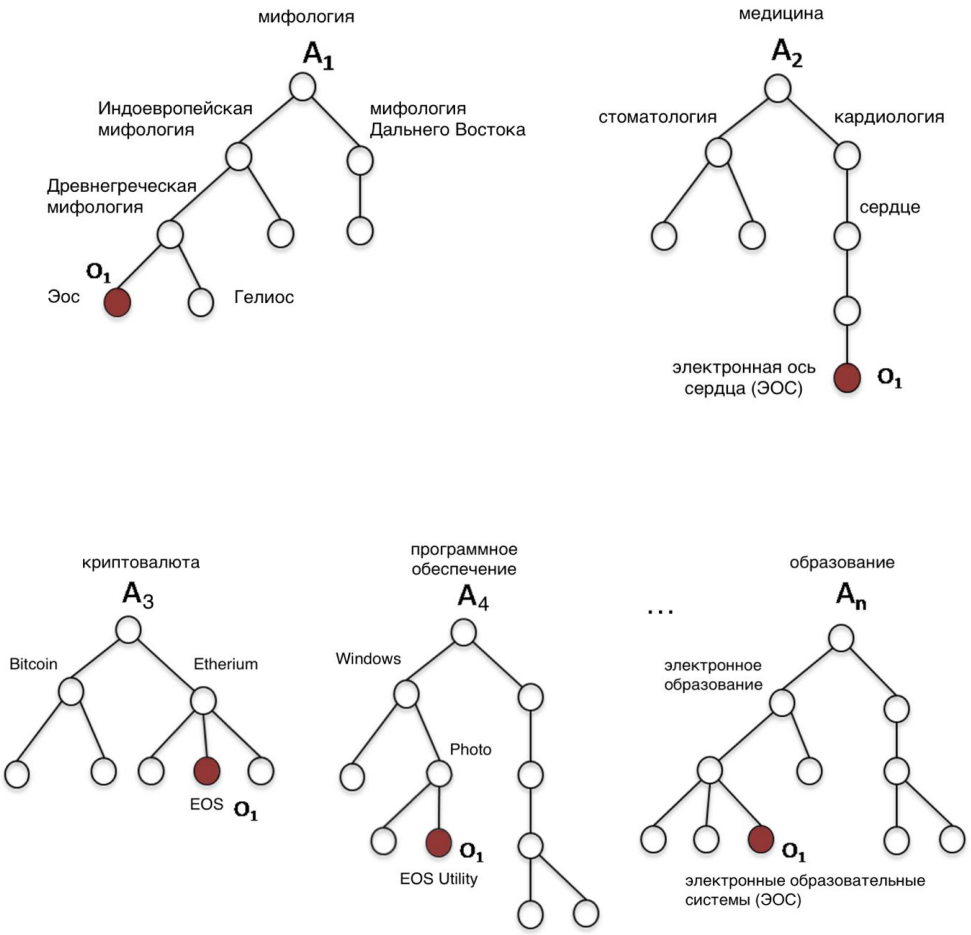


Рис. 5.4. Поиск в тезаурусе

На следующем шаге пользователь выбирает интересующую его предметную область, затем, в уже выбранной предметной области определяются ближайшие концепты, связанные с запросом пользователя. Найденные концепты служат критерием релевантности при ранжировании поисковой выдачи. На Рис. 5.5 представлена операция выборки лексических единиц для уточнения поискового запроса.

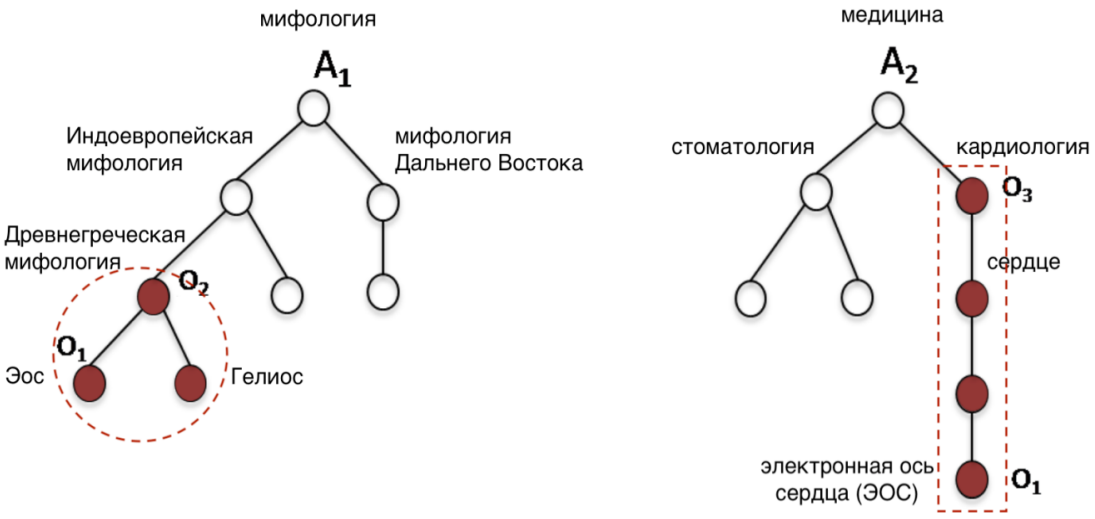


Рис. 5.4. Выбор понятий

Основные этапы подхода к информационному поиску на основе тезауруса представлены на Рис. 5.6.

Рис. 5.6. Основные этапы подхода к информационному поиску

Таким образом, если поисковой запрос был в области медицины, то поисковая система будет ориентироваться при ранжировании результатов на

наличие ближайших синонимических лексических единиц, таких как «кардиология», «сердце», «сердечная ось сердца», и при этом игнорировать документы, в которых встречаются «Гелиос», «древнегреческая мифология».

К преимуществам использования тезаурусы при информационном поиске также можно отнести возможность учета нескольких лексических единиц, обозначающих одинаковое понятие, например, полную форму сокращения «эос» - «электронная ось сердца».

Использование предложенного подхода позволит разрешить лексическую многозначность и существенно разгрузить поисковую выдачу, оставив лишь интересующую пользователя предметную область.

5.2. Информационная технология обработки научно-технических текстов в параллельном корпусе

Дж. Люгер выделяет уровни обработки естественного языка: графематический, морфологический, синтаксический, семантический, прагматический уровни и фоновые знания. При этом языковые уровни взаимосвязаны таким образом, что неоднозначность на нижнем уровне разрешается за счет привлечения информации с более высоких уровней [250].

Первые два уровня, как показал анализ параллельных корпусов, в которых одним из языков является русский, успешно реализованы на практике. Разметка многокомпонентных терминов должна опираться на результаты морфологической разметки, а сама являться основой для разметки номенклатурных наименований. Разметка семантических ролей не может быть реализована в научно-технических текстах, если не определены границы многокомпонентных терминов и номенклатурных наименований. Представление языковых объектов как системы взаимосвязанных компонентов позволит отражать все языковые особенности текстов, 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 систем обработки текстов на

разных языках одновременно.

Последовательность этапов обработки научно-технических текстов обоснована тем, что некоторые виды разметки используются как основа для других более сложных разметок. Так, морфологическая разметка является основой для терминологической разметки, в основе которой лежат структурные модели многокомпонентных терминологических единиц, а структурная разметка позволяет отсечь те структурные элементы научно-технических текстов, где нет терминов или их обработка значительно перегружает список терминов-кандидатов.

Разметка научно-технических текстов при добавлении в параллельный корпус на основе разработанных в исследовании моделей и методов происходит в несколько этапов, представленных на рис. 5.7.

Рис. 5.7. Этапы обработки научно-технических текстов в параллельном корпусе

На вход системы поступает научно-технический текст и его переводная

версия. На первом этапе осуществляется метаразметка текстов – приписывание экстралингвистических меток к каждому обрабатываемому тексту. К таким экстралингвистическим меткам можно отнести: жанр текста, название, автор, год издания, издательство и другие характеристики текста в зависимости от его жанра. Также обязательно необходимо указать язык оригинала, то есть язык, на котором написан текст, а также язык перевода. Данные метатекстовой разметки о жанре текста дают представление о его базовой структуре, наборе обязательных их факультативных элементов текста.

Морфологическая разметка осуществляется на основе библиотеки Pymorphy2, где каждой лексической единице приписываются все возможные тэги о ее морфологических характеристиках. Этап терминологической разметки реализуется на основе результатов морфологической разметки и подразумевает разметку многокомпонентных терминов и номенклатурных наименований в параллельном корпусе.

Стилистическая разметка служит для выделения русскоязычных машинных текстов и их фрагментов в корпусе, а также может быть использована как маркер синтаксической сложности научно-технических текстов, принадлежащих к разным научным направлениям. Кроме того, может представлять ценный ресурс для переводоведения благодаря возможности выявлять наличие переводческих трансформаций при переводе текстов с одного языка на другой.

Этап семантической разметки подразумевает возможность разметки синтаксических конструкций, по своему формальному представлению меньших, чем предложение, но выражающих одно значение, клише, или же отражающее особенности семантико-синтаксических связей компонентов предложения.

Дополнительно предусмотрена возможность анафорической разметки и разметки переводческих трансформаций, поделенных на три основных типа: лексические, грамматические и лексико-грамматические.

Этапы терминологической, стилистической и структурной разметок осуществляются на основе моделей и методов, разработанных в предыдущих

главах. Однако требование филологической корректности корпуса подразумевает постоянный контроль результатов автоматической разметки лингвистами, а при необходимости дальнейшей фиксации и устранения возникающих ошибок при обработке научно-технических текстов в параллельном корпусе.

5.3. Программные средства разметки и выравнивания научно-технических текстов в параллельном корпусе

Структурная разметка и выравнивание научно-технических текстов в параллельном корпусе. При выполнении структурной разметки в тексте на двух языках выделяются основные структурные элементы: название текста, информация об авторе, аннотация, ключевые слова, введение, основной текст, заключение и библиография (Рис. 5.8 и 5.9).

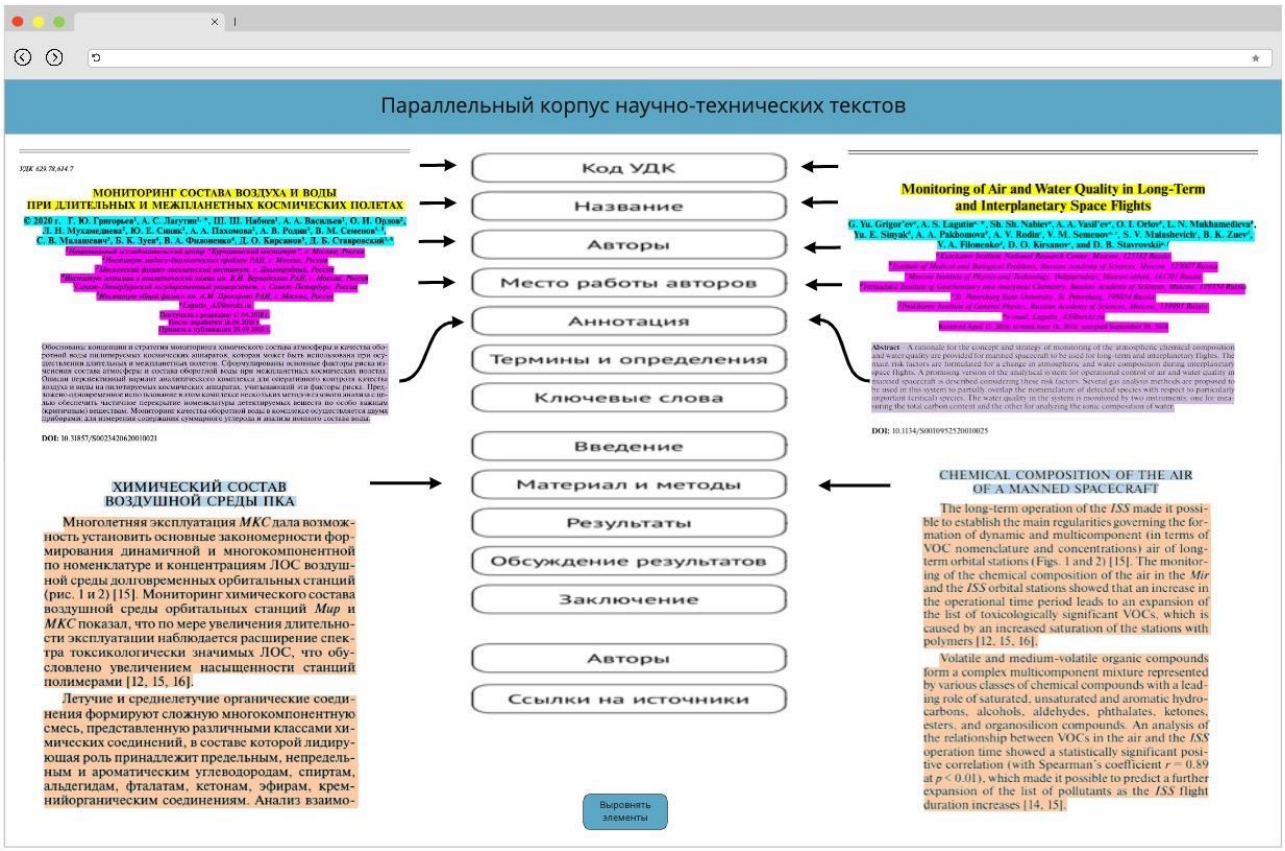


Рис. 5.8 – Визуализация работы системы структурной разметки и выравнивания

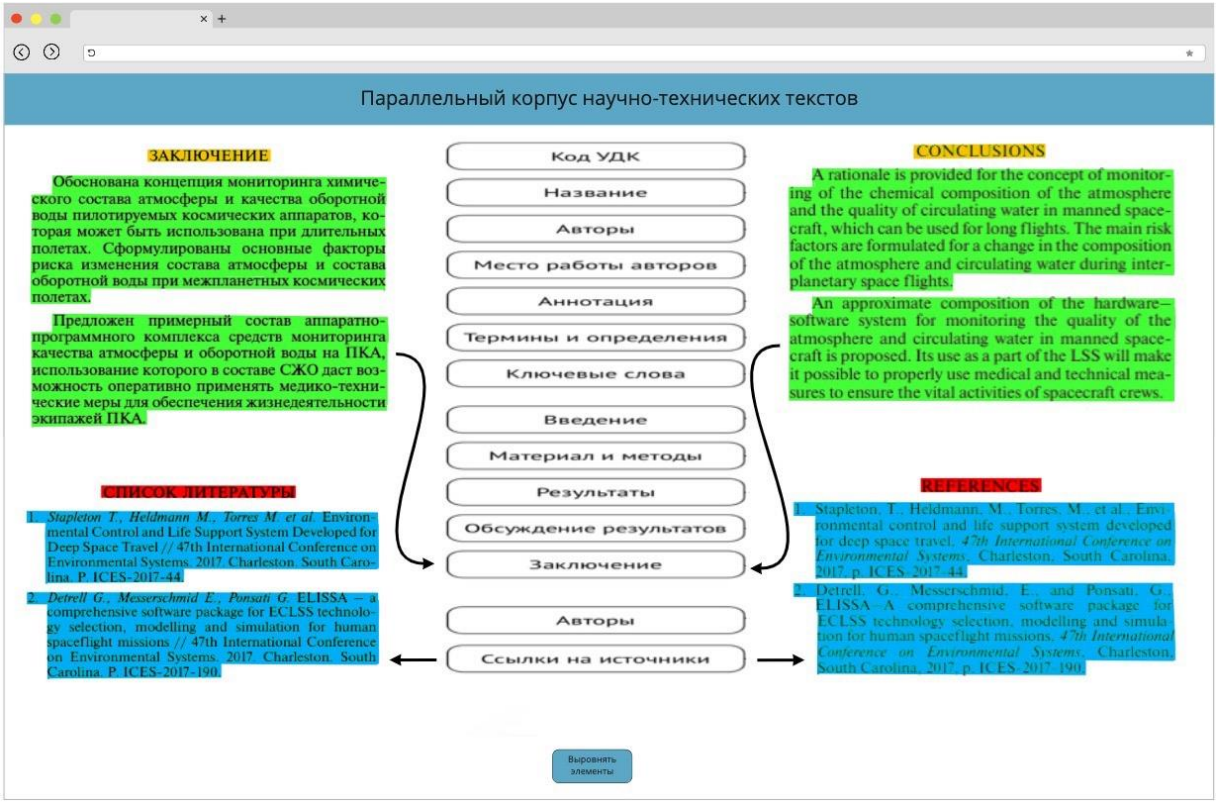


Рис. 5.9 – Структурная разметка

Посередине страницы располагается перечень структурных элементов, которые соединяются со структурными элементами в двух текстах при помощи стрелок.

Кроме того, в больших по объему стурутерных элементах можно выделять отдельные элементы. Так, например, научно-техническая статья содержит в себе три способа изложения текста: описание, повествование и рассуждение, которые отличаются способом изложения материала, а также элементами связности текста. Во введении можно выделить отлельными тегами обзор современного состояния проблемы - хронологический обзор текущего состояния знаний о каком-либо явлении, который предлагает направления для будущих исследований. Анализ текущего состояния развития некоторый области знаний во введении – это описание, изложение хода проведения эксперимента в части «Материалы и методы» относится к повествованию, а в части «Обсужение результатов» авторы используют тип изложения материала – рассуждение.

Таким оброзам, получаем несколько способов изложения текста: описание,

рассуждение, повествование, аннотацию и заключение, что может иметь критическое значение при создании наборов данных для машинного обучения.

При выполнении терминологической разметки размечающий выделяет одним цветом соответствующие друг другу термины в двух языках. Особое внимание уделено многокомпонентным терминам, перевод которых представляет определенные сложности.

Разметка и выравнивание многокомпонентных терминов в параллельных научно-технических текстах. Для извлечения и последующей разметки многокомпонентных терминов из англо- и русскоязычных текстов, разработано приложение, написанное на языке Python с использованием библиотек tkinter, nltk, pymorphy2, os, sys и других. Программа предоставляет пользователю интерфейс для ввода текстов, подлежащих анализу. Из полученных текстов путём морфологического анализа слов, которые в него входят, а также их взаимного расположения, выделяются словосочетания, которые могут являться многокомпонентными терминами.

Далее пользователь может вручную отобрать термины, классифицировать их и сохранить в базу данных, которая состоит из нескольких текстовых файлов, с которыми работает программа и которые можно анализировать с помощью функции поиска. Она предоставляет интерфейс для поиска терминов, соответствующих определённому набору характеристик.

На рис. 5.10 представлен алгоритм работы системы извлечения многокомпонентных терминов из параллельных научно-технических текстов, представленных в виде IDEF0-диаграмм третьего уровня декомпозиции разработанного алгоритма.

Интерфейс программы для извлечения многокомпонентных терминов из параллельных научно-технических текстов представлен на рис. 5.11. Работа в программе начинается в выборе способа ввода текстов: загрузка в виде файлов в форматах doc, docx, txt, pdf или ручной ввод текстов на английском и русском языках.

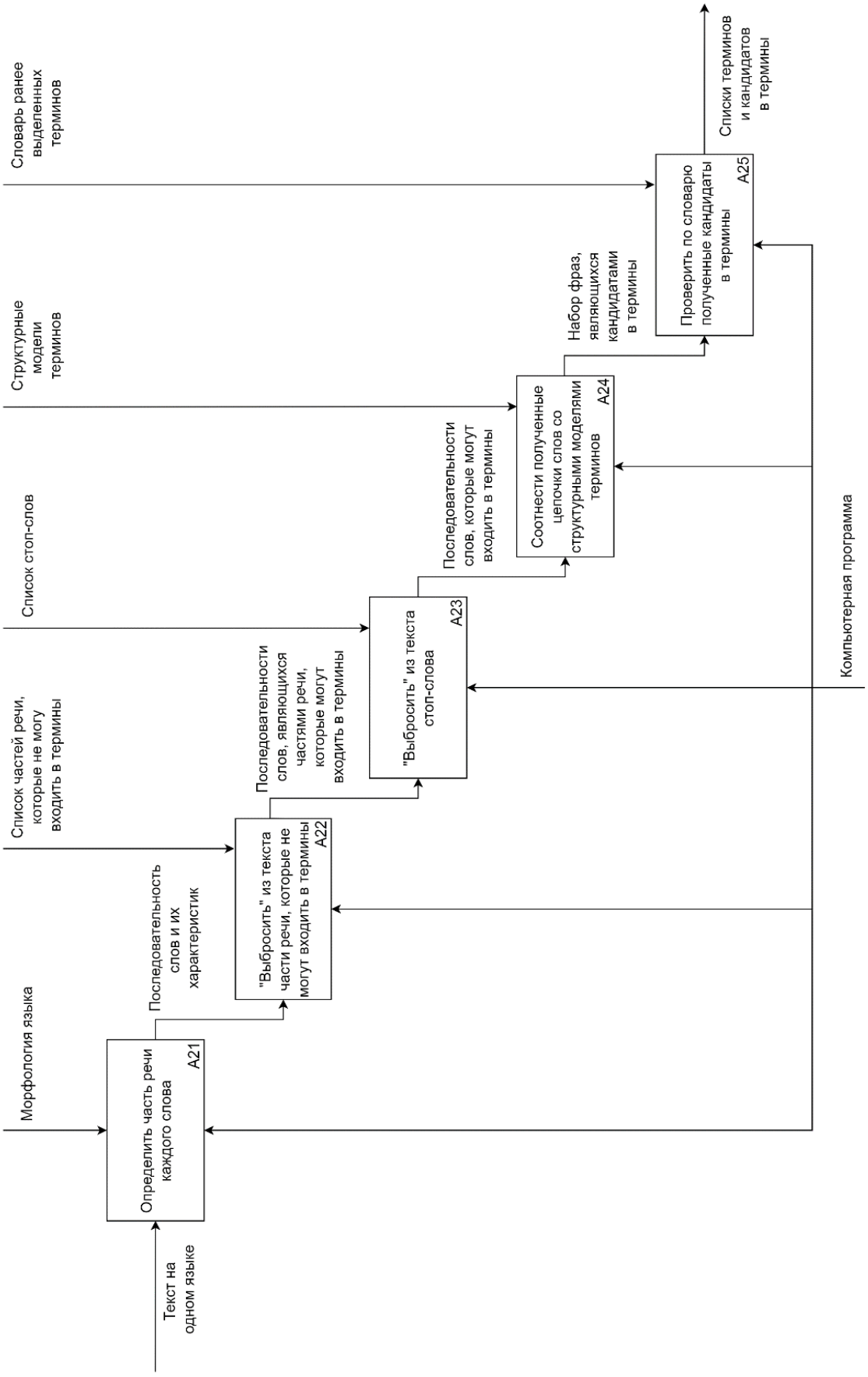


Рис. 5.10. Алгоритм работы системы извлечения многокомпонентных терминов из параллельных научно-технических текстов

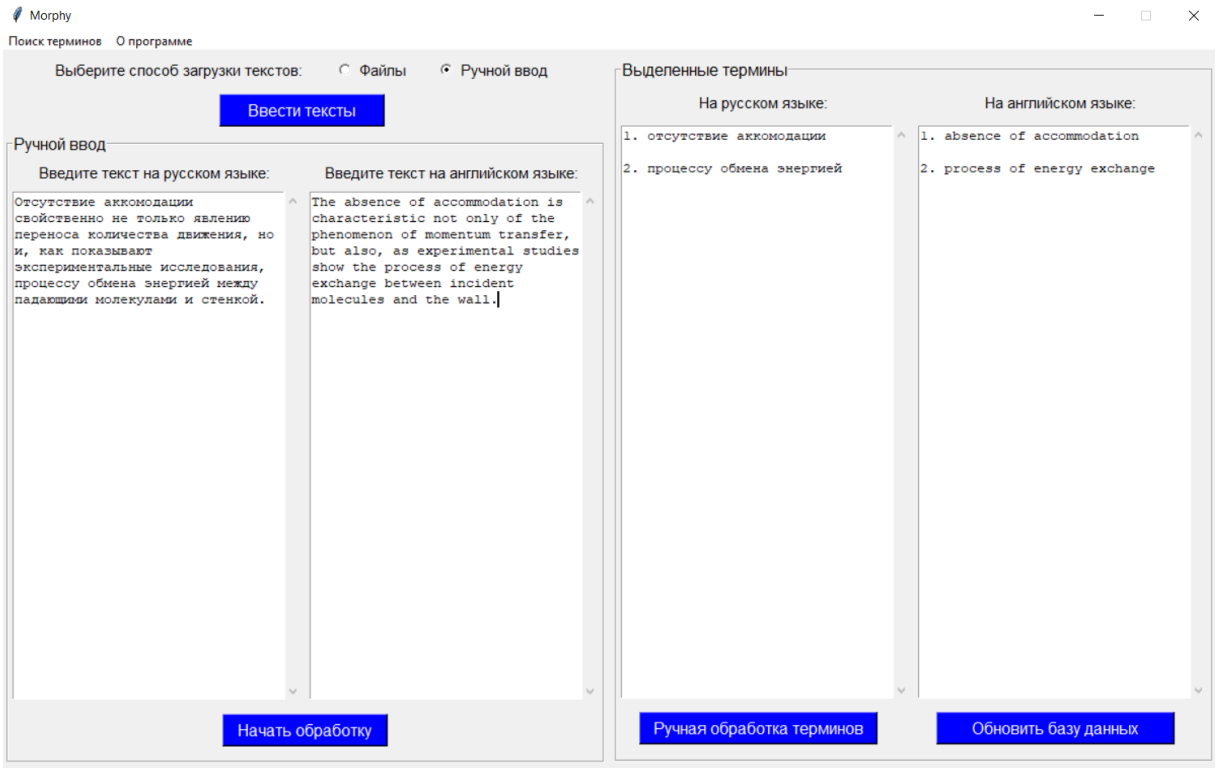


Рис. 5.11. Интерфейс системы извлечения многокомпонентных терминов из параллельных научно-технических текстов

В программе также реализована функция выявления количества повторений терминов в каждом тексте – повторяющийся по тексту термин будет выведен один раз, а количество его употреблений указывается в скобках рядом с термином. Извлечение терминов может производиться как с двух языков, так и с одного языка. После ввода данных следует нажать “Начать обработку”.

По окончании обработки текстов откроется диалоговое окно, представленное на Рис 5.12, где пользователю будет дана возможность вручную выбрать требуемые терминологические единицы, которые ранее не были сохранены в словарь корпуса.

После выделения терминологических единиц пользователь может приписать им лексические и грамматические характеристики, и в последствии сохранить в базу данных. Результаты отбора терминов появятся в окошке, представленном на Рис.13. На основе грамматических и лексических характеристик можно проводить поиск терминов в базе данных по определенным параметрам.

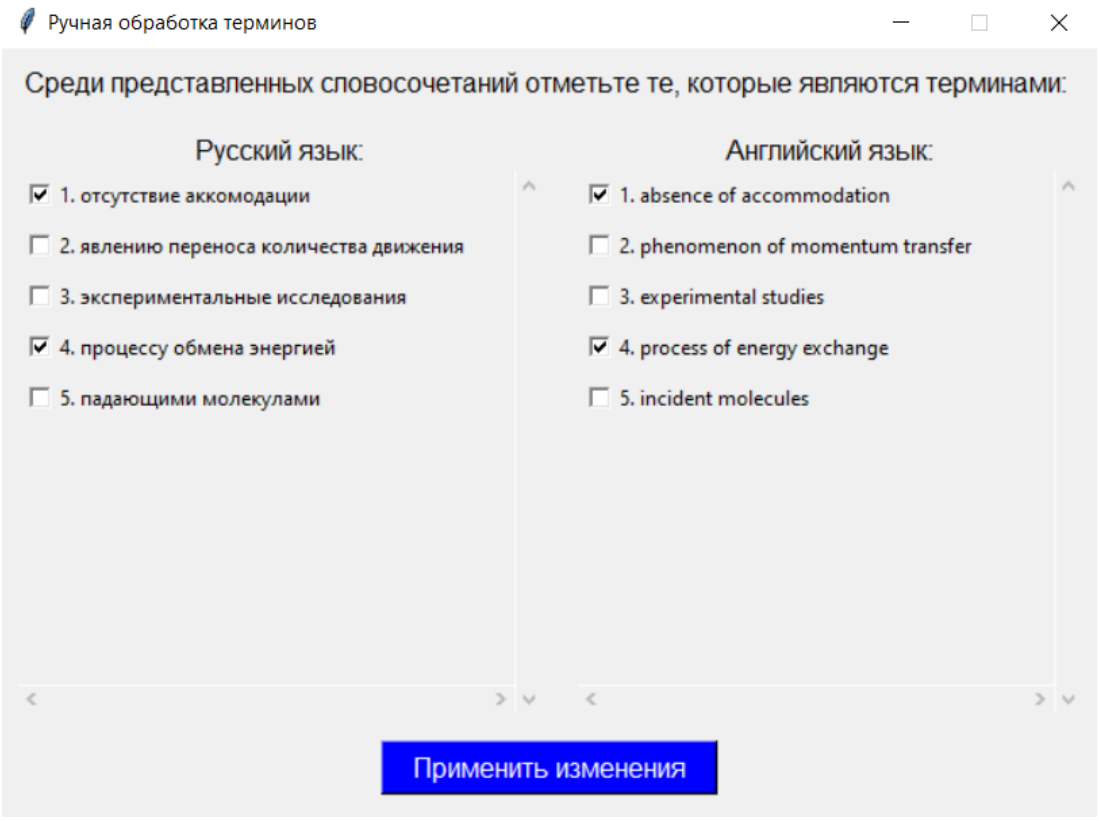


Рис. 5.12. Диалоговое окно выбора терминологических единиц

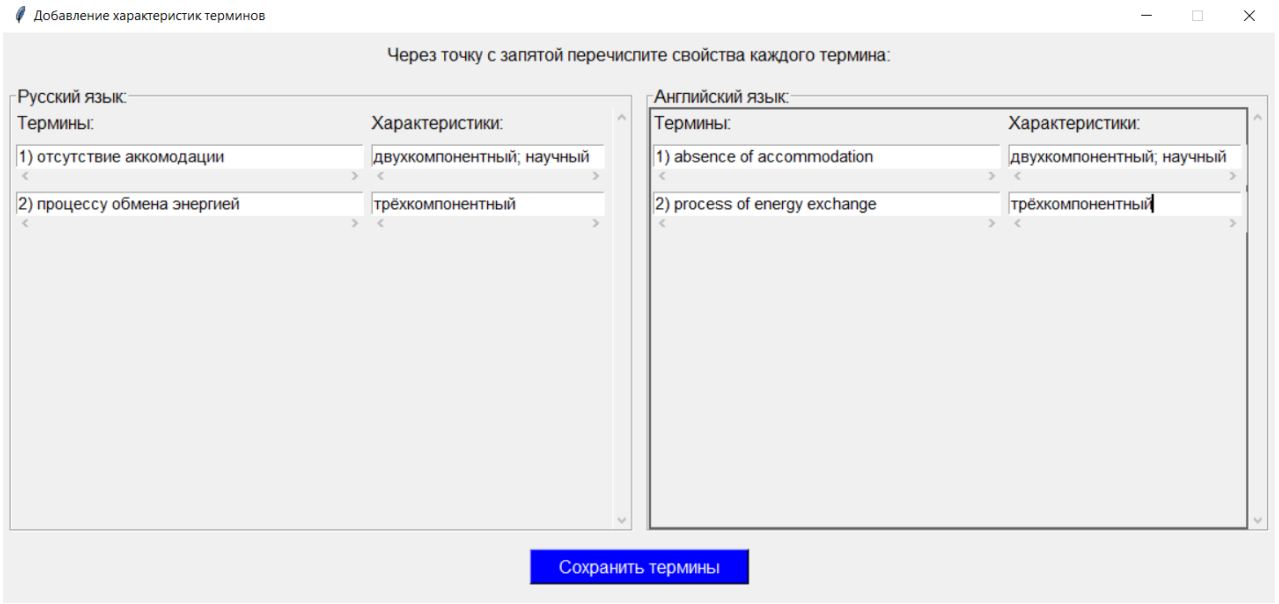


Рис.5.13. Окно ввода лексической и грамматической информации к терминам

Рассмотрим возможности практического использования предложенной системы извлечения многокомпонентных терминов из текста научно-

технической статьи по космонавтике и ее переводной версии. Предположим, что целью является исследование формальной структуры терминов в английском и русском языках. Результаты такого исследования представлены в таблице 5.1.

Таблица 5.1. Результаты анализа формальной структуры терминов научно-технической статьи по космонавтике

Морозов В. М. Каленова В. И. Управление спутником при помощи магнитных моментов: управляемость и алгоритмы стабилизации // Космические исследования. 2020 №3. С. 199-207	рус. яз.	англ. яз.
Всего слов в тексте	1037	1404
Найдено терминов-кандидатов всего	285	302
Найдено терминов вручную	56	35
Из них **однокомпонентных**	11	6
RU: спутник, магнитометр, орбита, время, работоспособность, эффективность, матрица, нуль, столбцы, моделирование, коэффициент EN: satellite, magnetometers, system, controllability, time, dimensions		
Из них **двухкомпонентных**	32	13
RU: круговая орбита, линеаризованная система, уравнение движения, магнитная система, малый спутник, магнитная катушка, космический аппарат, магнитный момент, линеаризованная модель, управляющий момент, геомагнитное поле, нестационарная система, стационарная система, математическое моделирование, гравитационное поле, система координат, уравнение движения, орбитальная система, плоскость орбиты, направляющий косинус, единичный вектор, угол наклона, плоскость орбиты, плоскость экватора, угол поворота, линеаризованное уравнение, управляемое движение, система уравнений, определитель матрицы, независимые столбцы, матрица коэффициентов, квадратичный функционал EN: LQR method, magnetic systems, magnetic coils, control moment, floquet theory, stationary system, mathematical modeling, gravitational field, circular orbit, coordinate systems, OY-axis, orbital plane, independent columns		
Из них **трехкомпонентных**	10	10
RU: линейная нестационарная система, собственный магнитный момент, внешнее магнитное поле, линейная нестационарная система, приближенная стационарная система, замкнутая периодическая система, главная центральная ось, постоянное магнитное поле, линейная обратная связь, стационарная управляемая система EN: magnetic orientation system, intrinsic magnetic moment, external magnetic field, spacecraft attitude control, angle of inclination, equations of motion, determinant of matrix, time-varying system, moments of inertia, resulting time invariant		
Из них **четырехкомпонентных**	0	2
RU: - EN: stabilization of relative equilibrium, linearized system of equations		

Также была проведена ручная проверка результатов работы системы

извлечения многокомпонентных терминов, которая показала наличие ошибок при обработке научно-технических текстов (Таблица 5.2).

Таблица 5.2. Анализ ошибок системы извлечения многокомпонентных терминов из параллельных научно-технических текстов

Морозов В. М. Каленова В. И. Управление спутником при помощи магнитных моментов: управляемость и алгоритмы стабилизации // Космические исследования. 2020 №3. С. 199-207		рус. яз.	англ. яз.
Всего действительных терминов		56	35
Всего терминов, не извлеченных системой		3	4
Результат программы	Термин	Описание проблемы и возможное решение	
1. Снабженная магнитная системой	Магнитная система ориентации	Программа, возможно, определила, что трехкомпонентный термин должен включать в себя именно прилагательное «снабженная», а не существительное «ориентации» по модели прил+прил+сущ, данную модель стоит поменять на прил+сущ+сущ.	
2. Function of geomagnetic field	Geomagnetic field	Были присоединены ненужные существительное «Function» и предлог «Of», возможно ошибка в модели, стоит создать четкую модель двухкомпонентного термина прил+сущ	
3. Original nonstationary system	Nonstationary system	Идентичная ситуация с пунктом 4, из модели прил+прил+сущ следует убрать первое прилагательное	
4. ЛКР	ЛКР	Программа не определяет аббревиатуры как термин, следует добавить соответствующую модель	
5. LQR	LQR	Идентичная ситуация с пунктом 4.	
6. Однородная система с постоянными коэффициентами	Линейная однородная система	Программа ошибочно определила термин, не определилось нужное прилагательное и добавились 3 ненужные части речи, в модели пятикомпонентного термина присутствует ошибка, оттуда нужно исключить предлог	
7. MATLAB	MATLAB	Программа не определяет аббревиатуры как термин, следует добавить соответствующую модель	

Говоря о самих результатах, можно выделить некоторые положительные и отрицательные моменты. Из текста на русском языке программа довольно хорошо извлекает термины, автоматически преобразовывая их ядерные

элементы терминов в именительный падеж единственного числа, однако не распознает аббревиатуры как в русском тексте, так и в английском. Также программа часто определяет только часть многокомпонентного термина, возможной причиной чего является то, что в базе нет определенной модели.

Разметка семантических ролей в параллельном корпусе научно-технических текстов. Программа реализована на языке Python 3.8+, а интерфейс программы – на QT. Программа позволяет создавать слой семантической разметки для любого загруженного текста путём разметки предложений по семантическим ролям. В программе доступен режим просмотра разметки на основе общей базы данных. В программу включена следующая терминология:

1. предикат – ядро семантической конструкции (пропозиции);

2. актант – значимый участник ситуации, заполняющий валентность предиката;

3. семантическая роль – обозначение совокупности участников ситуации, зависящих от значения предиката;

4. сочетание – пронумерованное множество слов, относящихся к одной семантической роли (например, терминологические словосочетания);

5. связь – пара сочетаний типа «предикат-аргумент»;

6. предложение – поиск отдельных лексем, сочетаний и семантических групп в контексте (указывается правый и левый контексты). Все элементы могут существовать только как часть предложения;

7. начальная форма – лексема, включающая в себя реализации всех словоформ в размеченном тексте.

Программа реализована в двух разных режимах: анализе и навигации, как представлено на рисунках 5.14-5.15.

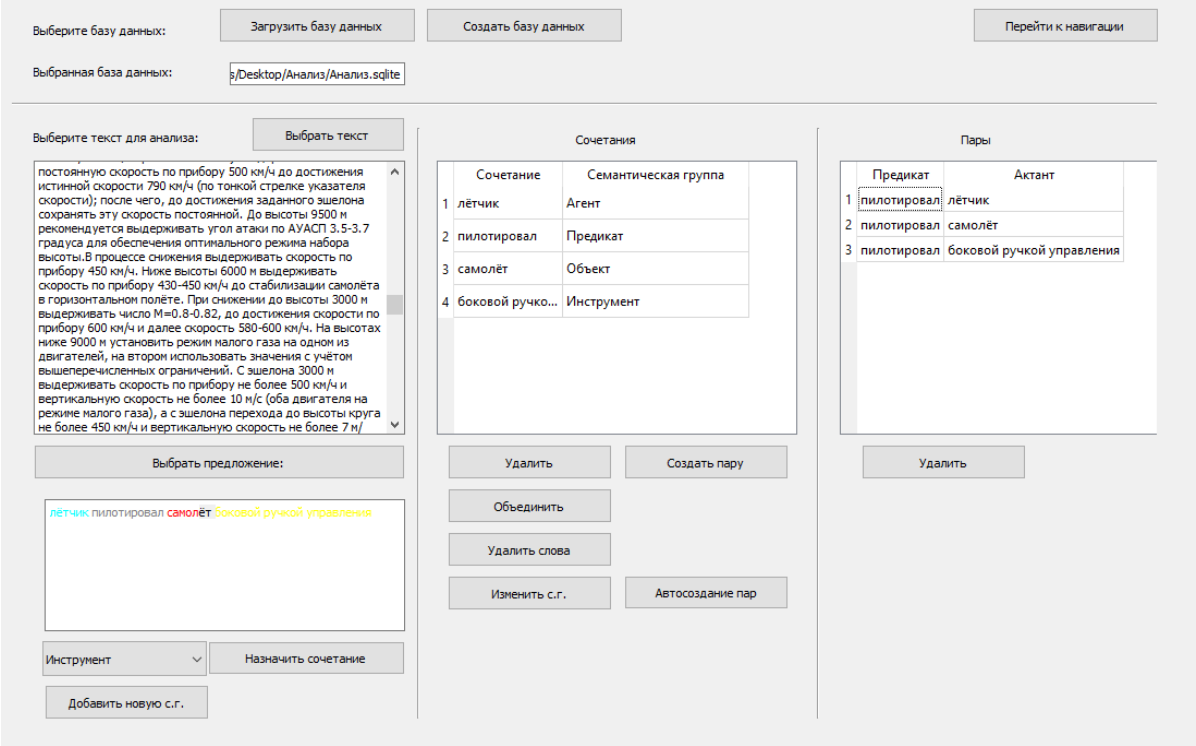


Рисунок 5.14 – Интерфейс программы в режиме разметки.

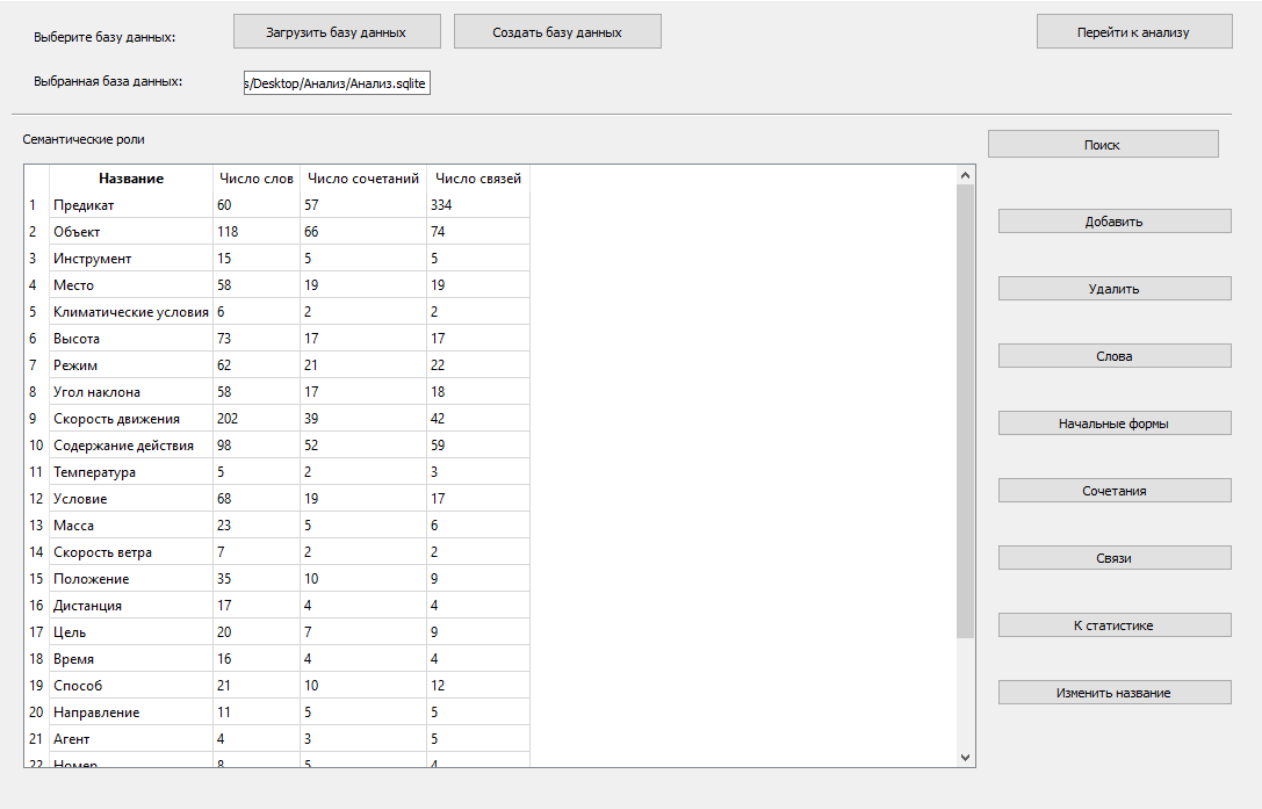


Рисунок 5.15 – Интерфейс программы в режиме навигации.

Режим анализа используется для осуществления самой разметки, а режим навигации включает в себя модуль статистического анализа и предоставляет

следующую информацию о размеченных данных:

1. количество размеченных слов;

2. количество размеченных лексем (начальные формы слова);

3. количество размеченных предложений;

4. количество словосочетаний (в том числе, терминологических словосочетаний);

5. количество связей предиката с различными аргументами;

6. количество полученных в результате разметки семантических ролей.

Режим навигации реализуется через несколько взаимосвязанных таблиц, содержащих данные по разметке. Библиотека «Pymorphy2», загруженная в исходный код программы, автоматически осуществляет лемматизацию (приведение всех размеченных слов к начальной форме) и определяет часть речи. Таким образом, программа осуществляет элементы морфологического анализа для предоставления более подробной информации о размеченных лексических единицах, которая также содержится во взаимосвязанных таблицах. Для работы программе требуется подключение к файлу базы данных, куда будут сохранятся размеченные данные, и через который можно получить к ним доступ. Ниже представлены возможные запросы к базе данных:

1) получить все лексемы или терминологические словосочетания, реализованные в одной семантической роли;

2) получить все возможные пары «предикат-аргумент» с выбранным предикатом или аргументом;

3) получить все предложения, в которых встречается выбранная семантическая роль, слово или словосочетание;

4) получить список всех семантических ролей, в которых реализуется выбранное слово или словосочетание и др.

В ходе анализа можно создавать неограниченное количество баз данных с текстами для разметки. Предметная направленность текста не влияет на качество разметки, поэтому в программе можно создать несколько баз данных для различных областей знания. В таком случае перед началом разметки может

понадобиться лишь удаление из списка некоторых семантических ролей, автоматически загруженных в программу для разметки текстов по авиации и космонавтике и характерных только для данной предметной области. Разметку можно производить с разных устройств и одновременно несколькими участниками. По завершении разметки полученные базы данных можно совместить в одну. Программа также сохраняет всю информацию о своей работе в папку пользователя, благодаря чему можно получить доступ к содержимому последнего созданного файла в случае ошибки в функционировании программы.

Режим анализа. В процессе работы с программой в первый раз необходимо создать базу данных, в которую будет сохранятся вся информация о разметке. При повторной работе с программой можно открыть уже существующую базу данных с помощью кнопки «Загрузить базу данных» и продолжить работу с ней, или же снова создать новую базу при необходимости с помощью кнопки «Создать базу данных», как показано на рисунке 5.16. По строке с путём к базе данных можно определить, какая база открыта в данный момент. Последняя открытая база данных сохранится, и при следующем открытии программы она будет загружена автоматически. При нажатии на кнопку переключения режимов (разметка/навигация) меняется основная часть окна, в которой происходит работа.

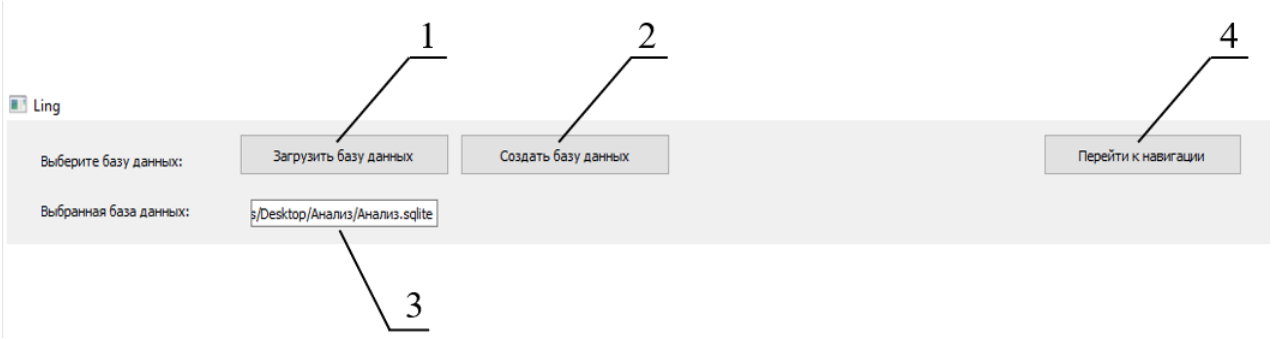


Рисунок 5.16 – Выбор / создание базы данных: 1 – выбрать существующий файл базы данных; 2 – создать новый файл базы данных; 3 – строка с путём к базе данных; 4 – переключение режимов работы с программой

После создания базы данных необходимо загрузить в неё текст или

несколько текстов для разметки в формате «txt». В ходе анализа в один файл базы данных будут загружаться размеченные данные, а в другом файле будут храниться тексты для разметки. В режиме анализа для работы с текстом требуется загрузить его из ранее созданной базы данных с помощью кнопки «Выбрать текст», как показано на Рис.5.17.

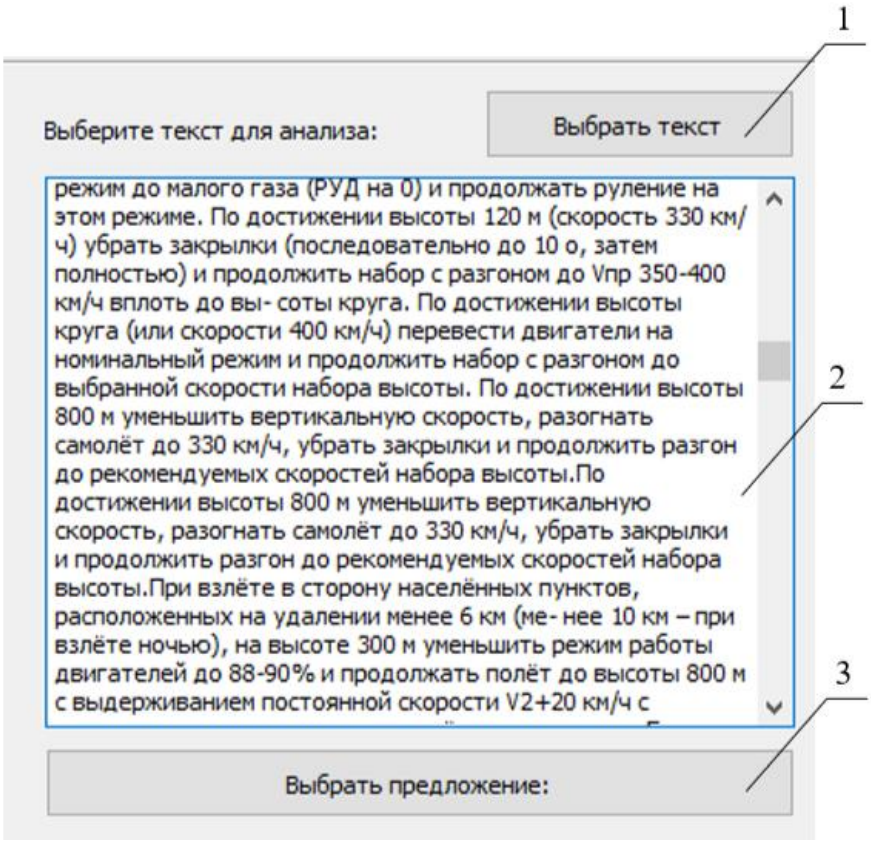


Рисунок 5.17 – Выбор текста и предложения для разметки: 1 – выбор файла с текстом; 2 – просмотр содержимого файла; 3 – выбор предложения в тексте

Выбранный текстовый файл целиком отображается в левом крайнем окне интерфейса. Редактировать выбранный текст в программе нельзя, поэтому перед загрузкой необходимо предварительно подготовить его к разметке: удалить лишние предложения, проверить наличие ошибок/опечаток, при необходимости расшифровать аббревиатуры, а также удалить из предложений лишние знаки препинания, которые могут расцениваться программой как идентификаторы окончания предложения, например, в сокращениях «шт.» или «т.ч.».

Предложением считается последовательность символов, разделенных точкой, вопросительным или восклицательным знаком. Если в середине предложения встречается один из перечисленных знаков препинания, система примет его за сигнал о конце предложения. Предложение будет обработано некорректно и в том случае, если в нём пропущен один из перечисленных знаков препинания. После того, как текст был подготовлен к разметке и загружен из базы данных в программу, можно приступить к разметке предложений. Необходимо поместить курсор в предложение, которое требуется выбрать и нажать кнопку «Выбор предложения», при этом целиком выделять предложение необязательно (Рис. 5.18).

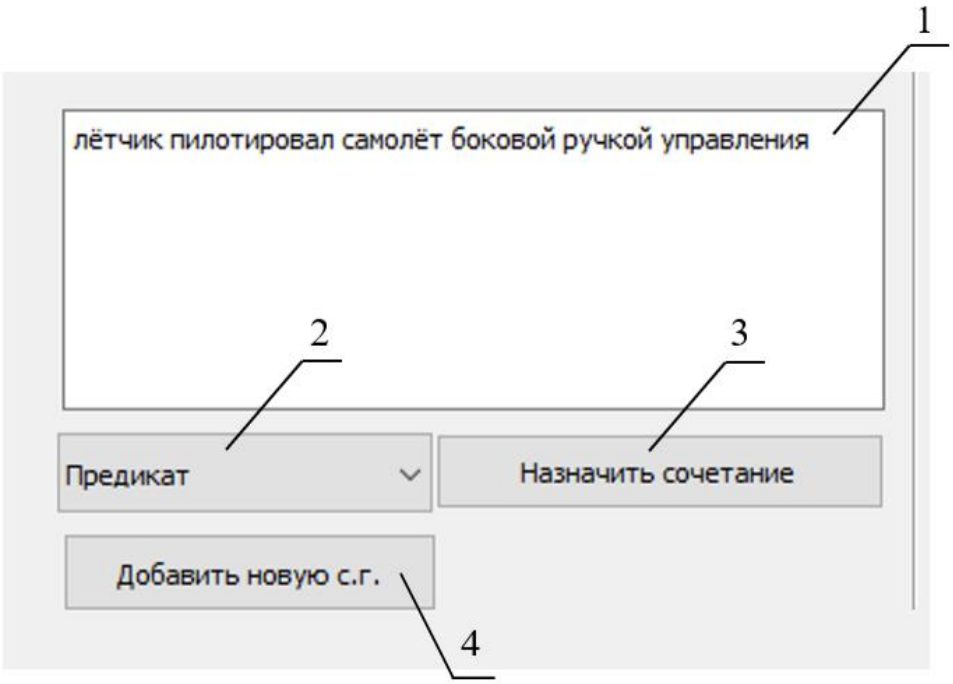


Рисунок 5.18 – Режим работы с предложением. Назначение сочетаний: 1 – просмотр предложения; 2 – выбор семантической группы; 3 – добавление сочетания; 4 – добавление новой семантической группы

После выбора предложения можно перейти к этапу назначения сочетаний типа «предикат-аргумент». В первую очередь необходимо выделить предикат путём наведения на него курсора или выделения всего слова. Далее, из выпадающего ниже меню следует выбрать роль «предикат» и нажать кнопку

«назначить сочетание». Предикат автоматически сохранится в окне разметки. Назначение аргументов осуществляется тем же способом: требуется выделить необходимый аргумент или навести на него курсор, выбрать для него подходящее название семантической роли из выпадающего меню и сохранить с помощью клавиши «Назначить сочетание». Если такой роли не существует, ее можно создать с помощью кнопки «Добавить новую с.г.». Функция создания семантической роли позволяет добавлять любые семантические роли и, соответственно, расширять инвентарь семантических ролей. Все размеченные семантические роли подсвечиваются различными цветами, как показано на рисунке 5.19. Когда предикат и все его аргументы выделены, можно перейти к работе в окне «Сочетания».

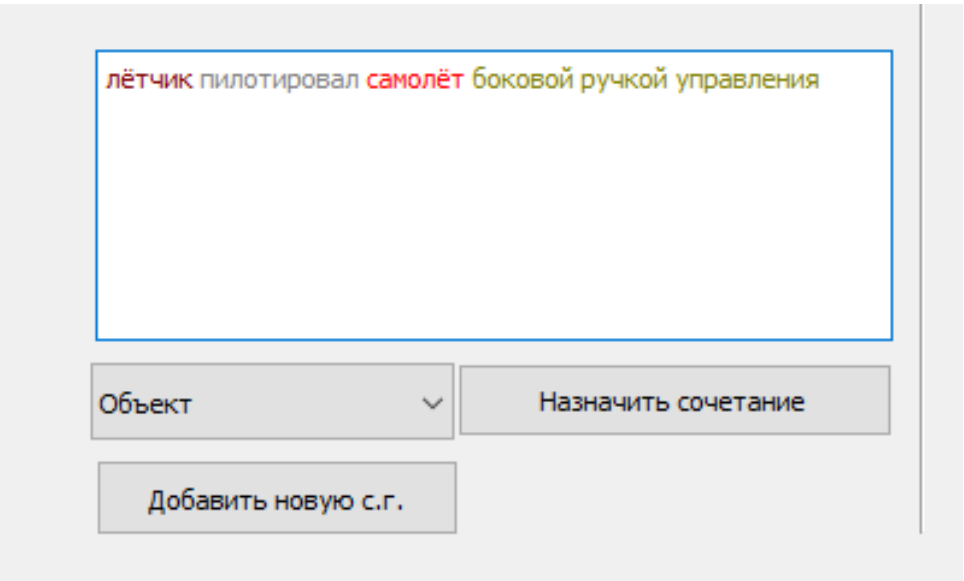


Рисунок 5.19 – Пример размеченного предложения

Работа с сочетаниями большей частью направлена на возможное исправление неточностей в разметке, что позволяет устранить какие-либо ошибки в разметке до того момента, как разметка предложения будет сохранена в базу данных. Кнопка «удалить» стирает из базы данных сочетание и все связи с ним. Кнопка «объединить» создает новое сочетание из всех выбранных слов в разделе «сочетания». Кнопка «удалить слова» позволяет выборочно удалить

слова из словосочетания, если они были размечены и соотнесены с предикатом некорректно. Кнопка «Изменить с.г.» позволяет изменить тип связи, например, «предикат-агент» на «предикат-инструмент». Кнопка «создать пару» объединяет в пару сочетания «Предикат-актант», если предикатов несколько и их нужно соотнести с определёнными аргументами, а кнопка «автосоздание пар» в предложении с одним предикатом автоматически создает все связи между выбранным предикатом и актантами, как показано на рисунках 5.20 и 5.21. После нажатии на кнопку «создать пару» или «автосоздание пар», все данные о размеченном предложении и всех выбранных парах «предикат-аргумент» сохраняются в базу данных.

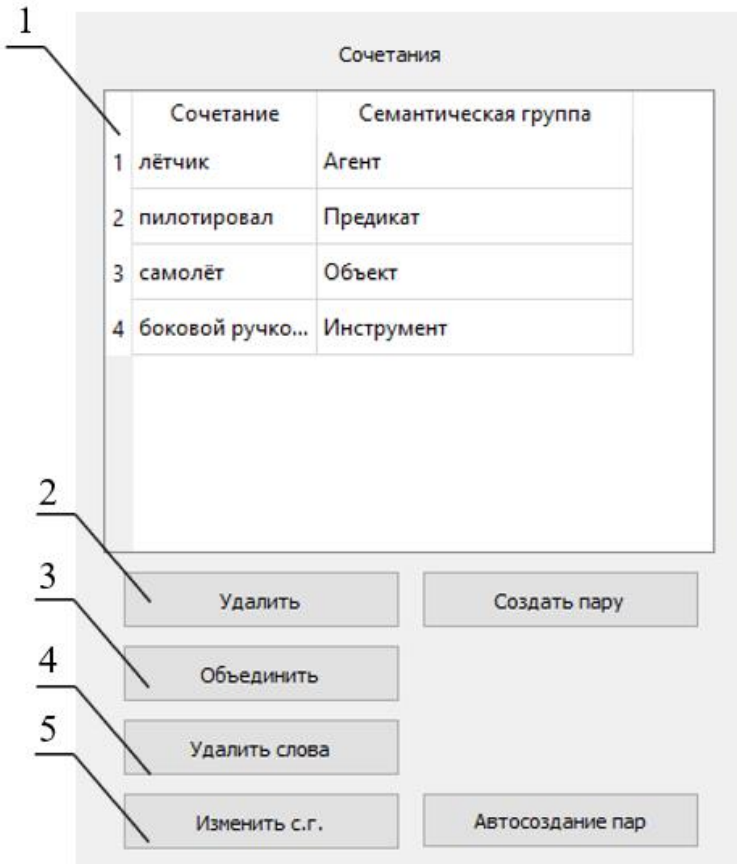


Рисунок 5.20 – Работа с предложением. Просмотр сочетаний: 1 – таблица просмотра сочетаний; 2 – удаление выбранных сочетаний; 3 – объединение слов из списка выделенных сочетаний с выбором новой семантической роли; 4 – удаление выделенных слов; 5 – изменение семантической группы выбранного сочетания

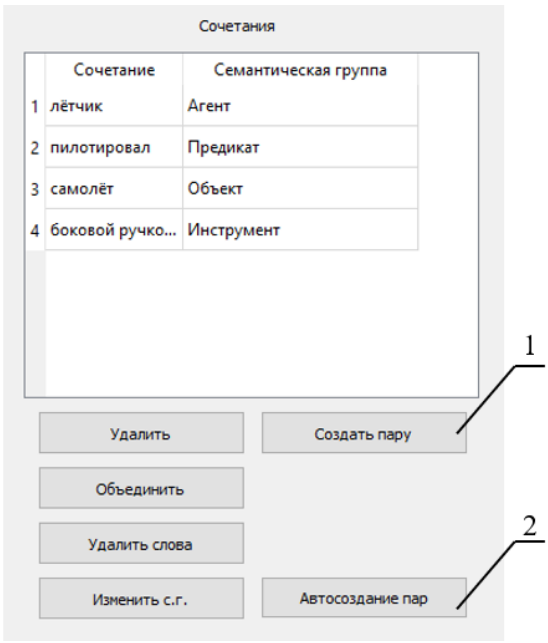


Рисунок 5.21 – Режим работы с размеченными сочетаниями: 1 – создание пары из вручную выбранного предиката и актанта; 2 – автоматическое создание пар в предложении с одним предикатом

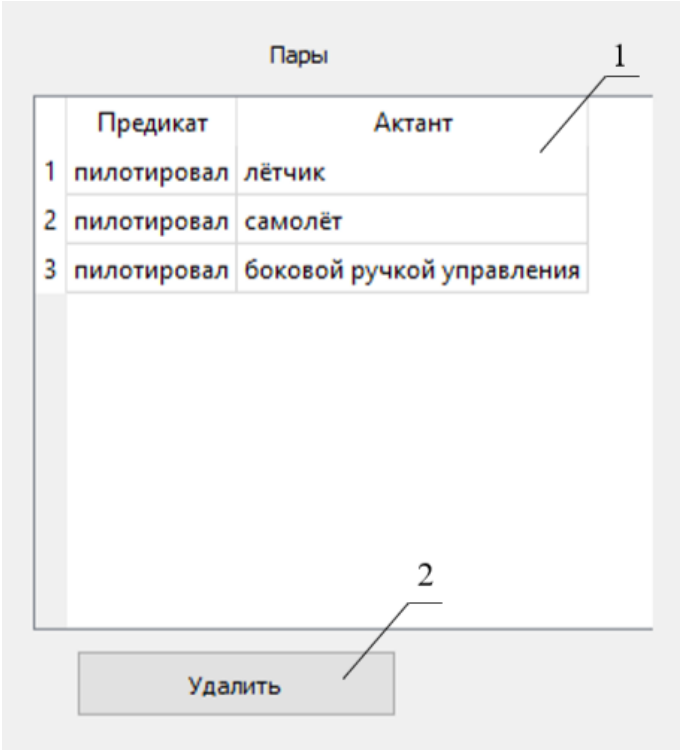


Рисунок 5.22 – Работа с предложением. Просмотр связей: 1 – просмотр размеченных сочетаний; 2 – удаление выбранной связи

Режим навигации. В режиме навигации предоставляется доступ ко всем

размеченным данным в форме нескольких взаимосвязанных таблиц, которые также содержат некоторые статистические данные о размеченных предикатах и аргументах. Раздел «Статистика» представляет собой основную таблицу, в которой можно посмотреть информацию о содержимом базы данных и перейти к другим таблицам. Между таблицами существуют аналогичные переходы, которые осуществляются через кнопки, содержащие названия самих таблиц. При нажатии на каждую из таких кнопок формируется новая структура, согласно типу кнопки и выделенным элементам таблицы. К примеру, при выделении двух предложений в таблице «Предложения» и нажатии на кнопку «Слова» сформируется таблица слов, которые входят в каждое из выделенных предложений. Кнопка «начальная форма» выведет все размеченные лексемы, то есть все начальные формы слов, хранящихся в базе данных. Если необходимо вывести только определённые лексемы, реализованные, например, семантической ролью «Инструмент», необходимо также вернуться к основной таблице «Статистика», вывести список семантических ролей, выделить среди них запись «Инструмент» и нажать на кнопку «Начальные формы».

Существуют кнопки, которые не являются частью переходов, например, кнопка «Поиск». Поиск реализован по лексеме, от которой можно перейти ко всем остальным структурам, с которыми она связана. Кнопки «Добавить» и «Удалить» позволяют создать новую или удалить уже имеющуюся семантическую роль. С помощью кнопки «Изменить название» можно отредактировать название для какой-либо семантической роли, например, заменить название «Погодные условия» на «Климатические условия». При переходе к списку предложений также появляется кнопка «Удалить», которая удалит выбранное предложение и все связанные с ним элементы из базы данных. С помощью кнопок «Перейти к анализу» и «Перейти к навигации» можно быстро переключаться между режимами навигации и анализа.

На рис. 5.23 визуализированы все виды разметок, реализованных в параллельном корпусе научно-технических текстов.

Структурная	Введение, абзац 2, предложение 1									Введение, абзац 2, предложение 2			
Стилистическая	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	wit h	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. 5.23. Визуализация слоев разметки и выравнивания научно-технических текстов

5.4. Использование параллельного корпуса в исследованиях по лингвистике и информатике

Параллельный корпус научно-технических текстов обладает широким исследовательским потенциалом для изучения качественных и количественных показателей научно-технических текстов, которые могут стать основой при решении практических задач в лингвистике и информатике. В качестве примеров приведены результаты исследований по влиянию структурных особенностей терминов на естественность научно-технических текстов, способов перевода модальных глаголов на основе параллельных текстов стандартов, а также использование корпуса как набора данных для разработки специализированного чат-бота.

В настоящее время одним из ключевых факторов ранжирования сайтов в поисковой выдаче является качество размещенных на нем текстов, которое оценивается не только пользователем сайта, но и поисковой машиной, у для которой разработаны собственные критерии качества текста и инструменты его оценки. Одним из таких критериев является естественность текста. Под естественностью принято понимать критерий качества текста, который определяет закономерность расположения слов, где частота слова обратно пропорциональна его месту в тексте. Проверку естественности текста проводят с помощью закона Ципфа, который можно сформулировать следующим образом: если для какого-нибудь довольно большого текста составить список всех слов, которые встретились в нем, а потом ранжировать эти слова в порядке убывания частоты их появления в тексте, то для любого слова произведение его ранга и частоты появления будет величиной постоянной:

$$c = f * r,$$

где f - частота встречаемости слова в тексте; $\boldsymbol{r}$ - ранг слова в списке; $\boldsymbol{c}$ - эмпирическая постоянная величина (коэффициент Ципфа).

Закон Ципфа гласит, что частота словесных лексем в большом корпусе естественного языка обратно пропорциональна рангу. Фактический график ранговой частоты текста на естественном языке в некоторой степени отклоняется от идеального распределения Ципфа, особенно на двух концах диапазона. Отклонения могут зависеть от языка, темы текста, автора, от того, был ли текст переведен с другого языка, и от используемых правил правописания

Для сравнительного анализа естественности текстов на основе закона Ципфа выбраны тексты художественного, публицистического и научного стилей, а результаты их проверки на естественность приведены в таблице 5.3.

Таблица 5.3. Естественность текстов разных стилей

	Стиль	Источник текстов	Естественность, %
1	художественный	главы из серии романов Джоан Роулинг о Гарри Поттере, роман И.С. Тургенева «Отцы и дети» и др.	65-80
			80
2	публицистический	30 текстов с сетевого издания РИА Новости	70-85
3	научный	30 статей из научной электронной библиотеки «КиберЛенинка».	32-78

На основе таблицы 5.3 видно, что тексты научного стиля имеют самые низкие показатели естественности, что можно попробовать объяснить широким использованием терминологической лексики. Многокомпонентные термины состоят из нескольких слов, что делает их использование более редким, чем однословные термины. В результате, эти термины вносят искажения в распределение частотности лексики и снижают точность анализа текста.

Стоит отметить, что при квантитативном анализе научно-технических текстов, учет частотности каждого слова в законе Ципфа приводит к неточностям при оценке частотности сложных терминов. Например, термин «центр масс» состоит из двух слов, но обозначает конкретное понятие в физике, которое не должно рассматриваться как два отдельных слова. Если данный

термин будет разделен на два слова - «центр» и «масса», то это будет означать уже два других понятия. Для уменьшения этого влияния можно использовать различные подходы, например, рассматривать многокомпонентные термины как отдельные единицы или применять дополнительную обработку, направленную на выделение и учет компонентов терминов.

Если разбиение текста на слова является простым вычислительным процессом, то разбиение на значимые фразы, к которым можно отнести и многокомпонентные термины, является более сложной задачей. Вполне очевидно, что можно попытаться использовать N-граммы, которые в настоящее время являются наиболее распространёнными и быстрыми методами разбора текста. Крупные массивы данных N-грамм стали широко доступны для анализа, в частности, в рамках проекта Google Books. К сожалению, все N-граммы не справляются с одной важной задачей: при подсчете они накладываются друг на друга, что скрывает базовые частоты слов. Следовательно, невозможно правильно назначить ранжируемые значения частоты употребления для N-грамм, объединенных по всем значениям N.

Для проверки гипотезы в влиянии многокомпонентных терминов на естественность текста выбраны 30 научных статей по космонавтике. Далее при помощи системы извлечения многокомпонентных терминов, описанной в главе из отобранных статей найдены многокомпонентные термины. Между элементами термина пробел заменен на нижнее подчеркивание, чтобы представлять их как одну единицу. Также все ядерные элементы многокомпонентных терминов приведены в начальную форму: единственное число, именительный падеж. Следующим шагом был анализ текстов в оригинальном виде и в изменённом. (Рис. 5.24 и 5.25) Естественность текста выросла с 72 до 84%.

Как показали результаты проверки оригинального текста и измененного, наблюдается повышение качества текста, а именно естественности текста.

Естественность вашего текста: 78%
Всего слов: 444
Отфильтровано стоп-слов: 112
Отображено: 42
Тошнота: 4.24

#	Слово	Вхождений	По Ципфу	Соответствие	Рекомендации
1	спутник спутника, спутник, спутников, спутником	18	18	100%	
2	система системы, система, системе, системами	17	9	53%	-8
3	ось осями, ось, оси, осью	15	6	40%	-9
4	движение движением, движение, движения	13	5	39%	-8
5	гравитационный гравитационной, гравитационный, гравитационного	9	4	45%	-5
6	центр центра, центр, центре	8	3	38%	-5
7	момент моментом, момент, моментов, момента	8	3	38%	-5
8	закон закон, закона, законы	7	3	43%	-4
9	земля земли, земля, землей	7	2	29%	-5
10	масса масс, масса	7	2	29%	-5
11	электромагнит электромагнитами, электромагнит, электромагнитах, электромагнитов	7	2	29%	-5

Рис. 5.24 Естественность текста научно-технической статьи по космонавтике

Обработка терминов позволила более точно определять уникальные термины и уменьшить повторяемость их отдельных частей при анализе по закону Ципфа. Также это позволило более точно оценить влияние многокомпонентных терминов на общую частотность слов и получить более точные результаты при анализе текста. Все эти процессы были проведены с целью улучшения качества анализа текста и получения более точной информации о частоте употребления слов. Результаты обработки позволили получить более точное распределение частотности слов в тексте. Также результаты анализа на закон Ципфа показали, что наиболее частотные слова являются ключевыми в тексте.

Естественность вашего текста: 82%

Всего слов: 406
Отфильтровано стоп-слов: 112
Отображено: 34
Тошнота: 4.12

#	Слово	Вхождений	По Ципфу	Соответствие	Рекомендации
1	спутник спутника, спутник, спутников, спутником	17	17	100%	
2	ось осями, ось, оси, осью	15	9	60%	-6
3	система система, системы, системе, системами	12	6	50%	-6
4	движение движением, движение, движения	10	5	50%	-5
5	закон закон, закона	6	4	67%	-2
6	земля земли, земля, землей	6	3	50%	-3
7	гравитационный гравитационной, гравитационный, гравитационного	6	3	50%	-3
8	режим режима, режим, режиме	6	3	50%	-3
9	угол угла, угол, углами	6	2	34%	-4
10	электромагнит электромагнитами, электромагнит, электромагнитах, электромагнитов	6	2	34%	-4
11	момент моментом, момент, момента	5	2	40%	-3

Рис. 5.25 Естественность предварительно обработанного текста научно-технической статьи по космонавтике

Итак, термины, состоящие из двух и более слов, имеют большое влияние на анализ текста на закон Ципфа. Правильное понимание значений терминов является ключевым для понимания контекста и смысла текста в целом. При анализе текста с использованием терминов, необходимо учитывать их специальное значение и использовать специализированные методы анализа текста. Такой рост в показателе естественности текста позволяет приблизиться к «идеальному» распределению Ципфа. Результаты показали, что после обработки многокомпонентных терминов происходит существенное улучшение равенства, определяющего закон Ципфа, и снижение их влияния на общую частотность

лексики. дополнительная обработка многокомпонентных терминов позволила получить более точную информацию о частоте употребления слов и повысить качество анализа текста. Как предполагалось ранее, слитное написание терминов уменьшает количество отдельных слов в тексте, что в свою очередь может упростить анализ на закон Ципфа. Эти методы могут быть важными инструментами при анализе текстов, содержащих многокомпонентные термины.

Сравнительный анализ способов перевода модальных глаголов с русского на английский языки проводился на материале 5130 предложений, содержащих средства выражения модальности в оригинальном тексте. 1500 предложений отобрана из текстов научно-технических статей, 2630 примеров употребления и перевода модальных глаголов в параллельных учебных текстах, а также 1500 – из текстов стандартов. В таблице 5.4 приведены результаты частотного анализа употребления модальных глаголов в англоязычных научно-технических текстах.

Таблица 5.4. Употребление модальных глаголов в англоязычных научно-технических текстах

	Научно-технические статьи		Учебники и учебные пособия		Стандарты	
	кол-во	%	кол-во	%	кол-во	%
must	65	4,33	49	1,86	60	4,00
have to	61	4,07	50	1,90	59	3,93
to be to	168	11,20	183	6,96	58	3,87
shall	0	0	0	0	506	33,73
should	171	11,40	131	4,98	236	15,73
need	61	4,07	179	6,81	93	6,20
can	535	35,67	975	37,07	137	9,13
could	51	3,40	89	3,38	5	0,33
to be able	45	3,00	64	2,43	55	3,67
will	157	10,47	396	15,06	67	4,47
would	0	0	185	7,03	0	0
may	147	9,80	246	9,35	222	14,80
might	39	2,60	83	3,16	2	0,13
ought to	0	0	0	0	0	0
Всего	1500		2630		1500	

На таблицы 5.4. можно увидеть ряд особенностей употребления модельных глаголов в научно-технических текстах разных жанров. Во-первых, в научно-технических текстах не используется модальный глагол *ought to*, во-вторых, самым частнотным глаголом в текстах стандартов является глагол *shall,* однако не было выявлено ни одного случая его употребления в текстах научно-технических статей и учебной литературы. В-третьих, самым частотным модальным глаголом в этих текстах является глагол *can*.

Не менее интересные результаты получились при анализе способов перевода модальных глаголов на русский язык. Из проанализированных 39 примеров употребления модального глагола *might* в научно-технических статьях, не было выявлено вариантов его перевода, то есть в 100% он не передается средствами русского языка. Также выявлено, что при переводе текстов учебников и учебных пособий на русский язык глагол *need* приобретает такие значения, как «*необходимо*» - 48%, «*следует*» - 9%, «*должен*» - 7%, «*понадобиться*» - 6%, «*не обязательно*» - 6%, «*требовать*» - 5%, «*мочь*» - 3%, «*приходиться*» - 2%, «*(не) нужно*» - 2%, «*не обязан*» - 2%, или вовсе *не переводится* – 10%. Стоит обратить внимание на отклонение переводных эквивалентов модального глагола «*need*» от зафиксированных в словарях и выявленных при анализе фактических переводов.

Проанализированы 150 примеров модального глагола «*can*» в текстах стандартах на примере нормативной базы программной инженерии.

*Sometimes it takes time to implement a chosen set of controls and during that time the level of risk may be higher than **can** be tolerated on a long-term basis.*	*Иногда требуется время, чтобы внедрить выбранный набор средств управления, и в течение этого периода времени уровень риска может быть выше, чем он **должен** быть в долгосрочной перспективе.*

Можно заметить, что в данных примерах модальный глагол «*can (not)*» переводится на русский язык в значении «*(не) мочь*» - 75%, «*можно ли*» - 1%, «*не удается*» - 5%, которые соответствуют своим подлинным значениям.

Модальный глагол «*could*» переводится на русский язык в значении «*могли бы*» - 5%. Модальный глагол «*can*» приобретает значения, не соответствующие своей исходной значении, таких как, «*должен*» - 3%, или вовсе не передается модальность путём перевода на русский язык– 2%. Модальный глагол «*can*» переводится, как «*могли*» - 2%, «*cannot*» - «*не могли*» во времени Past Simple - 1%. На рисунке 5.26 представлена схема соответствия способов выражения модальности в английском и русском языках на примере нормативной базы программной инженерии.

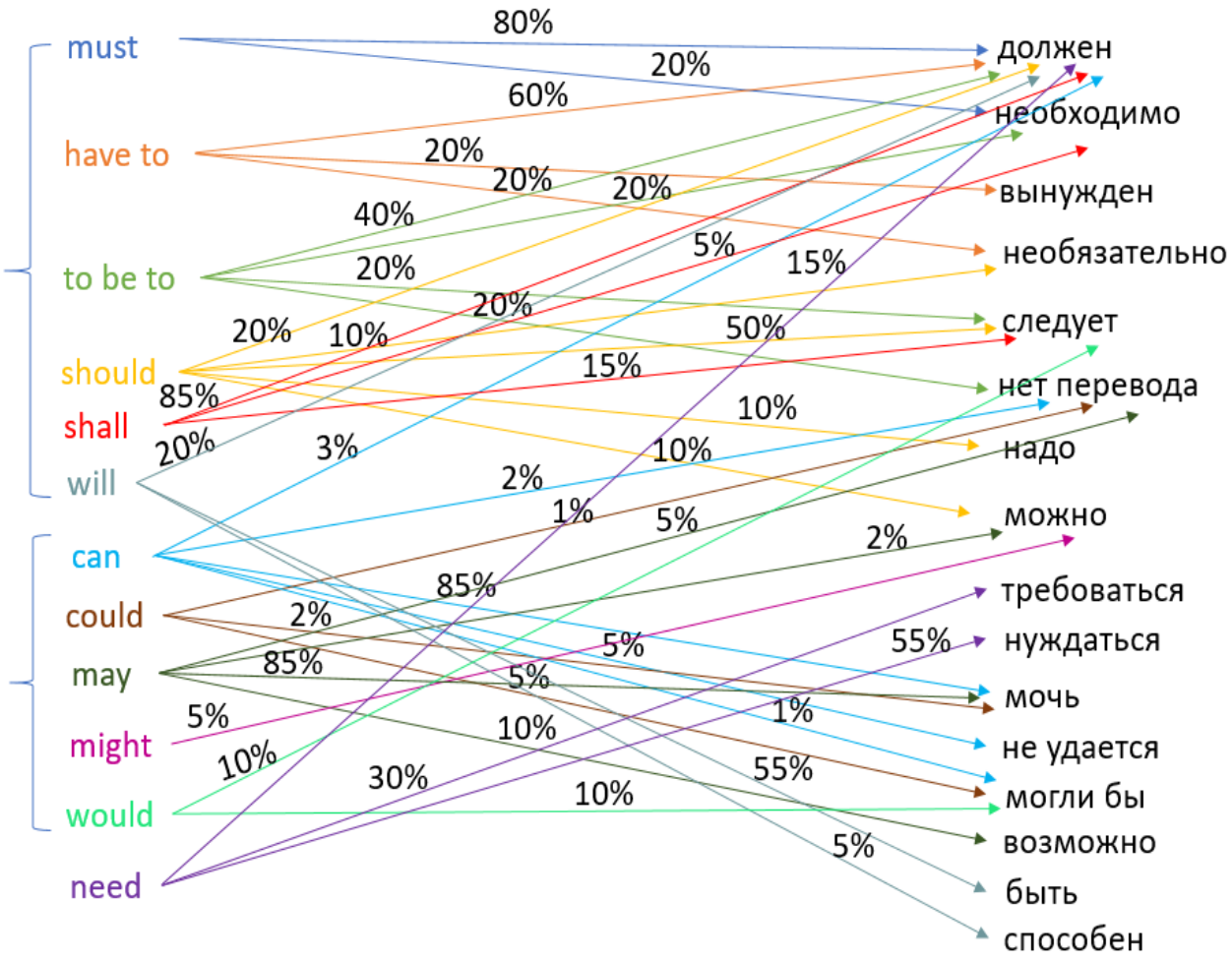


Рис. 5-26. - Схема соответствия способов выражения модальности в английском и русском языках на примере нормативной базы программной инженерии

5.5. Выводы по главе

Разработан прототип параллельного корпуса научно-технических текстов

на английском и русском языках. Предложен новый подход к организации системы обработки корпусных данных.

Получены следующие результаты:

1. Разработана концепция построения системы управления корпусными данными, которая дает возможность значительно расширить круг пользователей корпусом за счет расширения функционала корпуса, а также возможности доступа не только к самим параллельным текстам, но и результатам их обработки. Реализована возможность автоматической и автоматизированной разметки научно-технических текстов. Такие функциональные возможности позволяют использовать параллельный корпус при решении задач по автоматической обработке текстов, построению математических моделей языка и целого ряда другие задач.

2. Предложена информационная технология обработки научно-технических текстов, где каждый новый уровень разметки текстов основан на результате предыдущего уровня: структурная, морфологическая, терминологическая разметки, разметка номенклатурных наименований, семантико-синтаксическая разметка, разметка машинных текстов.

3. Разработана система разметки многокомпонентных терминов и номенклатурных наименований из параллельных англо- и русскоязычных научно-технических текстов. Разработанная система позволяет значительно сократить время на обработку терминологии в научно-технических текстах, а ее словарь при постоянном наполнении является основой для создания многоязычной терминологической базы данных.

4. Разработана система обработки семантико-синтаксических структур в научно-технических текстах. На примере разметки семантических ролей показано, что каждая предметная область может иметь свой уникальный набор семантических падежей. Описаны функциональные возможности системы при создании нового перечня уникальных семантических ролей.

5. Предложена система структурной разметки и выравнивания научно-технических текстов. Показано, что при выравнивании система ориентирована

на композиционную структуру самих научно-технических текстах, а также дополнительно проверяется на основе терминологической разметки терминов в параллельном корпусе научно-технических текстов.

6. Проиллюстрированы варианты использования параллельного корпусу при решения прикладных задач в области лингвистики и информатики на примерах теории перевода, CEO-копирайтинга и генеративного искусственного интеллекта.

Основные результаты к разделу опубликованы в работах [2, 5, 11, 13, 26, 27, 29, 33, 71, 73, 74].

6. ИСПОЛЬЗОВАНИЕ РАЗРАБОТАННЫХ МОДЕЛЕЙ И МЕТОДОВ ДЛЯ РЕШЕНИЯ ПРАКТИЧЕСКИХ ЗАДАЧ ПО ОБРАБОТКЕ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ

6.1. Метод обработки учебных пособий для создания лексического тренажера по дисциплине «Иностранный язык»

Стремительное развитие информационных технологий оказало значительное влияние на лексикографию, которой требовались десятки лет на организацию сбора, обработки, хранения и представления лексического материла. Использование современных технологий позволило сократить время на внесение изменений в словарь от нескольких лет до нескольких минут, а также значительно увеличить объемы содержащейся в словарях информации [251]. Так, объём словника электронного словаря Мультитран на сегодняшний день составляет приблизительно 12 млн. слов, Prompt - 11 млн. слов и ABBYY LingvoOnline - 18 млн. слов; количество рабочих языков: Мультитран - 12, Google - 102, Prompt – 13, ABBYY LingvoOnline - 28.

Не смотря на значительные успехи электронной лексикографии, использование многочисленных электронных ресурсов часто представляет значительные трудности не только для студентов и начинающих переводчиков, но и опытных специалистов, особенно при переводе специальной научно-технической литературы. Решение проблемы может состоять в возможности использования специализированных словарей, однако их количество также стремительно возрастает: Prompt – 16 словарей, ABBYY LingvoOnline – около 250 словарей, Мультитран – около 2500 словарей. Таким образом, поиск подходящего переводного эквивалента в рамках узкой специализации даже при помощи специализированных словарей занимает значительное время, а отдельные узкоспециализированные переводные словари представляют собой ограниченный класс лексикографических ресурсов [252]. Особую сложность представляет изучение специальной лексики в группах с разным уровнем владения иностранным языком.

В связи с вышесказанным актуальным является создание лексического тренажера по изучению специальной терминологии, содержащего все лексические единицы и задания на их запоминание, которые должны быть освоены в рамках дисциплины «Иностранный язык (английский)» для студентов технических специальностей МГТУ им. Н.Э. Баумана. При этом, проблематика создания учебных словарей и лексических тренажеров довольно широко освящена в научной литературе и имеется значительный спрос на такие ресурсы [253, 254], актуальными является создание методов и программных средств автоматической и / или автоматизированной генерации предметных словарей и лексических тренажеров.

Актуальность создания лексического тренажера по освоению специальной лексики обусловлена следующими факторами:

- длительный срок изучения иностранного языка в МГТУ им. Н.Э. Баумана: бакалавриат – 6 семестров, магистратура – 2 семестра, специалитет – 7 семестров. При этом, начиная с 5 семестра бакалавриата и специалитета, а также 1 семестра магистратуры изучается английский язык в профессиональной сфере.

- большое разнообразие узких предметных областей. В 2022/2023 гг. в МГТУ им. Н.Э. Баумана насчитывается 479 образовательных программ, подавляющее число которых имеют свою уникальную узкопрофессиональную терминологию, что в свою очередь свидетельствует об использовании большого числа разных учебных пособий по иностранному языку.

- разработка пособий преподавателями университета. Учебные пособия пишут преподаватели иностранного языка в МГТУ им. Н.Э. Баумана, т.е. требуется затрачивание значительных временных и человеческих ресурсов для создания комплектов учебных и практических материалов.

- использование нового учебного пособия каждый семестр препятствует возможности повторения пройденного ранее лексического материала с целью его повторения перед итоговыми аттестациями.

- сложность использования современных словарей и терминологических

баз данных из-за их универсальности, с одной стороны, и перегруженности избыточной для студентов информацией, с другой стороны, а также отсутствие инструментов для запоминания специальной лексики.

Таким образом, цель создания лексического тренажера – оказание содействия студентам инженерных специальностей МГТУ им. Н.Э. Баумана в освоении специальной лексики при подготовке к практическим занятиям и итоговым аттестациям по английскому языку.

Основные задачи разработки лексического тренажера для студентов инженерных специальностей МГТУ им. Н.Э. Баумана включают:

– сбор языкового материала, необходимого для освоения учебной программы дисциплины «Иностранный язык»;

– форматирование и разметка полученного языкового материала в виде, пригодном для машинной обработки;

– разработка алгоритмов лексического тренажера, соответствующие лингводидактическим требованиям лексического тренажера;

– создание статического сайта / приложения для размещения языкового материала и создание лексического тренажера.

Из перечисленных задач наиболее рутинной и требующей значительных временных и человеческих затрат является сбор и разметка языкового материала. Учебные пособия по дисциплине «Иностранный язык» написаны на полностью английском языке, что приводит к необходимости разработки метода извлечения лексики из учебного пособия, подбора переводных эквивалентов на русском языке и генерации тренажера по изучению извлеченной лексики.

Метод обработки учебных пособий для создания лексического тренажера по изучению специальной терминологии представлен на рис. 6.1. Рассмотрим использование метода обработки текстов учебных пособий на основе учебного пособия «English in the Digital Age» [255] (в 4-х частях), которое используется на 1 и 2 курсах инженерных специальностей на кафедре «Английский для машиностроительных специальностей» МГТУ им. Н.Э. Баумана.

Рис. 6.1. Метод обработки учебных пособий для создания лексического тренажера по изучению специальной терминологии

Каждая часть учебного пособия состоит из трех модулей, содержащих современные аутентичные тексты, целью которых является развитие коммуникативной, лингвострановедческой и социокультурной компетенций. Издание также содержит ссылки на аудио- и видеоматериалы из открытых источников англоязычной информационно-образовательной среды и разнообразные лексико-грамматические упражнения. На изучение каждого модуля отводится 5 семинарских занятий, где 1-4 занимаются по учебному пособию, а на 5 занятии проводят рубежный контроль и защиту презентации по теме модуля.

На первом этапе проводится морфологическая разметка текста учебного

пособия, которая в последующем является основой для извлечения и разметки многокомпонентных терминов в тексте, генерации тестовых заданий и отражения грамматической информации о лексической единице в словаре тренажера.

На втором этапе проводится структурная разметка текста учебного пособия. Модель научно-учебного текста представлена в главе 2. Ее использование позволяет определить структурные элементы, которые не содержат специальной лексики (формулировки заданий, грамматические комментарии, задания на перевод на английский язык), а также структурные элементы, содержащие основную лексику модуля. Таким образом, можно будет сформировать перечень всей лексики, используемой в пособии, и основной лексики для изучения в каждом модуле учебного пособия. Основной языковой материал учебном пособии «English in the Digital Age» представляет собой лексику из разделов «Essential vocabulary», «Vocabulary related to…», «Useful notes», «Module's-Topic Related Vocabulary».

На третьем этапе осуществляется разметка многокомпонентных терминов как одной лексической единицы при создании заданий тренажера, а затем формирование списка основной лексики модуля. В результате обработки текста учебного пособия по каждому из 12 модулей выделено: 1 модуль – 67, 2 модуль – 66, 3 модуль – 51, 4 модуль – 39, 5 модуль – 141, 6 модуль – 124, 7 модуль – 83, 8 модуль – 102, 9 модуль – 38, 10 модуль – 70, 11 модуль – 70, модуль 12 – 49 единиц основной лексики, владение которой студент должен продемонстрировать на семинарских занятиях и промежуточных аттестациях.

На следующем этапе к каждой лексической единице модуля на основе параллельного корпуса научно-технических текстов подбирается 1-3 переводных эквивалента. При этом используется принцип соответствия предметной области учебного пособия и соответствующего подкорпуса, а также частотность использования переводного эквивалента в подкорпусе. Так, например, на основе словаря Мультитран к лексической единице *skin*, которая используется в шестом модуле, посвященному космонавтике, даны следующие

переводные эквиваленты: в общем словаре: *мех; кожура; кожица; оболочка; обшивка; корка; плена; наружный слой; шкурка; кляча; жулик; скряга; кожа; шкура; снять кожуру; покрыть кожей; саба; снять кожу; снять шкуру; шелуха.* В словаре «Техника» ресурса Мультитран *окалина; покрытие; поверхностный слой; кожа; обшивочный лист; очищать от изоляции,* а в словаре «Космонавтика» вообще отсутствует. Уточнение запроса путем добавления слова *spacecraft skin* приводит к тому, что нужно смотреть уже две словарные статьи. Безусловно подбор переводного эквивалента может и очевиден для специалиста в данной предметной области и не представляет сложностей - *обшивка*, но студентам как минимум требуется время на его подбор или они вообще выбирают неправильный перевод. На основе данных их параллельного корпуса научно-технических текстов, тематическое деление которого основано на направления подготовки МГТУ им. Н.Э. Баумана, у лексической единицы *skin* наиболее частотными являются переводные эквиваленты по космонавтике *обшивка, обшивочный лист,* а биомедицинским технологиям – *кожа, кожный покров.*

Затем на основе подкорпуса осуществляется подбор примеров употребления лексической единицы в контексте, а также его перевод. Приоритетными контекстами являются те, которые содержат не только основную лексику модуля, но и специальную терминологию изучаемой предметной области, так как эти контексты используются для автоматической генерации заданий тренажера. В противном случае могут быть сформированы контексты, куда по смыслу может подходить практически любое слово, например:

The phenomenon of <<b>…<d>> has always been very exciting, both for its fundamental scientific interest and because of its many applications.

Феномен <<b>…<d>> всегда был очень захватывающим как из-за его фундаментального научного интереса, так и из-за его многочисленных применений.

В лексическом тренажере представлено несколько видов заданий для

овладения специальной лексикой: карточки, выбор верного переводного эквивалента и заполнение пропусков. Генерация заданий реализуется на основе следующих правил: варианты ответов подбираются из одного модуля, обладают одинаковыми морфологическими, грамматическими и синтаксическими характеристиками. Например, при формировании задания на выбор правильного переводного эквивалента все варианты ответов будут относиться к одной части речи, числа, использоваться в рамках одной предметной области, иметь одинаковую формальную структуру, т.е. если правильный ответ является словосочетанием, то и другие варианты ответов также будут словосочетаниями и т.п.

Каждая лексическая единица сопровождается транскрипцией и аудиофайлом для прослушивания данной единицы в британском варианте произношения. К полученным словарным статьям обязательно добавлялась грамматическая категория части речи, что позволило составлять соответствующие лингводидактическим нормам задания для тренажера автоматически и без использования способных проводить грамматический парсинг лексемы алгоритмов искусственного интеллекта для обработки естественного языка. Для тех же целей в примерах употребления лексема обязательно выделялась тегами: открывающим «<<b>» и закрывающим «<d>>». В каждой словарной статье указывался модуль учебного пособия, к которому относится слово.

Словарные статьи обрабатывались и загружались на сервер в виде объекта JavaScript, содержащего объекты JSON вида:

```
{"word": string,
  "translation": string,
  "transcription": string,
  "pos": string,
  "examples": [{
        "eng": string,
        "ru": string
```

}],

"module": number }

где поле "word" хранит лексему на английском языке, поле "translation" хранит перевод или переводы, разделенные запятой, поле " transcription" хранит транскрипцию лексемы, поле "pos" хранит часть речи лексемы, поле "examples" хранит массив с одним элементом – объектом JSON, хранящим поля «eng» и «ru», поле "eng" хранит пример употребления лексемы на английском языке, поле "ru" хранит пример употребления лексемы на русском языке, поле "module" хранит номер модуля, к которому относится лексема..

Словарные статьи лексического тренажера содержат грамматический материал о лексической единице и аудиозаписи британского варианта произношения. Пример словарный статьи показан на рис. 6.2.

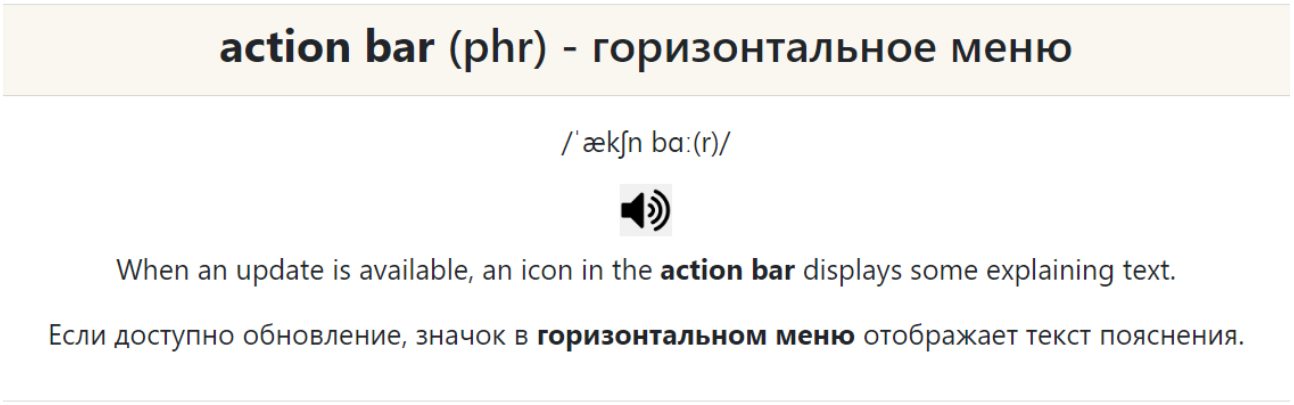


Рис. 6.2. Словарная статья к терминологической единице «action bar»

При заполнении пропусков используются контексты, внесенные в словарь лексического тренажера, а основная лексика имеет соответствующую разметку и при формировании заданий является правильным ответом, например:

His invention was a long way away where we are today with our mobile phones, <<b>tablets<d>> and laptops but it was a significant development in the use of calculating devices.

Его изобретение было далеко от современных телефонов, <<b>планшетов<d>> и ноутбуков, но оно было значительным прорывом в области счетных устройств.

Другими вариантами являются *abacus*, *calculator*, *display* и *абак, калькулятор, монитор* соответственно на русском языке.

Для разработки лексического тренажера выбраны следующие технологии:

– HTML — стандартизированный язык разметки документов для просмотра web-страниц в браузере. Использовался как основной инструмент верстки web-страниц сайта ЭУС.

– CSS — формальный язык описания внешнего вида документа (web-страницы). Использовался для описания дополнительных свойств элементов web-страниц.

– Bootstrap 4 — свободный набор HTML- и CSS-шаблонов для создания сайтов и web-приложений. Классы Bootstrap были использованы для стилизации большинства элементов сайта, что позволило сократить затраченное на верстку время к минимуму при сохранении качественного внешнего вида web-страниц ресурса.

– JavaScript — мультипарадигменный язык программирования, поддерживающий слабую динамическую типизацию и автоматическое управление памятью. Был использован в качестве основной технологии для разработки поведения сайта и хранения словарных статей.

– Git — распределённая система управления версиями. Использовалась для удаленной разработки.

– JSON — основанный на JavaScript текстовый формат обмена данными. Использовался для хранения словарных статей в удобном не занимающем большого количества оперативной памяти формате.

– Go — компилируемый многопоточный язык программирования. Использовался для разработки программы-парсера для получения аудиозаписей произношений слов с сайта «Oxford learner’s dictionary» и Google Translate API.

– Python — высокоуровневый язык программирования общего назначения

с динамической строгой типизацией и автоматическим управлением памятью. Использовался как основной инструмент для обработки словарных статей и тегирования предложений-примеров.

Основными принципами проектирования словаря являются простота и доступность, так как минималистичный и интуитивно понятный графический интерфейс ресурса фокусирует внимание студента на изучении языкового материала.

Все страницы сайта лексического тренажера содержат навигационную панель с поисковой строкой и ссылками на главную страницу сайта и лексический тренажер. На главной странице сайта расположены ссылки для поиска языкового материала по первой букве и по модулю учебного пособия, как показано на рис. 6.3.

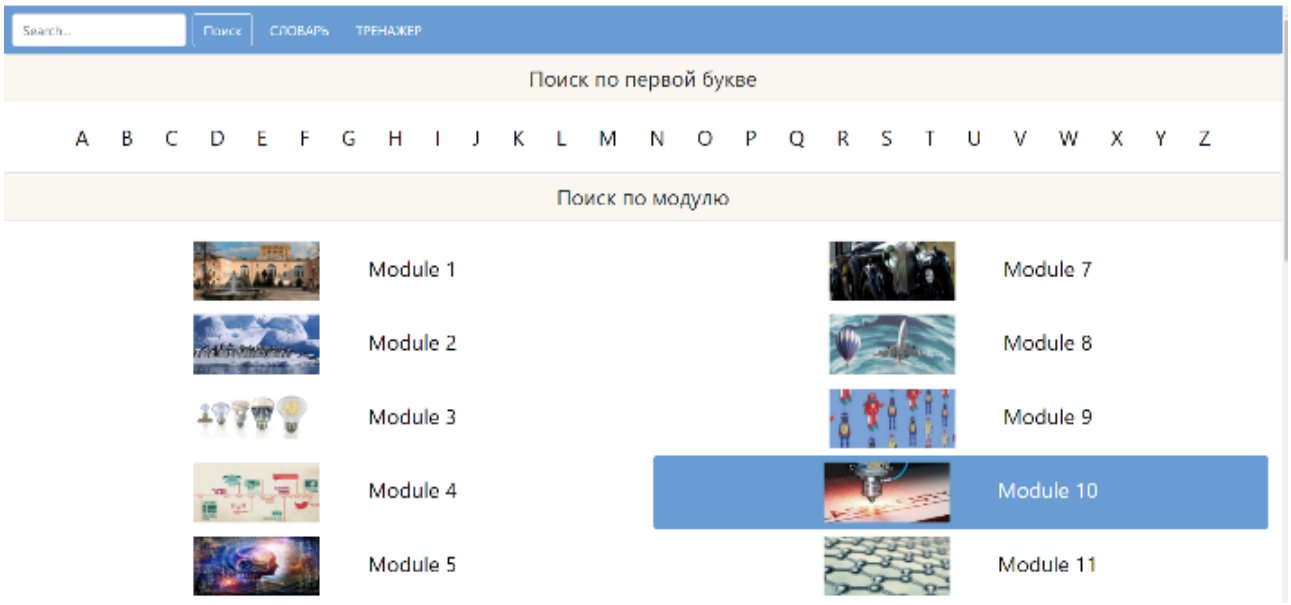


Рис. 6.3. Интерфейс лексического тренажера

Для поисковой строки реализован «живой поиск» методом поиска соответствий по средствам строгого регулярного выражения. Ссылки поиска по первой букве и по модулю ведут на соответствующие страницы с прямыми ссылками на искомые словарные статьи. Ссылки из списка всего языкового материала и из поисковой строки являются прямыми ссылками на искомые

словарные статьи.

Алгоритмы данного тренажера позволяют автоматически составлять более 100 000 типовых задач на перевод, выбор верного ответа, заполнение пропусков в предложении (Рис. 6.4). В лексическим тренажере присутствует классическая система карточек для заучивания лексики. Имеется система фильтрации задач по модулям в зависимости от потребностей учащегося.

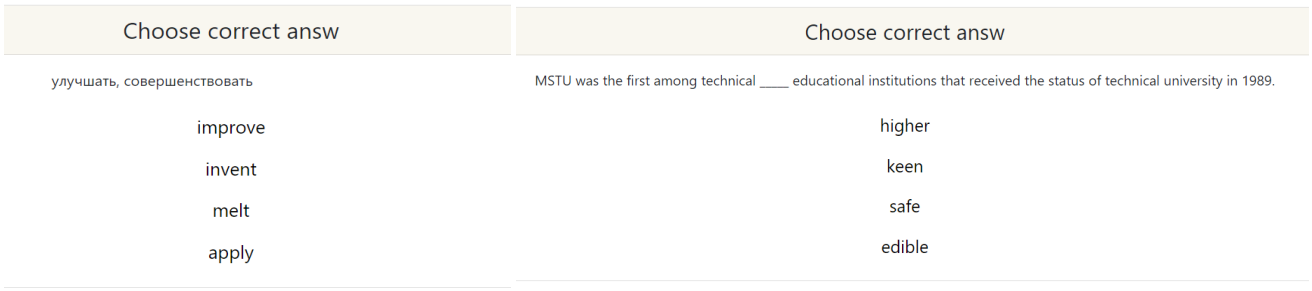


Рис. 6.4. Примеры сгенерированных заданий в лексическом тренажере

В таблице 6.1 представлены результаты сравнения возможностей лексикографических ресурсов для поиска переводных эквивалентов и изучения специальной лексики студентами технических специальностей МГТУ им. Н.Э. Баумана. Основным критерием при отборе лексикографических ресурсов была доступность и частота использования в студенческой среде.

Таким образом, не смотря на наличие больших лингвистических баз данных и словарей, которые содержат в себе значительные объемы информации о каждом ресурсе, для изучения специальной лексики в рамках дисциплины «Иностранный язык» целесообразно использовать ресурсы, привязанные к конкретным учебным пособиям. Предложенный метод обработки научно-учебных текстов может быть также использован для проверки разрабатываемых учебных пособий, контрольных работ и других материалов на наличие основной лексики и количестве ее употреблений в тексте пособия, или же использования незнакомой лексики при создании проверочных работ [256].

Таблица 6.1. Сравнительный анализ возможностей лексикографических ресурсов по изучению специальной лексики

	Ресурс / Критерий	Мультитран	Abby Lingvo	Promt	Словарь Google	Словарь Yandex	**Лексический тренажер**
1	Количество переводных эквивалентов в словаре	1-50	1-15	1-10	1-20	1-20	**1-3**
2	Соотнесенность с конкретной предметной областью	+	+	+	+	+	+
3	Привязка к конкретному учебному материалу / пособию	-	-	-	-	-	+
4	Использование карточек для изучения лексики	-	+	-	-	-	+
5	Тестовые задания для изучения лексики	-	-	-	-	-	+
6	Наличие транскрипции и грамматической информации	+	+	+	+	+	+
7	Примеры употребления изучаемой единицы в контексте	-	+	-	+	-	+

6.2. Методы выявления тенденций развития научных направлений (на материале анализа публикаций по газовому топливу)

Количество научных публикаций в мире ежегодно увеличивается, причем основная доля их роста приходится именно на развивающиеся страны. Всего в 2018 г. количество новых научных статей выросло на 5% – до 1,6 млн, включенных в обширную базу данных Web of Science. Этот показатель стал самым крупным за всю историю мировой науки, т. е. обновление информации о результатах научных исследований происходит так часто и интенсивно, что в плотной информационной среде все труднее отделить псевдонаучную информацию от истинно научной [257]. Для работы с большим количеством публикаций создаются инструменты их автоматической обработки, например, библиометрический анализ, который позволяет выявлять тенденции развития научных направлений.

Цель исследования – это изучение способов анализа научных тенденций на основе публикаций в сфере компримированного природного газа, включая его

применение в качестве моторного топлива. Для достижения поставленной цели необходимо: найти источники и инструменты для формирования базы научных работ по теме; провести библиометрический анализ с целью выявления закономерностей; проверить выявленные тенденции с помощью поиска подтверждающей информации в Интернете; создать массив статей для дальнейшей обработки с помощью специального программного обеспечения; выявить в результате работы с массивом, чему исследователи уделяли особое внимание в разные годы в области компримированного природного газа.

Сбор данных проводился с помощью электронных ресурсов. Для получения библиометрических показателей осуществлялось подключение к базам данных РИНЦ, Scopus и Web of Science, а для обработки и визуализации полученных результатов использовалась программа VOSviewer.

Статьи, входящие в РИНЦ, представлены на сайте eLibrary.ru. Он разработан и поддерживается компанией «Научная электронная библиотека». Создав свою учетную запись, можно воспользоваться функциями, которые предлагает сервис, например, работа с подборками публикаций. База данных РИНЦ выполняет функцию не только инструмента для оценки деятельности учёных или научных организаций на основе индекса цитирования, но и авторитетного источника библиографической информации по российской научной периодике. В России эта база данных является одним из основных источников информации для оценки эффективности организаций, занимающихся научными исследованиями. Статьи, текст которых был необходим нам для соответствующего анализа, также можно скачивать с сайта научной электронной библиотеки «КиберЛенинка» [258].

Web of Science – поисковая интернет-платформа с реферативными базами данных. По окончанию сессии нами были получены библиометрические данные в формате, предназначенном для чтения программой VOSviewer, предназначенной для построения и визуализации библиометрических сетей. В данном случае она использовалась для построения карты терминов научной литературы. В отличие от eLibrary, на платформе Web of Science проводился

анализ зарубежных работ. На своем веб-сайте Web of Science определяет себя как «информационно-аналитическая платформа с информацией о ведущих международных публикациях» [259].

Подобно платформе Web of Science, Scopus также библиографическая и реферативная база данных. Её разработчиком и владельцем является издательская корпорация Elsevier. На платформе Scopus можно найти статистические данные по полученным результатам поиска [260].

Научные ресурсы, опубликованные после 1996 г., индексируются в БД Scopus вместе со списками пристатейных библиографических списков. Цитируемость в этой базе данных подсчитывается путём автоматизированного анализа содержания этих списков. Таким образом, по БД Scopus можно определить количество ссылок на все проиндексированные ресурсы, но только опубликованные с 1996 г.

В отличие от БД Web of Knowledge, в БД Scopus не используется понятие импакт-фактора (численный показатель цитируемости статей, опубликованных в определенном научном журнале), зато широко применяется индекс Хирша (количественная характеристика продуктивности учёного, группы учёных, научной организации или страны в целом, основанная на количестве публикаций и количестве цитирований этих публикаций) [261].

Программа VOSviewer – инструмент для построения и визуализации библиометрических карт. В нашем случае объектами карты являются термины. В программе образуются связи – каждая связь соединяет два термина и определяется за счет совместного появления двух терминов. Если два термина встречаются в одном текстовом корпусе рядом с друг другом в определенном порядке больше, то их связь сильнее. В данном контексте сочетание может означать семантическую близость. Корпусная лингвистика с помощью статистического анализа выявляет закономерности совместной встречаемости, на основе которых можно делать различные выводы, например, обнаруживать идиоматические выражения.

В рамках исследования подходов ученых к теме компримированного природного газа предполагается обнаружить закономерности совместной встречаемости технологий, использующихся в этой сфере, и применений сжатого природного газа. В процессе визуализации связи складываются в сети, а ключевые слова образуют кластеры. Карта строится таким образом, что термин может относиться только к одному кластеру. Кластеры, в свою очередь, образуются из набора близких терминов с выделяющимся количеством связей с друг другом. Частота встречаемости термина определяет его «вес», значимость слова передает его размер (чем больше слово и его круг, тем чаще оно встречается). Дополнительно можно измерить отображение размером термина, чтобы оно соответствовало количеству и силе связей термина, а не его частоте [262].

Для обработки текстов научных публикаций использованы методы извлечения специальных терминов английского и русского языков [2, 4] и лингво-статистические методы для получения квантитативных характеристик исследуемых текстов [23, 263], а также методы морфологического анализа текстов [264, 265].

Помимо прочего, проведен анализ источников и практики финансирования научно-исследовательских и опытно-конструкторских работ (далее – НИОКР), влияющих применение природного газа в качестве моторного топлива. В ходе работы применялась единая государственная информационная система учета научно-исследовательских, опытно-конструкторских и технологических работ гражданского назначения (ЕГИСУ НИОКТР). Материал исследования дополняет картину, так как не все научные исследования получают выход в виде публикаций или каким-либо другим образом. Их деятельность можно наблюдать по системе учета, а также по библиометрическим данным на основе ключевых слов, используемых в информационных картах. В дополнение к этому есть возможность получать информацию об исполнителях и заказчиках, а также представление об объеме НИР, исходя из выделенного финансирования.

По запросу «компримированный природный газ» (КПГ) составлена подборка из 4232 публикаций, в которую входили статьи не только в журналах, но и в сборниках конференций, а также патентные документы, диссертации, книги, отчеты и др. Однако большую долю в общем объеме этой подборки все ещё составляют статьи из журналов и из сборников научных трудов, а также материалы конференций и патентные документы.

Анализ показал, что самое большое количество статей было опубликовано в 2015 и 2017 гг. После 2017 г. наблюдается спад публикационной активности и эти годы можно охарактеризовать как её скачки. Похожая ситуация была обнаружена на графиках и в других источниках. 2018 г. был примерно таким же плодотворным. Большая часть публикаций посвящена энергетике и транспорту. Причем, если рассматривать энергетику и машиностроение по отдельности, то станет видно, что увеличение публикаций в 2018 г. обеспечили статьи по тематике «энергетика».

Особое внимание стоит уделить тематике «экология». Для начала, она была выделена вместе с тематиками «энергетика» и «машиностроение», и общий объем публикаций составил 1262 работы, а затем – отдельно по тематике «экология» количество публикаций составило 90. Для этой тематики большее количество публикаций было в 2017 г. – 11 публикаций, также выделяется и 2012 г. – 9 публикаций. Возможно, на количество работ по тематике «экология» в рамках КПГ в 2012 г. повлияла ухудшающаяся экологическая ситуация – незаконная вырубка леса, реализация опасных для природы проектов без обязательной экологической экспертизы. В дополнение к этому, на фоне аварии на АЭС «Фукусима» в Японии в предшествующем году, в начале 2012 г. у производителей природного газа были большие амбиции, в связи с изменением отношения к АЭС в сфере энергетики.

Корпорация «Газпром» открыла проект «Северный поток», имея высокие экологические стандарты, хотя много критики у экологов вызвал проект платформы «Приразломная» и трагедия при транспортировке платформы после бурения на Западно-Камчатском шельфе. Таким образом, детально рассмотрев

90 публикаций по тематике «экология», можно заметить, что самой распространенной как раз является тематика «охрана окружающей среды. Экология человека». На поисковый запрос с фразой *compressed natural gas* удалось найти 2023 статьи, опубликованных и готовящихся к публикации. Статистическая информация об этих работах была получения с помощью самого сайта.

При анализе российских статей из базы РИНЦ тоже был отмечен рост количества публикаций в 2017 г., но в дополнение к этому заметны ещё два пика – в 2014 и 2019 гг. В 2019 г. был поставлен рекорд по объему выбросов CO_2. В отчете Global Carbon Project, который был опубликован в нескольких журналах, отмечено, что трансформации, необходимой для резкого сокращения выбросов парниковых газов, не предвидится (Global Carbon Project – организация, количественно оценивающая глобальные выбросы парниковых газов и их причины) [266]. Возможно, это вызвало реакцию в исследовательских кругах, так как доля публикаций в 2019 г. по тематике «окружающая среда» заметно больше, чем в 2017 г., а публикаций по тематике «энергетика» меньше.

Ввиду ограничений сайта данные скачивались по частям, по 500 статей за раз. Затем все файлы были вместе загружены в программу *VOSviewer* со следующими параметрами: учитываются все ключевые слова, а не только указанные автором, каждое ключевое слово должно встретиться как минимум 5 раз. В получившийся карте оказалось 396 ключевых слов.

На общем графике кластеров (рис. 6.5) особенно выделяются три кластера, самыми частотными словами которых являются термины *optimization*, *performance* и *emissions*. Карта имеет 5 кластеров, они названы в честь своих самых ярких представителей: 1 – *optimization* (рис. 6.6), 2 – *emissions* (рис. 6.7), 3 – *performance* (рис. 6.8), 4 – *vehicles* (рис. 6.9), 5 – *compressed natural gas* (рис. 6.10.)

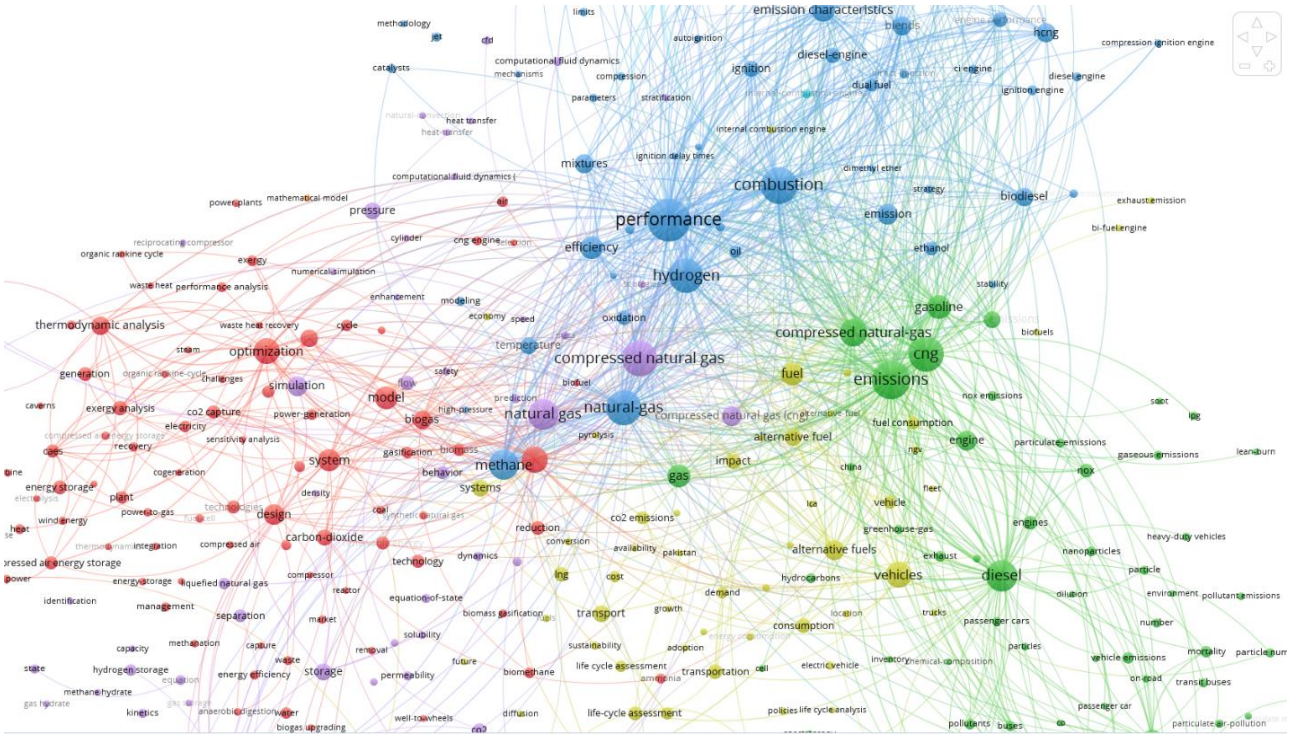


Рис. 6.5. Общая карта кластеров данных БД Web of Science

Рис. 6.6. Кластер *optimization*

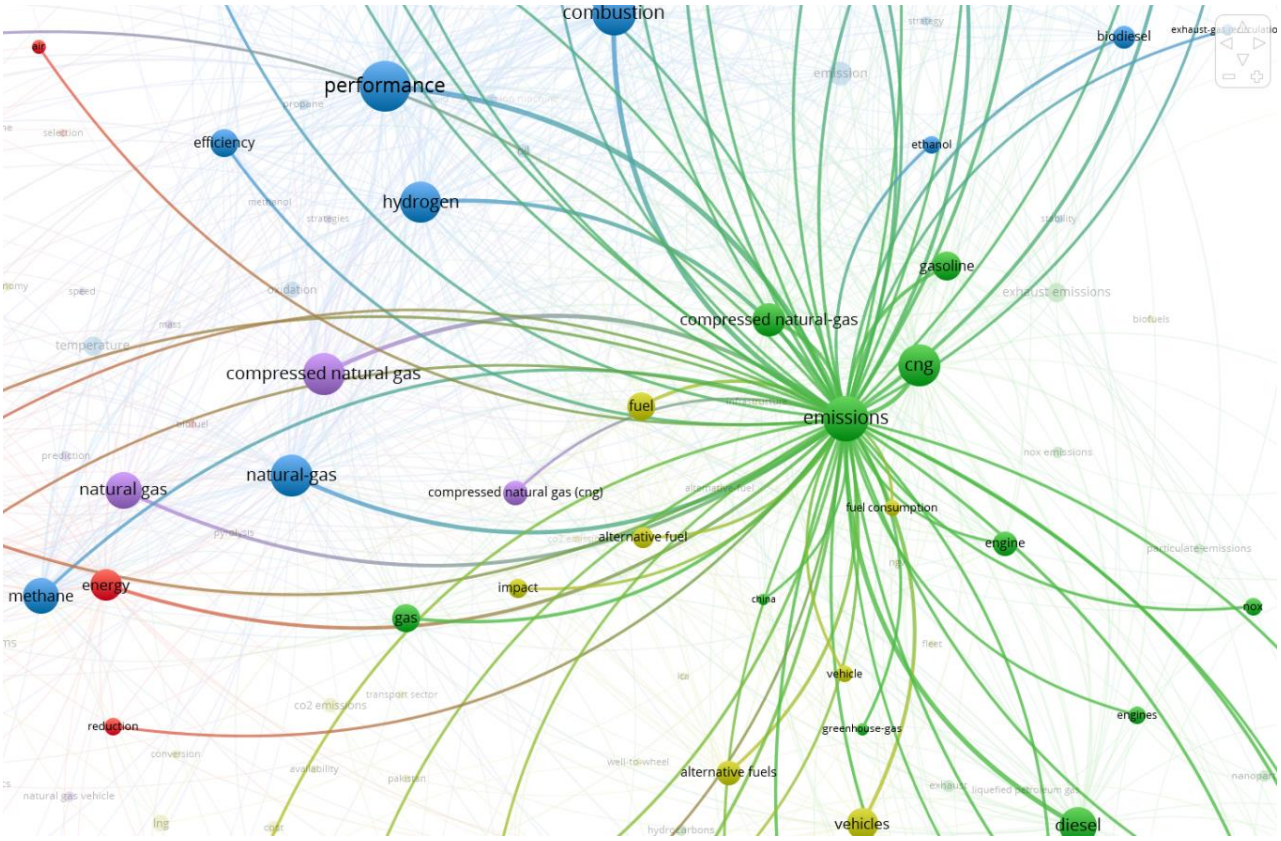


Рис. 6.7. Кластер *emissions*

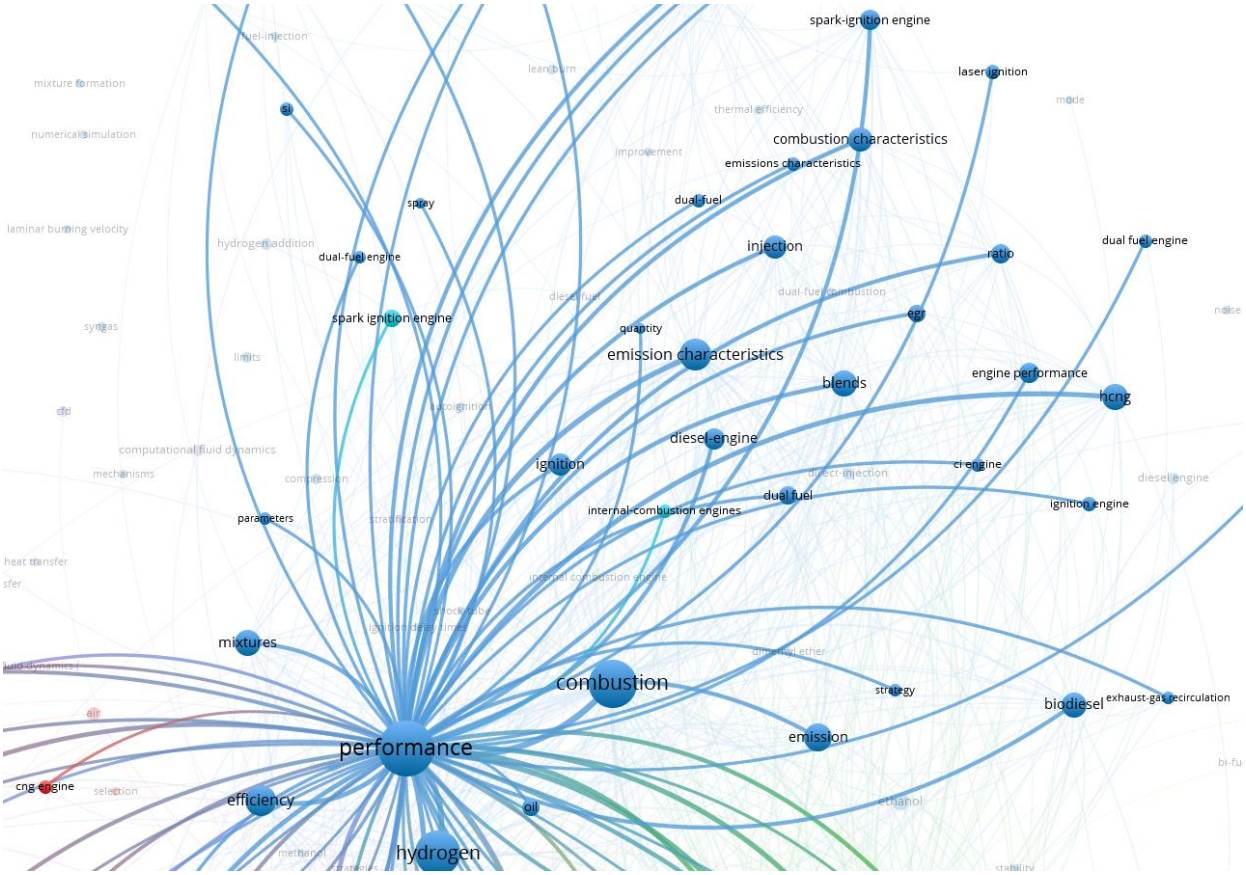


Рис. 6.8. Кластер *performance*

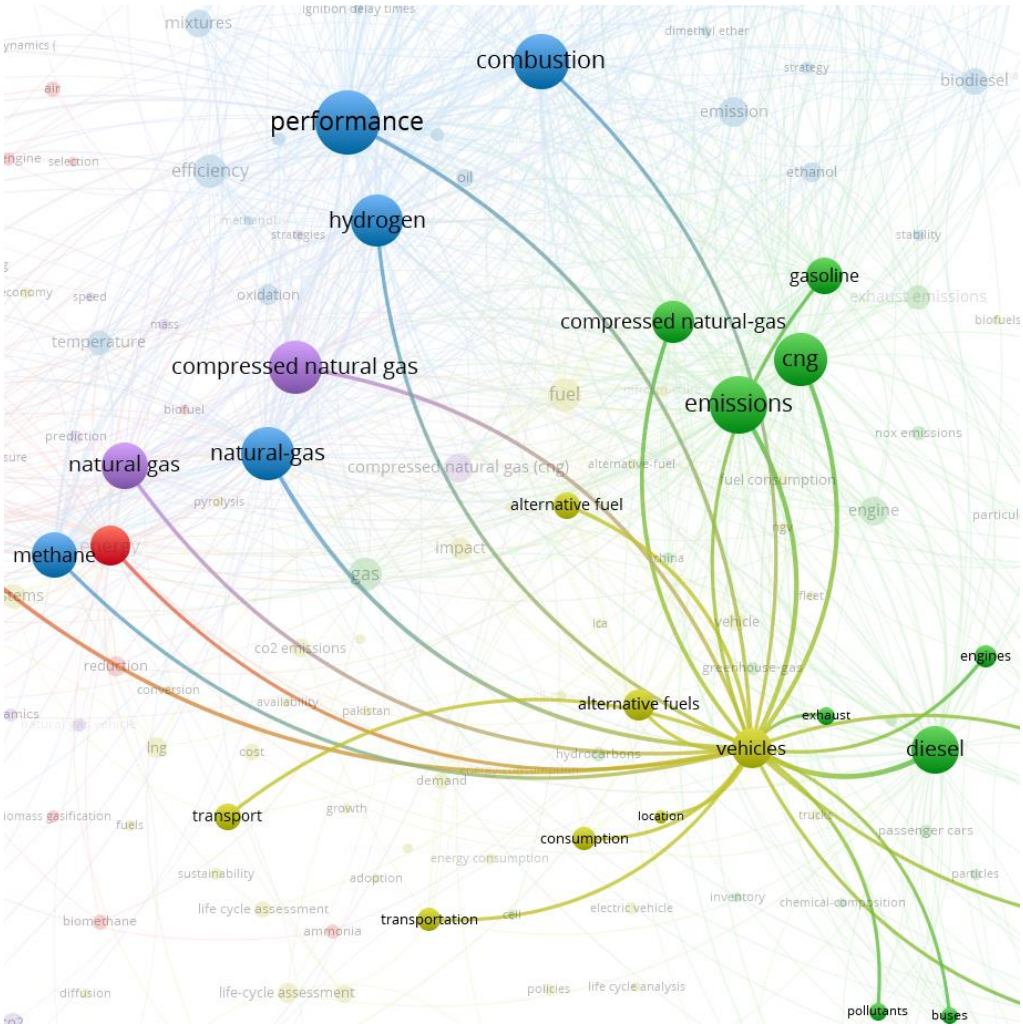


Рис. 6.9. Кластер *vehicles*

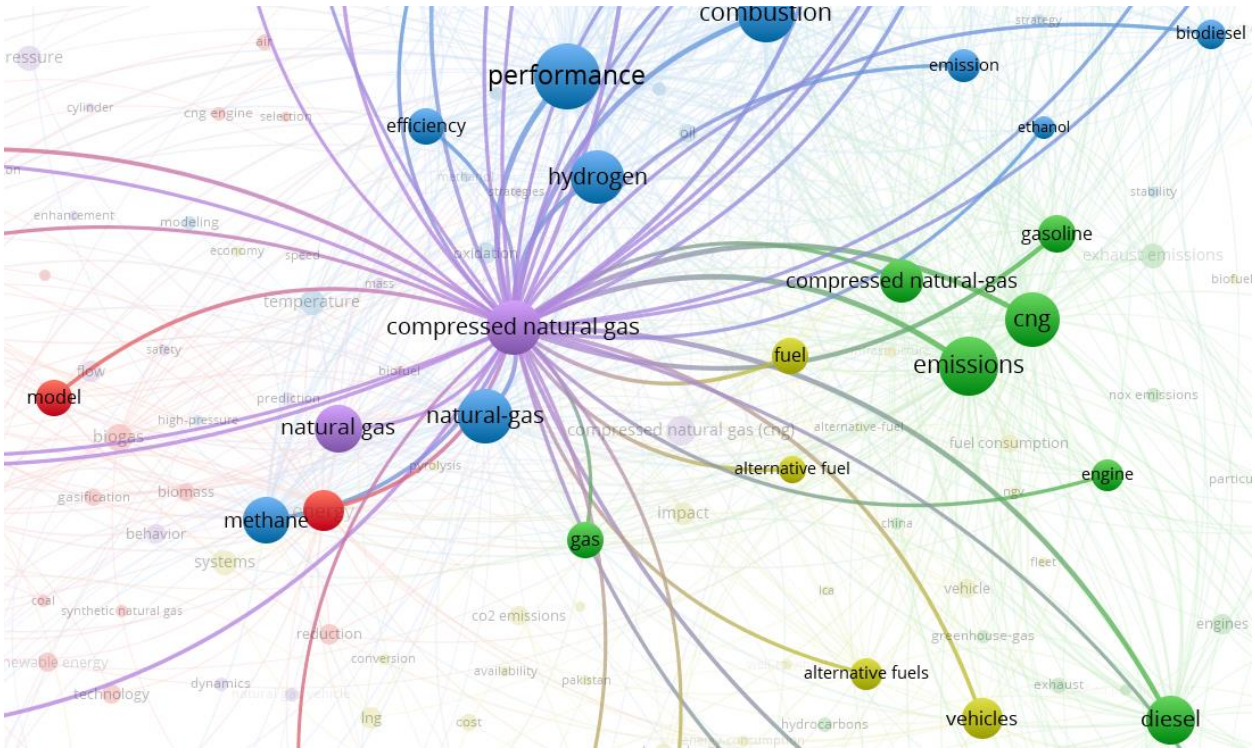


Рис. 6.10. Кластер *compressed natural gas*

Проанализировав получившиеся кластеры, можно увидеть, что основными аспектами в научных работах являются: оптимизация (систем, производства, циклов, моделей), выбросы (от двигателя, бензина, часто упоминаются вместе с газами), эффективность (горения, топлива, смесей, инъекций, зажигания), транспортное средство (транспорт, транспортировка, потребление, альтернативное топливо; близко находится к кластеру *emissions*, отсюда часто используются слова «дизельный», «выхлопы» и «двигатели») и сам КПГ. Являясь основной и «центральной» темой работ, «компримированный природный газ» имеет сильные связи с основными терминами всех кластеров, что можно наблюдать на получившейся карте (см. рис. 6.10). Это может свидетельствовать о том, что выборка для исследования статей полностью соответствует теме, в противном случае существовали бы отдельно стоящие кластеры, имеющие слабую связь с остальными.

По поисковому запросу *compressed natural gas* (поиск в названиях, аннотациях и ключевых словах) нами была обнаружена 3831 англоязычная работа. Скачанные файлы с библиометрическими данными были загружены в VOSviewer с аналогичными параметрами. В случае с выборкой из БД Scopus было найдено 19968 ключевых слов, но по условию (встречается минимум 5 раз) подошло 1939 слов. Получившаяся карта имеет 6 кластеров (рис. 6.11): 1 – compressed natural gas/diesel engines/engines (рис. 6.12); 2 – natural gas/gas industry/gases (рис. 6.13); 3 – gas emissions (рис. 6.14);, 4 – compressed air/energy efficiency/energy storage (рис. 6.15); 5 – gasoline/exhaust emission/air pollution (рис. 6.16); 6 – liquefied petroleum gas (рис. 6.17).

Рассмотрим, какие аспекты темы «компримированный природный газ» выделяются в работах, входящих в БД *Scopus*. Начнем с центрального кластера *compressed natural gas* (см. рис. 6.10), так как он представляет собой изначальный предмет исследования, термин КПГ имеет самое большое количество вхождений и больше всех связей. Программа *VOSviewer* отнесла к этому кластеру такие понятия, как: двигатель, сгорание, зажигание, топливо, смеси и др. У него есть сильные связи с ядерными словами других кластеров.

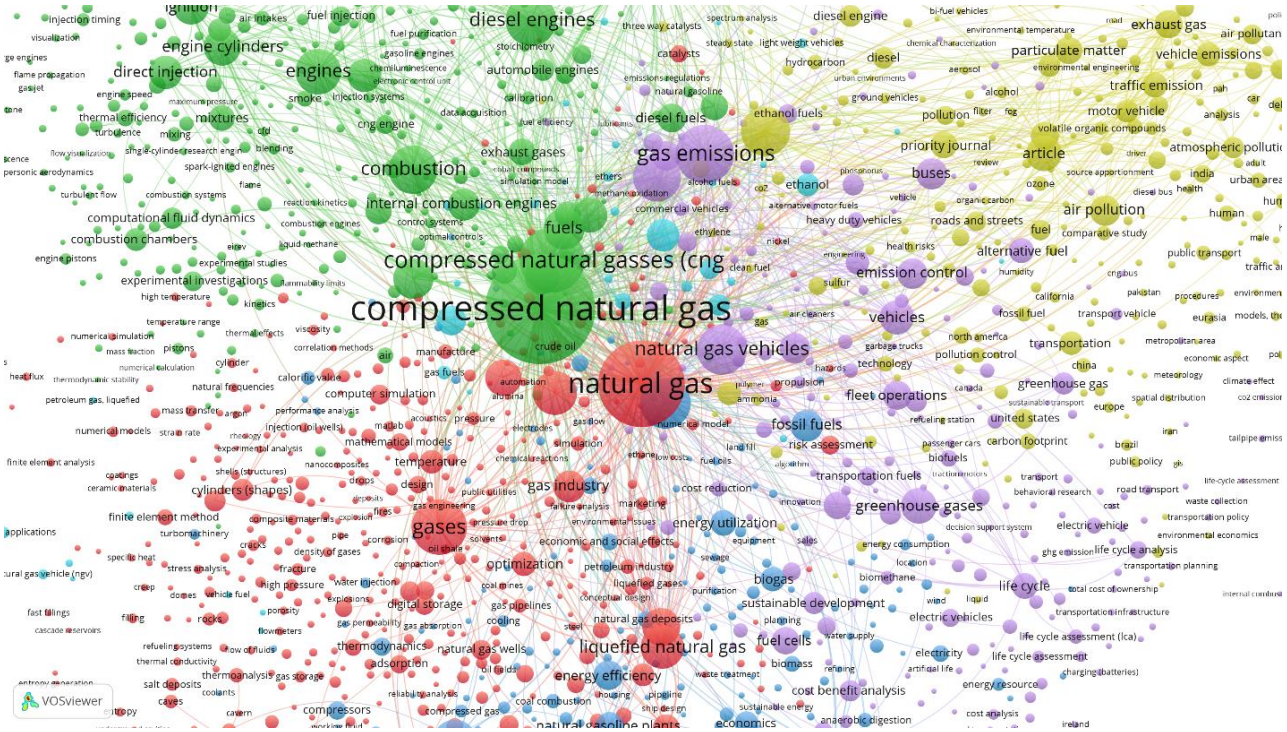


Рис. 6.11. Общий график кластеров данных *Scopus*

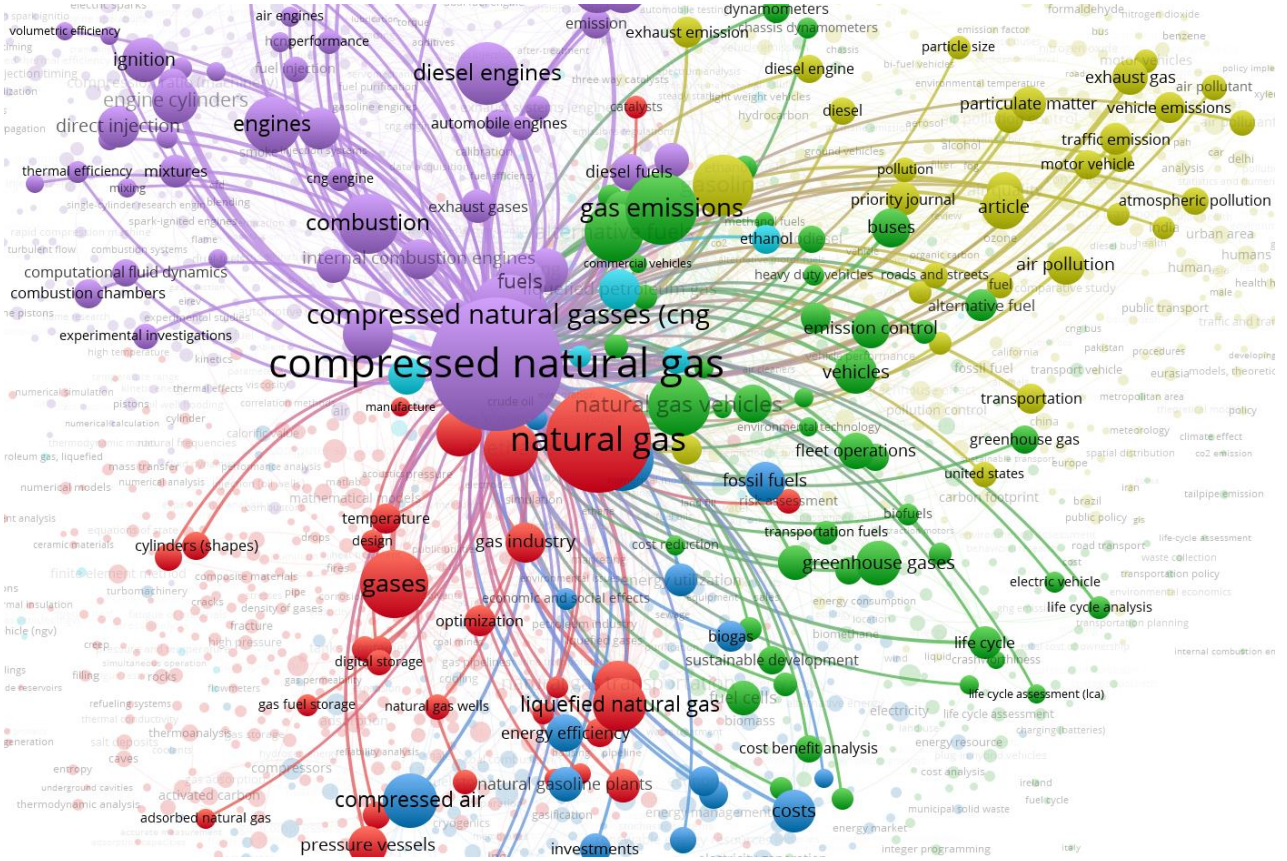


Рис. 6.12. Кластер *compressed natural gas*

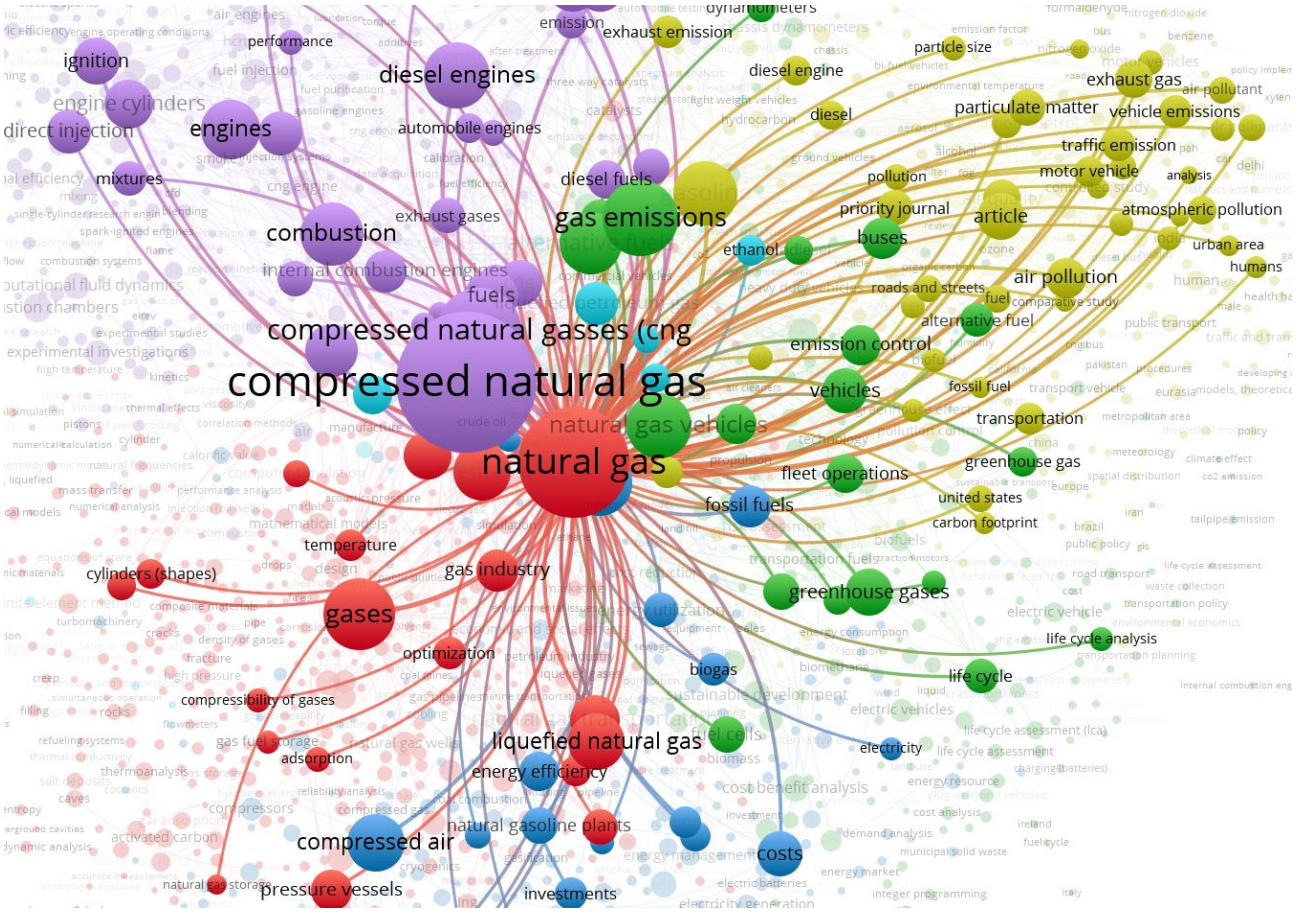


Рис. 6.13. Кластер *natural gas*

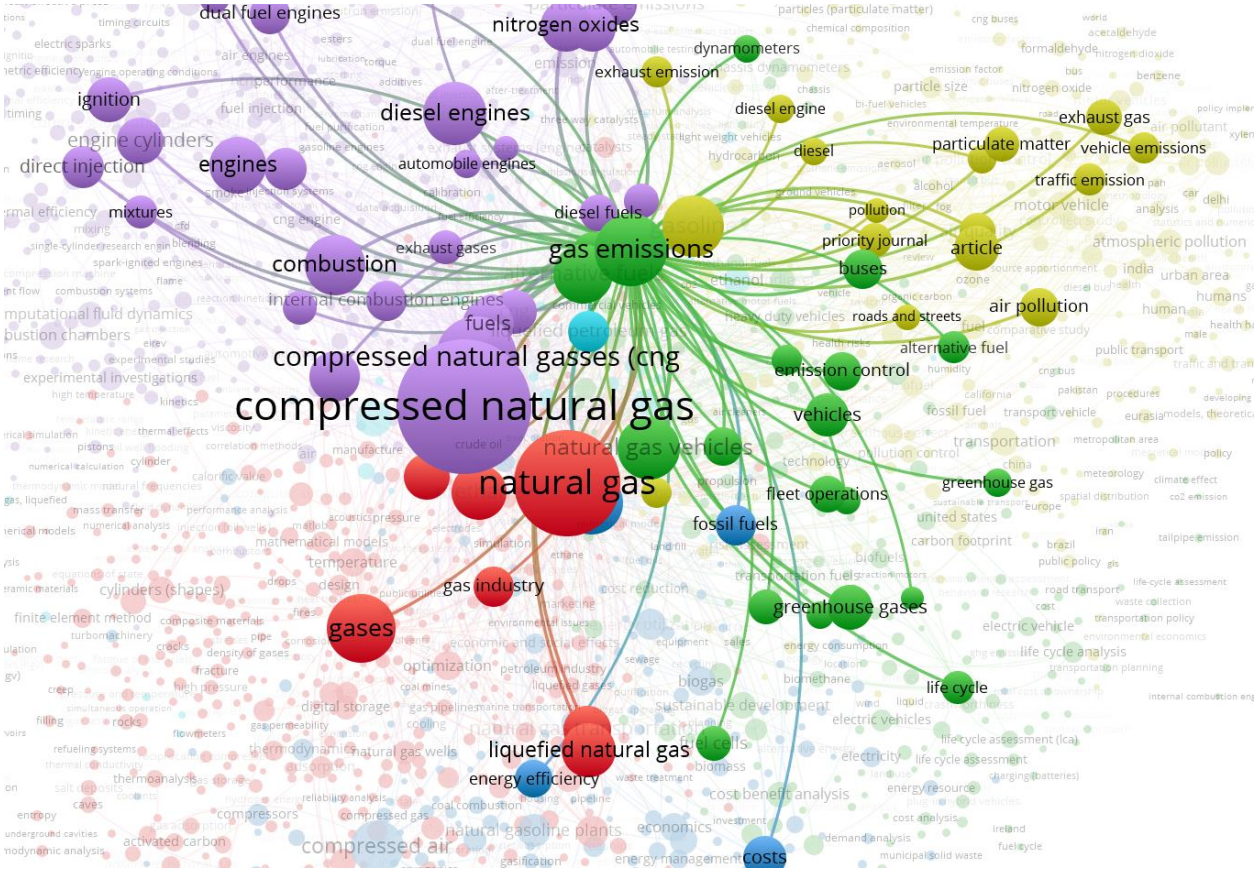


Рис. 6.14. Кластер *gas emissions*

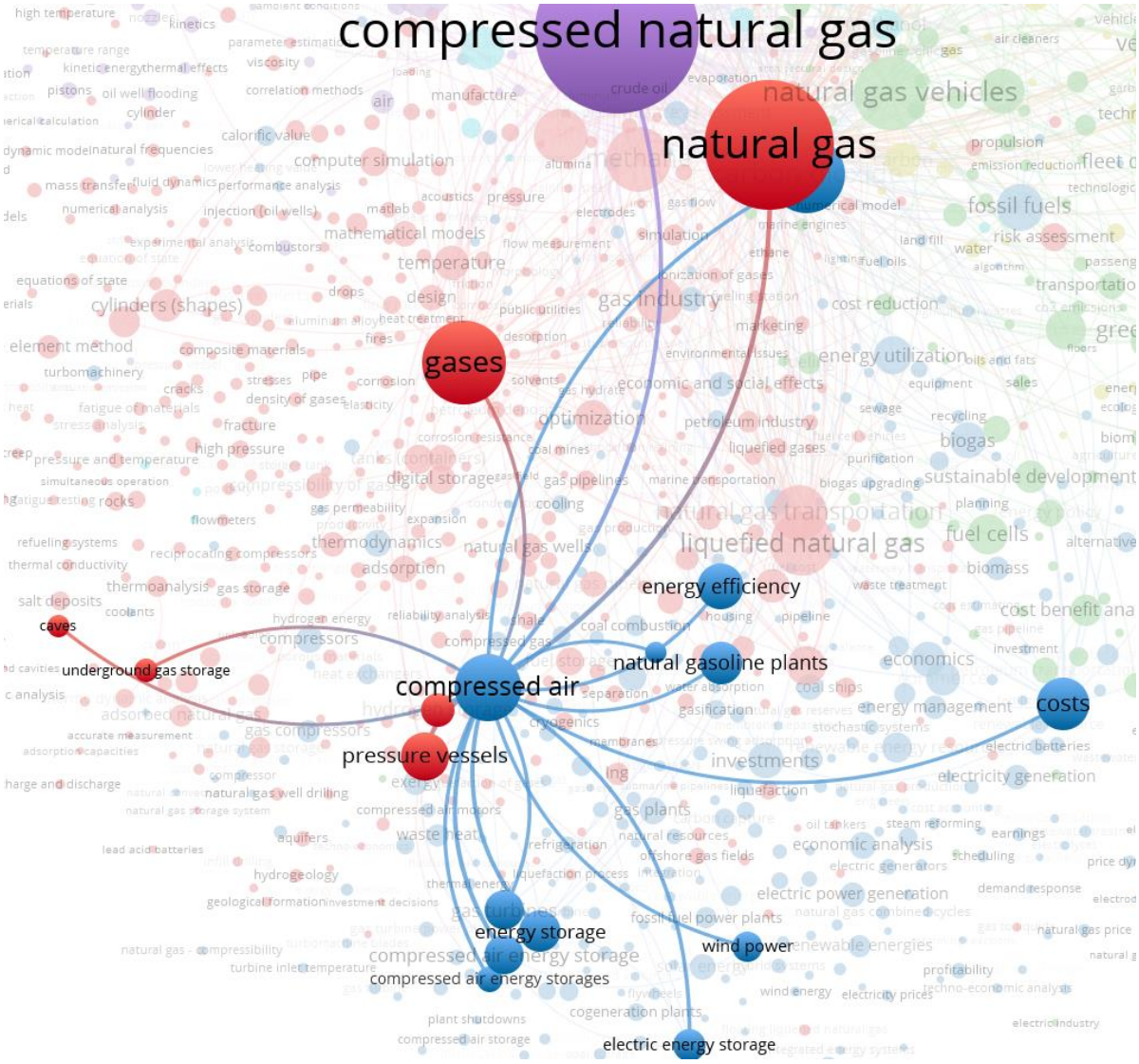


Рис. 6.15. Кластер *compressed air*

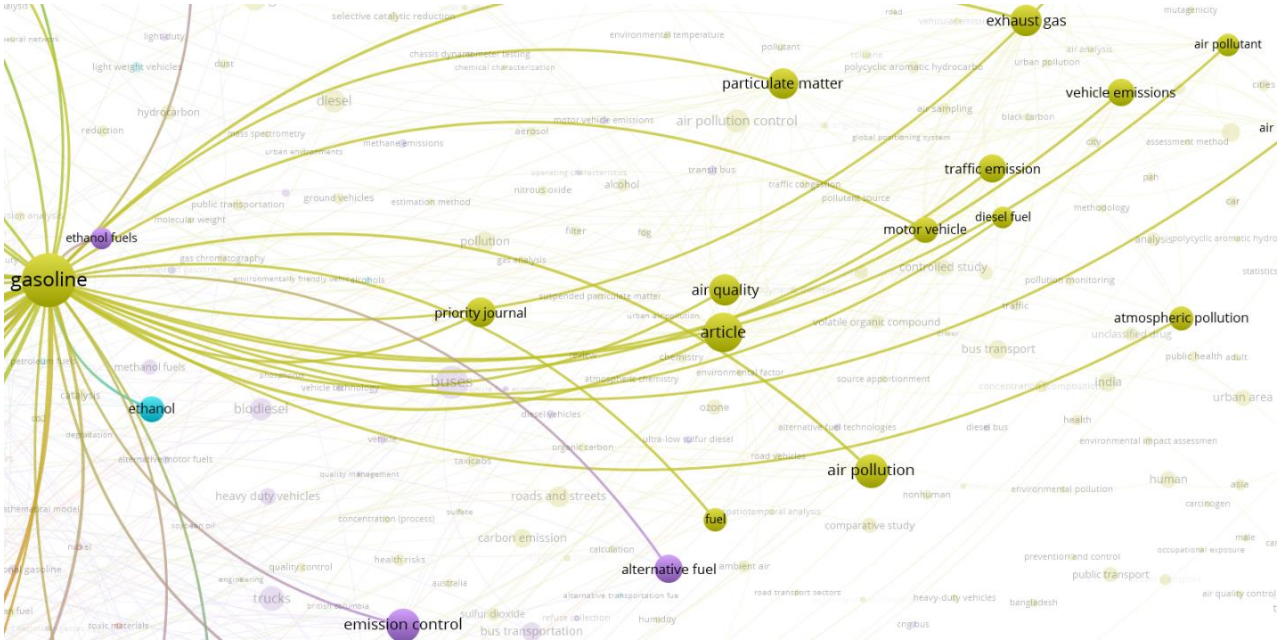


Рис. 6.16. Кластер *gasoline*

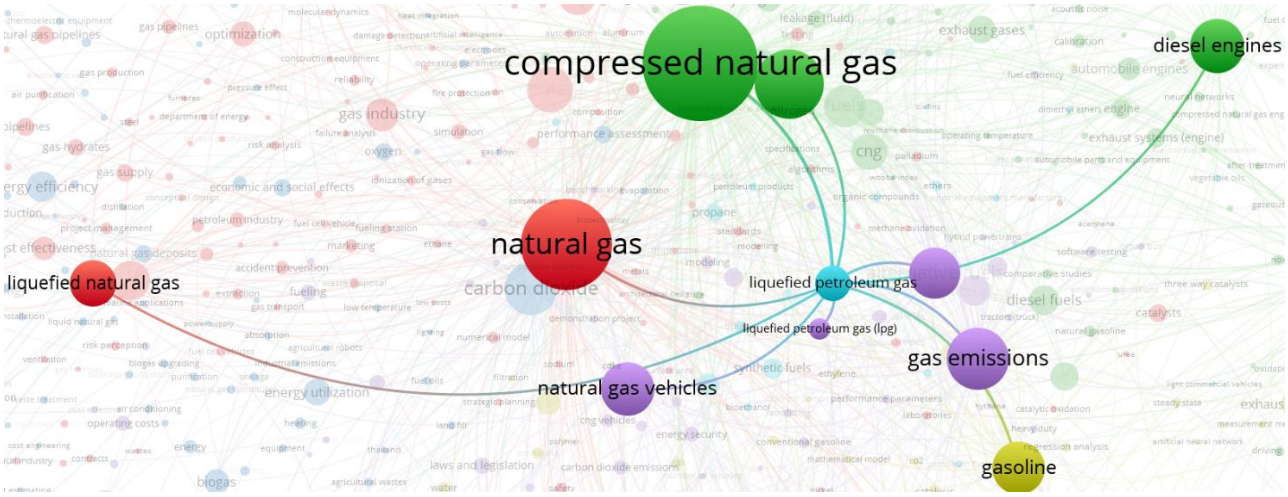


Рис. 6.17. Кластер *liquefied petroleum gas*

Кластер *natural gas* (см. рис. 6.13) имеет большую площадь пересечения с предыдущим кластером *compressed natural gas* (см. рис. 6.12), но он смещен в сторону остальных кластеров и у него нет узкой направленности. Непосредственно к этому кластеру относятся термины: газовая промышленность, оптимизация, газы, температура, сжимаемость газов и др. На этой карте (см. рис. 6.14), как и на карте БД Web of Science, есть кластер, имеющий отношение к выбросам. Примечательно, что он разделяет кластеры, посвященные природному газу, и кластер, посвященный бензину, который будет рассмотрен далее. В кластер *gas emissions* (см. рис. 6.14) входят термины: контроль выбросов, парниковые газы, транспортные средства, альтернативное топливо, автобусы, операции флота, топливные элементы и др. Далее идут кластеры небольшого размера, относительно уже разобранных, они могут остаться незамеченными на карте без цветных маркеров. Первый из них – *compressed air* (см. рис. 6.15), закономерно находится в стороне кластера *natural gas* (см. рис. 6.13). Входящие в него понятия: энергоэффективность, хранилище энергии, накопитель энергии сжатого воздуха, сила ветра, расходы и др. Такое подробное разделение стало возможно на карте БД *Scopus*, скорее всего, из-за того, что она включает данные 3831 работы, в то время как карта БД *Web of Science* основана на 2023 работах. Те же понятия, касающиеся накопления энергии и сжатого воздуха, можно найти в кластере *Web of Science optimization*

(см. рис. 6.6). Кластер *gasoline* (см. рис. 6.16) включает термины: загрязнение воздуха, атмосферное загрязнение, дизельное топливо, качество воздуха, транспортные выбросы и др. Это может свидетельствовать о том, что в работах о природном газе бензин зачастую упоминается вместе с проблемами экологии. Возможно, исследователи чаще сравнивают влияние на окружающую среду, нежели эффективность или другие характеристики. Так, можно сделать вывод, что вопрос экологии гораздо важнее для ученых при рассмотрении бензина и газа. Кластер *liquefied petroleum gas* (см. рис. 6.17) довольно незначителен, он не имеет терминов, которые относятся непосредственно к нему. Он находится между кластерами *natural gas* (см. рис. 6.13), *compressed natural gas* (см. рис. 6.12), *gas emissions* (см. рис. 6.14) и *gasoline* (см. рис. 6.16). Таким образом, сжиженный нефтяной газ является важным понятием, но не имеет большого количества посвященных ему публикаций.

В рамках изучения публикаций и практики финансирования НИОКР, влияющих на рост применения природного газа в качестве моторного топлива, нами был проведен анализ базы данных ЕГИСУ НИОКТР с 2014 г. и выявлено: 301 информационная карта реферативно-библиографических сведений, 284 сведений о результатах научно-исследовательских и опытно-конструкторских работ, 102 диссертации, 22 результата интеллектуальной деятельности, содержащих следующие ключевые слова и их производные: *топлив природн газ, природн газ топлив, газов топлив, газов моторн топлив, газомоторн, газов двиг, авт газ топлив, газозаправ, метанов топл, газодиз, газов дизельн, компримированн газ, компримирприродн газ*.

Наиболее популярные слова из перечня ключевых слов и их словоформ, которые встречаются в научных работах, отобранных для анализа из ЕГИСУ НИОКТР представлены на рис. 6.18, а на рис. 6.19 – облако словоформ, содержащихся в ключевых словах карточек НИОКР, результатов интеллектуальной деятельности (РИД), диссертаций и информационных карт реферативно-библиографических сведений (ИКРБС) для этих работ.

Рис. 6.18. Наиболее популярные ключевые слова, встречающиеся в НИОКР, связанных с использованием газомоторного топлива

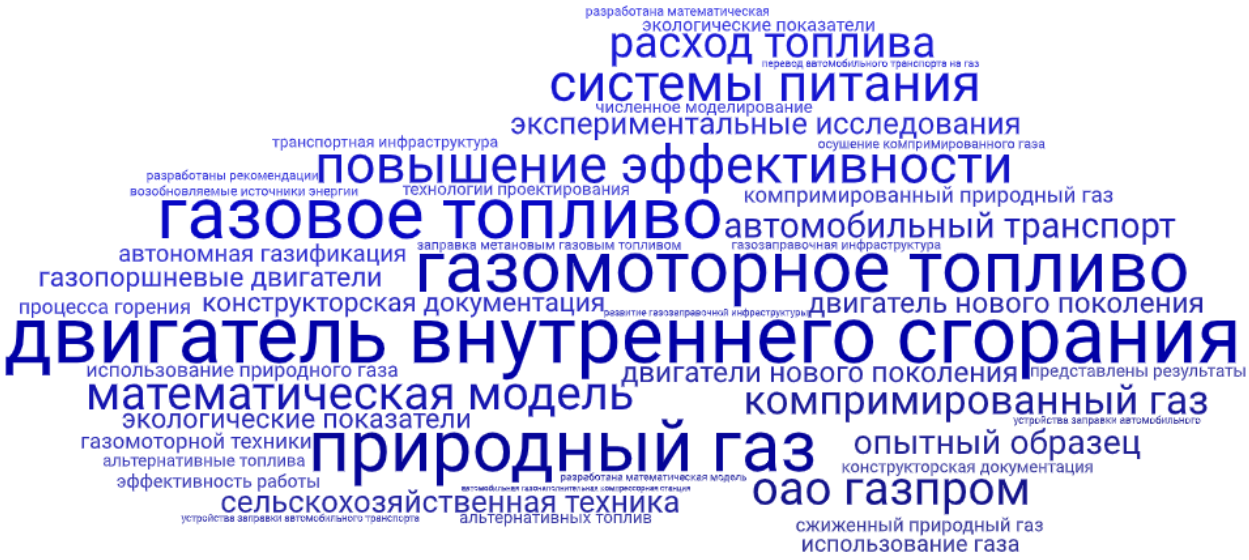


Рис. 6.19. Облако словоформ, содержащихся в ключевых словах карточек НИОКР, РИД, диссертаций и ИКРБС, связанных с использованием газомоторного топлива

Исходя из представленных изображений (см. рис. 6.18 и 6.19), видно, что публикации, попавшие в выборку для дальнейшего анализа, имеют прямое отношение к природному газу, топливу и двигателям.

Структура и объем выполненных исследовательских работ по видам (диссертация, НИОКР, РИД) в разрезе по годам представлена на рис. 6.20.

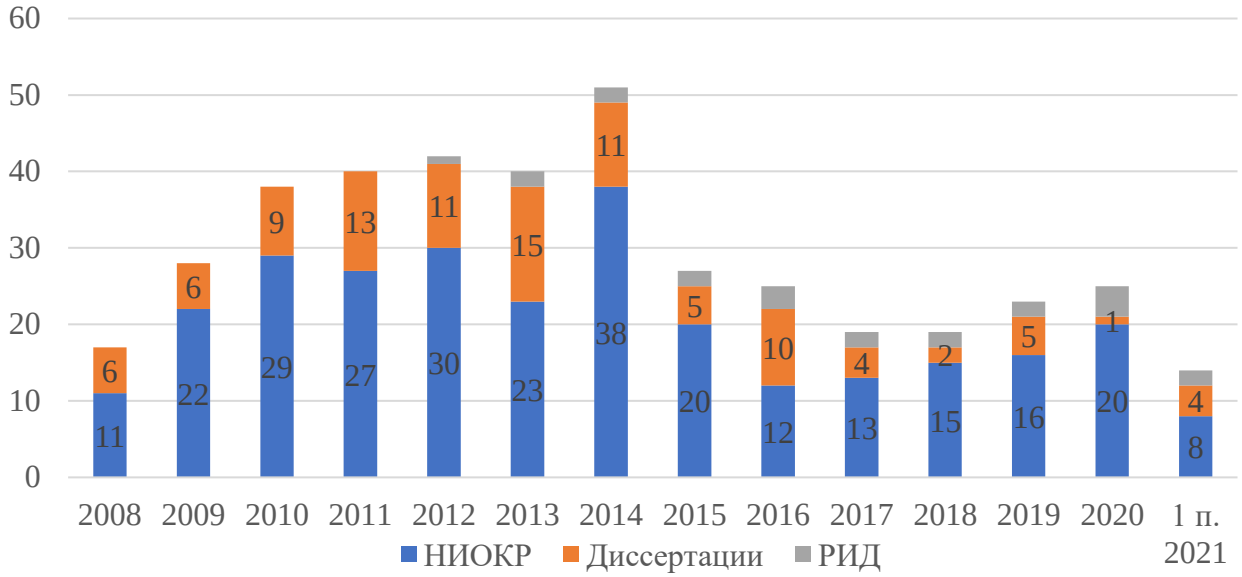


Рис. 6.20. Количество защищенных диссертаций, зарегистрированных НИОКР и РИД, связанных с использованием природного газа, газомоторного топлива и т.д.
(по оси У – количество исследовательских работ)

Материалом исследования выступали такие библиометрические данные, как: ключевые слова, аннотации, указанные тематики и т.д. Попробуем взглянуть на подходы ученых к теме КПГ с диахронической точки зрения и материалом будет служить совокупность самих публикаций.

Тексты статей находятся в распоряжении журналов, которые предоставляют доступ к полным текстам в основном через подписку. Часть статей публикуется без ограничений доступа и его можно получить через базу данных журналов (например, Scopus, Web of Science и eLibrary). Однако есть ограничения на объемы скачивания. Для того, чтобы избежать лишних трудностей и протестировать способ анализа текста нами был выбран сайт cyberleninka.ru. КиберЛенинка – русскоязычная научная электронная библиотека открытого доступа. Её основная задача – популяризация науки и научной деятельности. На момент проведения настоящего исследования в библиотеке насчитывалось около 2,5 млн статей.

По поисковому запросу «компримированный природный газ» было найдено 960 статей. С целью эффективного скачивания был написан программный код, позволяющий автоматически переходить по всем страницам, собирать ссылки на статьи, а затем скачивать их. Эта программа составлена с помощью Jupyter Notebook – инструмента для интерактивной разработки проектов. Всего было создано три программы, написанные на языке программирования python (при использовании Jupyter Notebook python является наиболее распространенным вариантом, но данная среда поддерживает и другие языки) [267]. Задача второй программы – преобразование скачанных статей в формате pdf в формат txt. Такая задача уже имеет решение, и большая часть программы состояла из готового кода.

Работа с текстами статей и обработка происходили через программу, которая один раз перебирала слова из *txt*-файлов и составляла из них словарь, где ключом является год, а значением – список слов, встретившихся в статьях этого года. Для поиска все слова прошли лемматизацию, чтобы разные формы одного слова не считались разными словами. На текущем этапе программа предназначена для поиска односоставных и двусоставных терминов. Пользователь делает поисковый запрос и получает две колонки, года, начиная с 2004 г., и частотность, количество словоупотреблений запроса в статьях каждого года. В файлах, полученных с сайта, много лишних символов, связанных с форматом pdf и разными обозначениями в статьях, «очистить» полностью данные весьма непростая задача, но она не обязательна для решения, так как это не влияет на результат – получение данных об определенном запрашиваемом термине. Программа специализирована на работу с сайтом cyberleninka.ru, потому что на разных сайтах статьи хранятся по-разному. Ввиду особенностей сайта нам не удалось скачать последние статьи. Таким образом количество рассматриваемых статей составило 842 (также были вычтены статьи 2021 г.).

Задавать поиск было решено по словам, оказавшиеся самыми частотными по результатам выше проведенного библиографического анализа (См. таблицу 6.2.).

Таблица 6.2. Частотность употребления терминов по годам

Год (кол-во статей)	Слова, частота					
	газовая индустрия	выбросы газа	бензин	загрязнение	транспорт	трактор
2004 (8)	0	0	0	0	4	0
2005 (4)	0	0	1	0	0	9
2006 (13)	2	0	10	2	25	4
2007 (13)	3	0	45	0	41	0
2008 (63)	11	1	144	23	416	44
2009 (52)	10	1	154	20	287	13
2010 (57)	18	2	179	21	487	23
2011 (62)	9	0	192	39	305	33
2012 (64)	9	3	124	16	238	32
2013 (90)	15	0	145	33	460	71
2014 (69)	26	1	104	36	346	64
2015 (77)	51	0	75	30	377	12
2016 (84)	37	1	88	43	471	49
2017 (96)	132	3	79	27	457	9
2018 (76)	164	3	66	27	505	3
2019 (72)	61	1	74	32	353	4
2020 (40)	12	0	46	21	104	3

Термин «газовая индустрия» в какой-то степени использовался почти всегда, но его употребление резко подскочило в 2017 и 2018 гг. Когда происходило активное развитие газовой промышленности, и в 2018 г. отмечается, что Россия вышла на рекордные объемы добычи и экспорта газа, доходы от сжиженного природного газа выросли на 83%.

В 2007 – 2015 гг. выделяется тенденция, связанная с бензином, пик которой наблюдается в 2011 г. (рис. 6.22). Рост и спад публикационной активности на протяжении нескольких лет означает, что причиной является постепенный процесс, без неожиданных сенсаций или происшествий. Например, в 2011 г. автомобильный метан вышел на мировой рынок, и этому предшествовали работы рассматривающие последствия появления нового типа двигателя. Однако двигатель с газожидкостными циклами не заменил полностью традиционный дизельный двигатель и интерес к ним медленно пошел на спад. Часто при сравнении бензина и КПГ поднимается тема загрязнения окружающей среды,

количество упоминаний загрязнения и бензина достигает самого большого значения в один и тот же год.

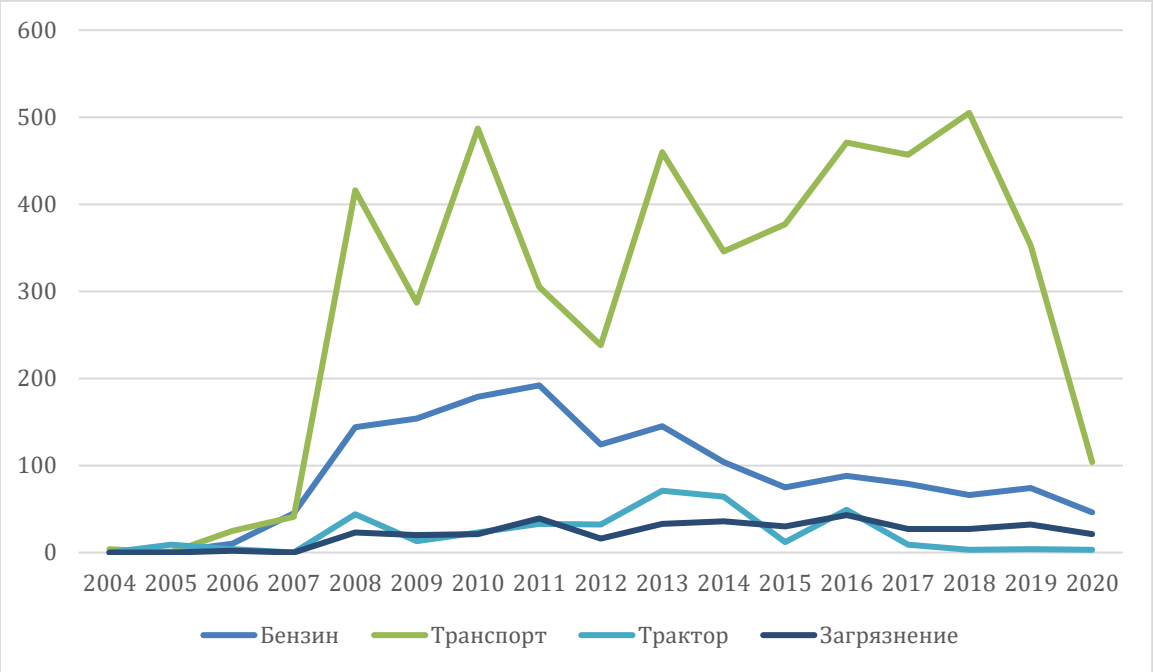


Рис. 6.22. Частотность слов по годам. (по оси У – количество исследовательских работ)

Изучение научных процессов в сфере газовой отрасли требует наличия различных компетенций, для тщательного рассмотрения темы нужна команда исследователей. Однако, поставленная в нашей работе цель заключалась в проверке способов выявления научных тенденций, используя различные инструменты. С помощью баз данных научного цитирования Web of Science, Scopus и eLibrary получены библиометрические данные для последующего анализа. В программе VOSviewer создана карта, показывающая связи терминов на основе совместного использования, которая дает представление о закономерностях использования понятий, давая возможность судить о подходах ученых к определенным явлениям. Так, с одной стороны, тема выбросов раскрывается в виде проблемы загрязнения транспортом с бензиновыми двигателями, а с другой – раскрывается как тема преимущества природного газа (расположение кластеров, посвященных бензину, природному газу и выбросам, вместе с принадлежащими им словами, показывает, в каком ключе

рассматриваются эти темы в публикациях). Библиометрический анализ по eLibrary выявил периоды роста и спада научной активности в данной сфере. Отдельный случай роста был продемонстрирован в исследовании НИУ ВШЭ.

Динамика запросов поисковой системе Google по регионам с 2011 по 2021 гг. показала особый интерес региона Нигерии, связанный с планами правительства сделать природный газ катализатором экономического развития страны. После сужения выборки более подробно мы рассмотрели тематику экологии. Для этого был разработан небольшой пакет программ для получения, обработки и исследования текстов статей из электронной библиотеки открытого доступа «КиберЛенинка». Чтобы выявить закономерности, внимание обращалось на изменение количества употреблений терминов в период с 2004 по 2020 гг. Само количество публикаций в год по теме компримированного природного газа изменялось, что свидетельствовало о популярности темы в целом. Частотность слов, в свою очередь, показывает, чему уделялось больше внимания, а значит, представлялось более важным в этой сфере на соответствующий момент.

По результатам анализа источников и практики финансирования научно-исследовательских и опытно-конструкторских работ было обнаружено, что в центре внимания исследований и разработок, касающихся применения природного газа в качестве моторного топлива, (помимо понятий «двигатель внутреннего сгорания» и газомоторное топливо») находятся понятия «двигатель нового поколения», «сельскохозяйственная техника», «повышение эффективности», «энергоэффективность». Из этого можно сделать вывод, каким образом исследователи подходят к этой теме, каковы их цели, в каком направлении движется эта сфера. Сам метод библиометрического анализа, по ключевым словам. Единой государственной системы учета показал свою надежность, результаты сходятся с результатами анализа данных из БД Web of Science и Scopus, в которых также выделяются экологический аспект и аспект энергоэффективности.

6.3. Метод формирования нормативного профиля требований к объекту сертификации

Сертификация – признанный в мире способ независимой оценки соответствия продукции, процессов и услуг установленным требованиям. Использование сертификации создает предпосылки для успешного решения ряда важных социальных и экономических проблем и задач. Главной целью сертификации программных средств и систем качества, обеспечивающих их жизненный цикл, является контроль и заверение качества технологий и продукции, гарантирование их высоких потребительских свойств [268].

Практическое проведение сертификации на этапе формирования программного обеспечения до настоящего времени в значительной мере заключается в ручном анализе экспертами больших объемов нормативной и проектной документации, представленной текстами на естественном языке. Это приводит к определенному субъективизму экспертных оценок, снижению их полноты и достоверности. Необходимость преодоления указанных проблем обусловливает актуальность разработки и использования компьютеризированных методов формирования нормативного профиля для сертификации программного обеспечения [269].

В настоящее время существует определенное количество методов и подходов представления требований, такие как использование формальных языков описания (OCL,BRML) [270], представление в виде графа в одном из средств поддержки требований (IBM Rational Requisite Pro, DOORS, Borland Caliber RM и некоторые другие) [271], составление графических моделей требований с использованием диаграмм (IDEF0, IDEF3, DFD, UML, OCL, SysML, ARIS) [272], формализация требований по использованию фасето-иерархических структур [273] и прочее. Такие представления могут быть удобными во время частого внесения изменений в требования, отслеживание выполнения, во время организации “привязки” к требованиям задач, персонала, тестов и кода [274]. Однако большинство требований нормативных документов носят качественный характер и отражают базовые и наиболее устойчивые

аспекты создания программного обеспечения. Во время реализации конкретных программных проектов эти требования должны быть детализированы и дополнены показателями, которые оценивают степень их выполнения. Наиболее распространенным является представление требований на естественном языке в виде сформулированных человеком высказываний, из содержания которых можно однозначно установить количественные и качественные критерии, которые выдвигают для определенных характеристик системы.

Требования к программному обеспечению представлены в виде текстов нормативной базы на естественном языке, однако, стандартные средства семантической обработки текстов не могут придать нужного уровня глубины и полноты нормативного профиля, так как не учитывают композиционную структуру и обобщенно-отвлеченный характер лексики текстов стандартов. В настоящее время под обработкой естественного языка понимают обработку документов, представляющих собой текст в виде набора предложений, при этом структурные особенности текста документа зачастую не рассматриваются [275, 276]. Сложность информационного поиска в текстах стандартов также обусловлена обобщенно-отвлеченным характером лексических единиц не только в текстах требований, но и названий заголовков и подзаголовков [277].

В связи с вышесказанным актуальной становится задача обработки документов с выраженной иерархической структурой, к которым относятся документы нормативной базы программной инженерии. Учет композиционных и стилистических особенностей таких текстов позволит разработать новые инструменты, позволяющие повысить эффективность работы эксперта, путем автоматизации рутинной работы [278].

При поиске требований в коллекции текстов стандартов недостаточно получить лишь список релевантных документов в качестве поисковой выдачи из-за значительного объема и высокой сложности документов. Повышение эффективности поиска в таких документах можно достичь, если в качестве поисковой выдачи будут получены не только документы, но и цитаты из них – точные дословные выдержки из текста, имеющих смысловую завершенность.

Цитаты можно получить с помощью анализа композиционной структуры текстов стандартов и нормативной документации, а потом могут быть уточнены через использование семантического анализа. В результате должна быть получена компактная поисковая выдача, где отсечен значительный объем информации, нерелевантный запросу.

Формирование нормативного профиля как одного из основных этапов сертификации программного обеспечения. Одной из первых и, безусловно, главных задач как при разработке, так и сертификации информационно-управляющих систем является составление требований к системе. Как показывает опыт индустрии информационных технологий и анализ работ в указанной области [279-280], вопросы, связанные с созданием корректных требований к системе и грамотное управление требованиями имеют критически важное влияние на проекты и возможность их успешной реализации.

Нормативный профиль содержит международные и отраслевые стандарты, а также различные регулирующие документы, разработанные в определенной предметной отрасли и определяющие функциональность широкого круга систем. Часть требований, сформулированных в указанных документах, также может быть использована для конкретного проекта и должна входить в частный профиль требований этого проекта. Задачу разработки требований к конкретному продукту можно рассматривать как задачу разработки частного профиля требований на основе общего профиля через использование необходимых требований из общего профиля в качестве параметров. Необходимо заметить, что следует различать три вида нормативных профилей. Общий нормативный профиль - совокупность международных и национальных стандартов или других нормативных документов, используемых в программной инженерии. Частный нормативный профиль – это совокупность нескольких (или подмножество одного) базового стандарта и других нормативных документов с четко определенными и гармонизированными подмножествами обязательных и факультативных возможностей, предназначенных для реализации заданной функции или группы функций. Нормативный профиль к объекту сертификации

– это совокупность требований нормативной базы, предъявляемых к объекту сертификации и предназначенных для соответствия этим требованиям.

При подготовке к проведению оценки программного обеспечения на соответствие требованиям эксперт осуществляет распределение требований по типам и группирует их по удобному ему принципу, например, требования к программным компонентам системы, требования к характеристикам надежности и тому подобное.

В [268] выделяют следующие варианты формирования нормативного профиля, как комбинации нормативных документов и их частей из профиле-образующей базы: формирование нормативного профиля на основе только одного нормативного документа, взятого из профиле-образующей базы в полном объеме; формирование нормативного профиля на основе определенной части только одного нормативного документа, взятого и профиле-образующей базы; формирование нормативного профиля на основе двух и более нормативных документов, взятого из профиле-образующей базы в полном объеме; формирование нормативного профиля на основе частей двух или более нормативных документов, взятого из профиле-образующей базы в полном объеме; формирование нормативного профиля на основе одного базового профиле-образующего документа, дополненного частями одного или нескольких нормативных документов из профиле-образующей базы.

Нормативная база сертификации программного обеспечения представлена нормативными документами трех уровней [270]: международного, принятого международными организациями; регионального, принятого региональной организацией по стандартизации; национального, принятого национальными органом по стандартизации. Обзор нормативной базы программной инженерии в разработке систем с интенсивным использованием программного обеспечения представлен в [281].

Модель ядра семантической целостности для автоматизации процедуры сертификации. Наличие у лексической единицы двух и более

значений, исторически обусловленных или взаимосвязанных по смыслу и происхождению, а также эквивалентность всего объема лексических единиц или их значений обуславливают необходимость разработки специальных средств, способных, с одной стороны, расширить запрос путем добавления синонимических лексических единиц в запрос, а с другой - выбрать нужное значение полисемантической лексической единицы [282]. Таким образом, становится необходимым синтез ядра семантической целостности запроса пользователя, отражающий отношения между предметными и предикатными лексическими единицами [283].

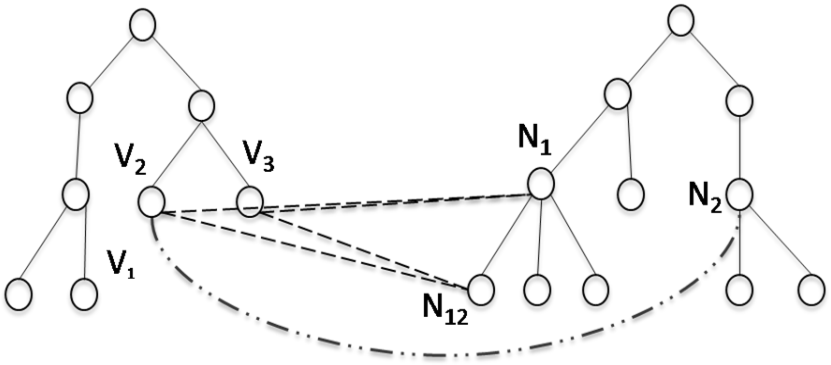
Обобщенная модель ядра семантической целостности должна обладать достаточной универсальностью, в частности, быть пригодной для сжатия текстовой семантической информации, что является основой построения моделей для экспертной системы поддержки принятия решений сертификационных аудитором при экспертировании программного обеспечения. Она отражает содержание предложения на понятийном уровне, независимо от применяемых лексических единиц и синтаксических конструкций, их сочетающих.

Модель ядра семантической целостности *C* для формирования нормативного профиля при сертификации программного обеспечения целесообразно представить следующим образом:

$$C =< V^{S_i}, N^{S_j} >$$

где V – предикатная лексическая единица, N– предметная лексическая единица, S – множество семантических падежей, состоящее из 8 элементов: s1 – агент, s2 – объект, s3 – контрагент, s4 – адресат, s5 – пациенс, s6 – результат, s7 – инструмент, s8 - источник; $i, j = (\overline{1,8})$.

На Рис. 6.23 изображена модель ядра семантической целостности, где пунктирными линиями показаны возможные семантические сочетания слов, а пунктирной линией с двумя точками – слова, семантически не сочетающиеся друг с другом в текстах языка стандартов.



V_1 – проводить	N_1 – документация к ПО
V_2 – отмечать	N_{12} – отчет о тестированиях ПО
V_3 – включать	N_2 – верификация ПО

Рис. 6.23. Модель ядра семантической целостности

Рассмотрим семантическую сочетаемость предметных и предикатных лексических единиц на следующих примерах:

В отчете о тестированиях (N_{12}) должны быть отмечены (V_2) все расхождения между проектом и реализацией, обнаруженные в процессе тестирований

Отчет о тестированиях программного обеспечения (N_{12}) должен включать (V_3) следующие пункты, как на уровне модуля, так и на уровне основного проекта…

Так предметные лексические единицы N_{12} и N_1 требуют семантический падеж результат s_6, N_2 - инструмент s_7. Глаголы отмечать V_2 / включать V_3 в значении «отражать некий результат в виде «документа / отчета» сочетаются с N_{12} и N_1, но не сочетаются с глаголом проводить V_1, которое имеет значение «пользоваться инструментом «верификация», так как он требует семантический падеж инструмент s_7.

Модель ядра семантической целостности предложения учитывает семантическую сочетаемость слов естественного языка, а также позволяет решить проблему многозначности глаголов из-за учета семантических характеристик предметных лингвистических единиц, сопрягаемых с глаголом.

Описанная выше модель ядра семантической целостности языковых объектов из предметной области сертификации программного обеспечения является формальной основой для создания системы онтологий, содержащих знания концептуального характера о смысловой структуры нормативной базы программной инженерии, подвергающиеся процедуре сертификации. Применение модели ядра семантической целостности позволяет обеспечивать возможность реализации онтологического среза и формирования на его основе отчетов в ответ на запросы аудитора сертификационного центра.

Синтез онтологической системы для формирования нормативного профиля. Онтологическую систему для формирования нормативного профиля при сертификации программного обеспечения показано на рис. 6.24. В ее состав входит: онтология критериев качества программного обеспечения, онтологии стандартов, онтология предметных лексических единиц, онтология предикатных лексических единиц, онтология запроса, онтология ядра семантической целостности [284-286].

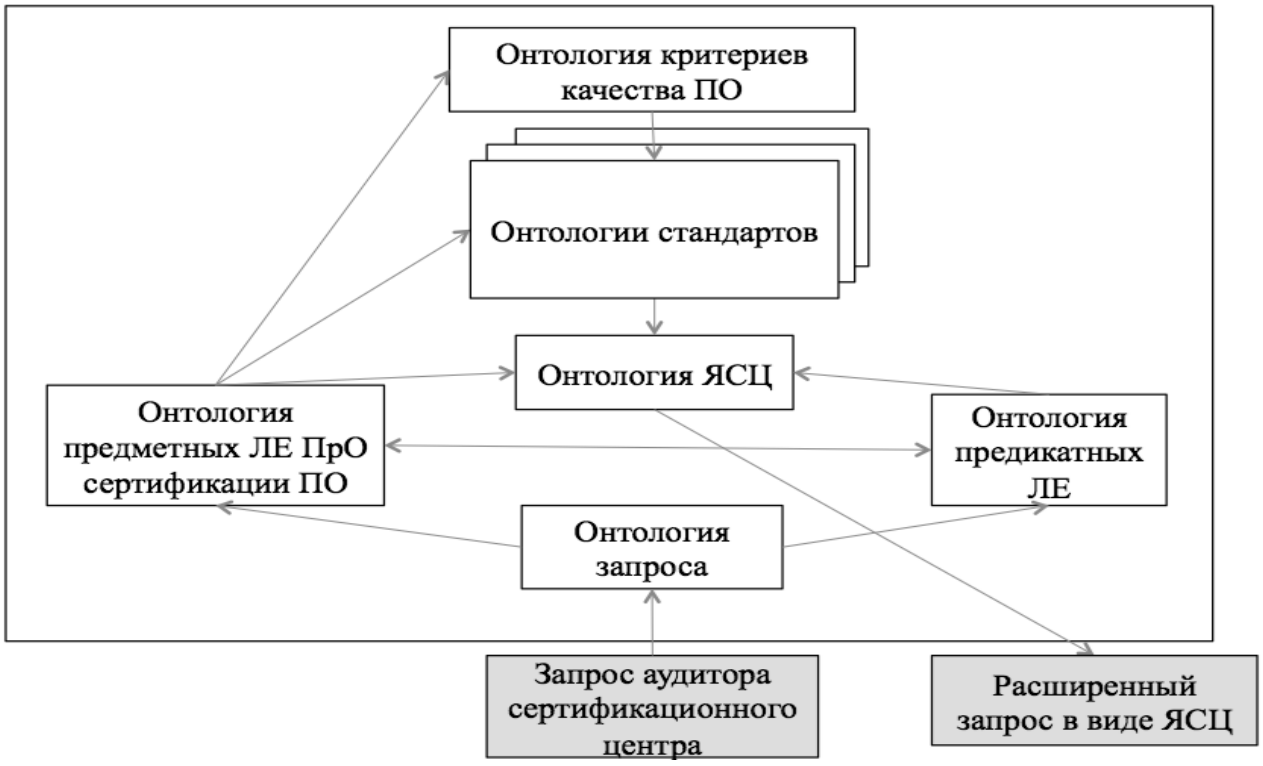


Рис. 6.24. Онтологическая система для формирования нормативного профиля требований к программному обеспечению

На вход подается онтология запроса A_1, сформированная на запрос аудитора сертификационного центра, где N_1 – предметная лингвистическая единица, V_1 – предикатная лингвистическая единица, выделении из запроса, поданного в виде полного вопроса. Затем необходимо выделить предметную и предикатную лингвистические единицы N_1 и V_1. В онтологии предметной области сертификации программного обеспечения осуществляется поиск предметной лингвистической единицы N_1.

Предположим, что запрос аудитора сертификационного центра звучит следующим образом: *Что включает в себя отчет о верификации программного обеспечения? Отчет о верификации программного обеспечения* является предметной лексической единицей, а *включать* – предикатной лексической единицей соответственно. На Рис. 6.25 представлен фрагмент онтологии предметных лексических единиц, и выделена лексическая единица из запроса.

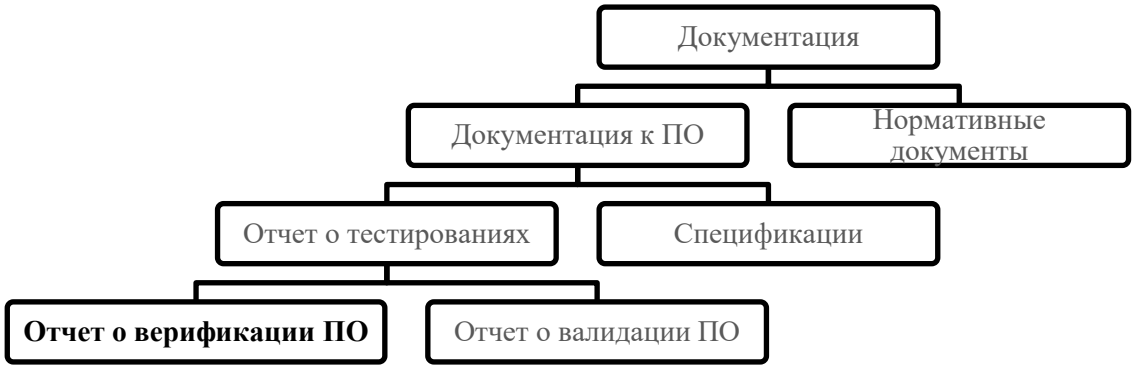


Рис. 6.25. Фрагмент онтологии предметных лексических единиц

На следующем этапе (Рис. 6.26) запрос необходимо расширить запрос путем добавления родовидовых понятий, относящихся к запросу.

Рис. 6.26 Фрагмент онтологии предметных и предикатных лексических единиц

Таким образом, к поисковому запросу добавим предметные лексические единицы *документ* и *отчет о тестированиях*, и предикатные лексические единицы, обозначающие фиксацию информации, – *содержать, отмечать, указывать*, а ядро семантической целостности *С* будет иметь вид:

С = (отчет о тестировании программного обеспечения × включать)∨ (отчет о тестировании программного обеспечения × отмечать) ∨ (отчет о тестированиях программного обеспечения × содержать) ∨ (отчет о тестировании программных обеспечение × подавать) ∨ (отчет о тестировании программного обеспечения × отображать) ∨ (отчет о тестировании программного обеспечение × указывать) ∨ (отчет × включать) ∨ (отчет ∧ отмечать) ∨ (отчет × содержать) ∨ (отчет × подавать) ∨ (отчет × отображать) ∨ (отчет × указывать) ∨ (документ × включать) ∨ (документ × отмечать) ∨ (документ ×содержать) ∨ (документ × подавать) ∨ (документ × отражать) ∨ (документ × отмечать).

На основе перечня текстовых файлов стандартов и перечня запросов для информационного поиска требований к объекту сертификации проводится поиск требований. Результатом этого этапа является перечень цитат из текстов стандартов, содержащих требования к объекту сертификации. Пример релевантной поисковой выдачи на запрос аудитора сертификационного центра [из стандарта МЭК 60880]:

Отчет о тестированиях программного обеспечения

8.2.3.1.3.1 В отчете об тестированиях программного обеспечения должны быть представлены результаты верификации, описанные в спецификации тестирований программного обеспечения и устанавливающие, работает или нет программное обеспечение в соответствии со спецификацией проекта программного обеспечения.

8.2.3.1.3.2 В данном документе должны быть отмечены все расхождения между проектом и реализацией, обнаруженные в процессе тестирований.

8.2.3.1.3.3 Отчет о тестированиях программного обеспечения должен включать следующие пункты, как на уровне модуля, так и на уровне основного проекта…

В связи с тем, что лексические единицы в текстах стандартов носят обобщенно-отвлеченный характер, а значимая информация содержится в названии самого стандарта либо его разделов, также проводится поиск стандартов, в заголовках которых содержатся предметные лексические единицы. Пример нерелевантной поисковой выдачи на запрос аудитора сертификационного центра [из стандарта МЭК 60880]:

Программные аспекты отчета о валидации системы

10.3.1 В отчете о валидации системы должны быть отражены результаты программных аспектов валидации системы.

10.3.2 В отчете должны быть указаны техническое обеспечение, программное обеспечение и конфигурация использованной системы, а также использованное оборудование и его калибровка и использованные модели при моделировании.

10.3.3 В данном отчете также должны быть указаны любые отклонения.

10.3.4 В данном отчете должны быть обобщены результаты валидации системы.

10.3.5 В данном отчете должна быть дана оценка соответствия системы всем требованиям.

На следующем этапе на основе перечня требований происходит структурирование нормативного профиля в соответствии с классификатором нормативной базы или вариантов формирования нормативного профиля, как комбинации нормативных документов и их частей из профиле-образующей базы. Результатом этого этапа является нормативный профиль в виде текстового файла, в котором представлен структурированный перечень требований к объекту сертификации.

На Рис. 6.27 представлена структура нормативного профиля для

сертификации программного обеспечения. Такое представление нормативного профиля упрощает работу аудитора сертификационного центра за счет представления нормативного профиля как совокупности цитат к объекту сертификации из текстов стандартов. То есть аудитор работает не с полными текстами стандартов, а с точными дословными выдержками из текстов стандартов.

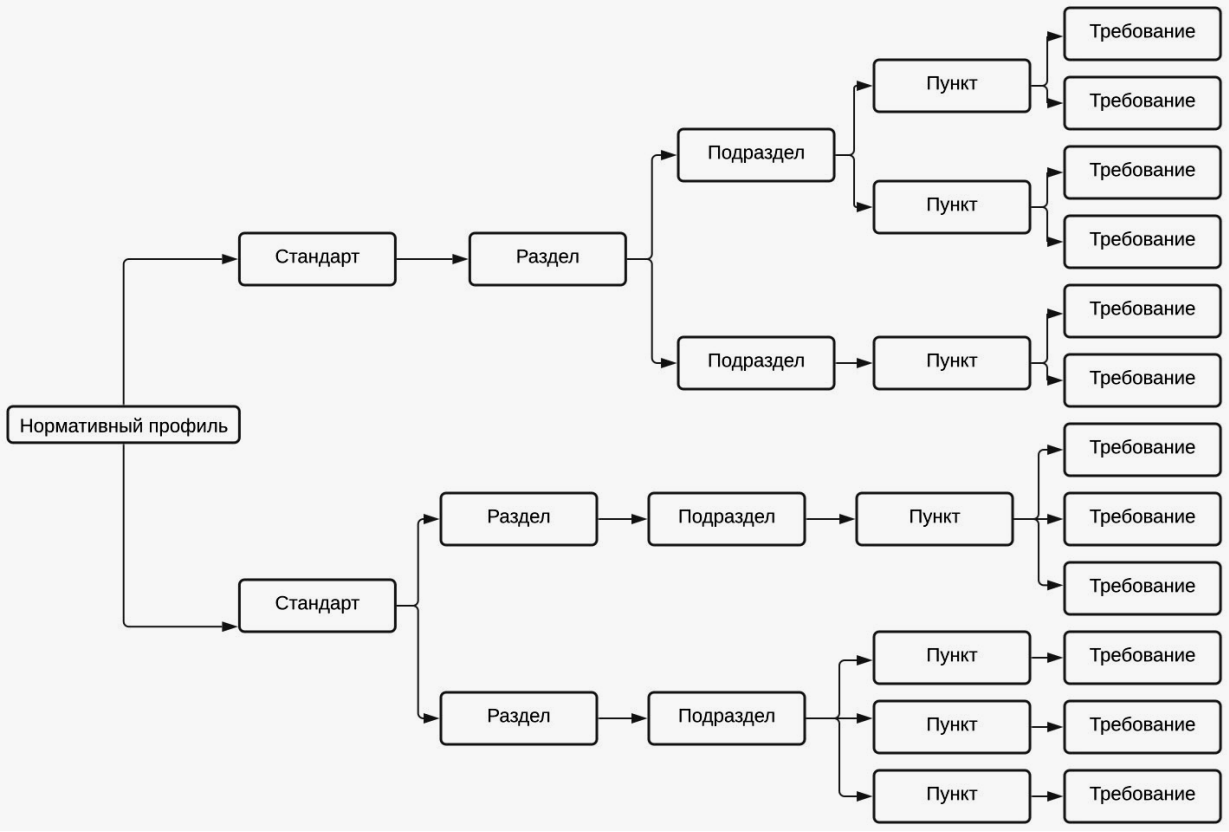


Рис. 6.27. Структура нормативного профиля при сертификации программного обеспечения

6.4. Метод информационного поиска в базе знаний о конструкции летательных аппаратов на основе падежной грамматики

Поддержка принятия решений в сложных ситуациях необходима в различных областях деятельности, где требуются обоснованные, логически доказуемые аргументы, в том числе в оперативном и стратегическом управлении технологиями на авиационных предприятиях. При принятии решений необходимо проведение детального моделирования последствий предполагаемого решения, поиск оптимального пути достижения заданного

результата при помощи технологий имитационного моделирования. С подобными задачами успешно справляются экспертные системы – комплексы программных средств, способные частично или полностью заменять специалиста при решении сложных задач, возникающих в процессе диагностики, проектирования или эксплуатации аэрокосмической и другой техники [39, 287].Такие системы основаны на знаниях, полученных в процессе взаимодействия с экспертами в конкретной предметной области.

Параллельно с развитием различных направлений исследований в области искусственного интеллекта происходит развитие информационных структур для представления знаний [288]. Появились новые способы описания и представления данных, возникли фреймовые, списочные и иерархические структуры. Представление знаний – это одно из направлений в исследованиях по искусственному интеллекту, изучающее способы описания объектов реального мира. Другое, не менее важное направление – это манипулирование знаниями. Сам процесс построения баз знаний достаточно сложен, плохо структурирован и носит итеративный характер, заключающийся в циклической модификации баз знаний на основе результатов ее тестирования [289-290].

Основной задачей проектирования базы знаний в области авиакосмического приборостроения является описание структуры летательного аппарата таким образом, чтобы оно предоставляло наиболее полную и непротиворечивую информацию об описываемом объекте. Конструкции самолетов и космических аппаратов требуют наличия сложных конструктивных элементов, процесс стыковки и крепления которых представляет собой не менее сложную задачу.

Если применить лингвистические модели в совокупности с математическим аппаратом, возможно достичь более ясного и результативного описания и решения сложной технологической проблемы. Информация воспринимается специалистом на когнитивном уровне посредством основных единиц этого уровня – понятий, идей, концептов, поэтому все чаще используются семантические модели для представления знаний, которые

воссоздают модель, схожую с процессами мышления конструктора, технолога [291]. Ярким примером подобной модели является семантическая сеть, под которой в инженерии знаний подразумевается граф с узлами, отображающими объекты предметной области, и дугами, обозначающими отношения между данными понятиями.

Необходимо создание модели, способной сохранять соответствие с описываемым технологическим процессом без снижения эффективности обработки информации. Для описания конструкций летательных аппаратов разработана интегрированная модель представления знаний – абстрактная иерархическая структура, объединяющая графовую структуру, фреймы и предикаты. Для реализации подобной структуры данных требуется принимать во внимание и предикатные символы, которые обеспечивают описание сущности, представленной в данной структуре. Пример такой структуры данных показан на рис. 6.28.

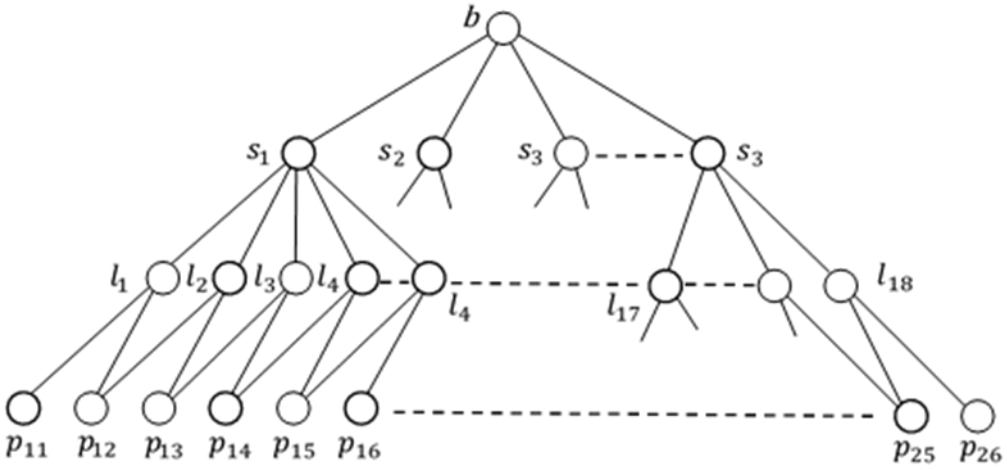


Рис.6.28. Абстрактная иерархическая структура для описания крыла летательного аппарата

Все знаний о предметной области представлены в виде графа, в узлах которого содержатся фреймы – минимально возможные описания сущностей, которые записаны с помощью предикатов и предикатных символов. Подобная структура гарантирует непротиворечивость описаний объектов и их отношений в реальном производстве.

Неотъемлемой частью любого управления является его документационное обеспечение, основа которого – однозначно понимаемая и непротиворечивая терминология. Стандартизированная терминосистема является важным условием технического прогресса, улучшения качества и надежности продукции. Единое терминологическое обеспечение позволит избежать недопонимания между специалистами как в рамках одной предметной области, так и за ее пределами.

Терминосистема – это некоторая лингвистическая модель, представляющая конкретную предметную область [292]. Аэрокосмическая терминология русского языка представляет собой сложную систему, включающую термины из различных областей знания и различных тематических пространств: обозначения субъектов аэрокосмической сферы (*космонав*т, *летчик, пилот*), наименования организаций и структур области авиации и космонавтики (*НАСА*, *ИАТА, ИФАТКА*), обозначения действий аэрокосмической деятельности (*взлет, посадка, пилотирование, пристыковка*), наименования деталей и частей летательных аппаратов (*фюзеляж, крыло, подкрылки, двигатель*), обозначение характеристик и состояний летательных аппаратов (*работоспособность, надежность, безопасность*). Стандартизация терминологии данной предметной области является трудоемким процессом, а отсутствие стандартизации приводит к появлению различных терминов для обозначения одних и тех же объектов и явлений, в связи с чем возникают явления синонимии и лексической многозначности.

При построении баз знаний интеллектуальных систем, особенно в области авиакосмического приборостроения, где процесс проектирования летательных аппаратов требует полного взаимопонимания между специалистами, синонимия является крайне неблагоприятным явлением. Различия в номинации объектов приводят к трудностям в сфере профессиональной коммуникации, поэтому базу знаний экспертной системы предлагается пополнять всеми возможными вариантами синонимов тех или иных узкоспециальных и общенаучных терминов.

В качестве иллюстрации приведем несколько примеров синонимичных терминов, выделенных на основе анализа синонимического словаря авиационных терминов Н.С. Шарафутдиновой [293]:

-*интерцептор – спойлер;*

-*пусковое устройство – пусковая кнопка – стартер;*

-*шкала дальности – дистанционная шкала;*

-*приборная доска–приборная панель;*

В аэрокосмической терминологии также наблюдается семантическая оппозиция номинаций «свой-чужой», откуда возникают синонимичные пары, например:

-*космонавт – астронавт (англ. astrounaut);*

-*космонавтика – астронавтика (англ. astrounautica);*

-*высотомер – альтиметр (франц. altimètre).*

Равнозначность подобных синонимов необходимо отразить в проектируемой базе знаний во избежание недопонимания между экспертной системой и ее пользователями.

Стоит отметить, что редко используемые аббревиатуры должны быть расшифрованы и понятны пользователям экспертной системы (например, *ШВП* – это *шасси на воздушной подушке* или *шаровой вытяжной парашют*).

Серьезные трудности в процессе поиска информации в базе знаний вызывают явления лексической многозначности – омонимия и полисемия. Под полисемией понимается сосуществование различных, но связанных между собой значений одного слова или фразы. Слово является многозначным, и разница между значениями данного слова часто оказывается вполне очевидной. Иногда сложно определить, является ли слово полисемантичным или нет, потому что отношения между словами могут быть неясными. Омонимией называется существование двух или более слов, имеющих одинаковое написание или произношение, но разные значения и происхождение.

Полисемантичные и омонимичные терминологические единицы в пределах области авиакосмического приборостроения относительно редки, но не исключены полностью, например:

Свеча – 1. Приспособление для воспламенения горючей смеси (в двигателях внутреннего сгорания); 2. Крутой взлет, подъем (фигура высшего пилотажа).

Хвост – 1. Хвостовое оперение летательных аппаратов; 2. Вытянутое из пыли и газа кометное вещество, образующееся при приближении кометы к Солнцу.

Так возникает проблема многозначности поискового запроса, которую можно решить путем выявления контекста конкретного термина. Обычно экспертная система ведет с пользователем диалог в вопросно-ответной форме, в процессе которого система может уточнить у пользователя о каком значении многозначного термина идет речь. Например, в процессе поиска информации диалоговая экспертная система может задать пользователю следующий вопрос: «*Свеча – фигура высшего пилотажа?*». От пользователя последует положительный или отрицательный ответ, что приведет к сужению пространства информационного поиска для нахождения нужного ответа, как представлено на рис.6.29.

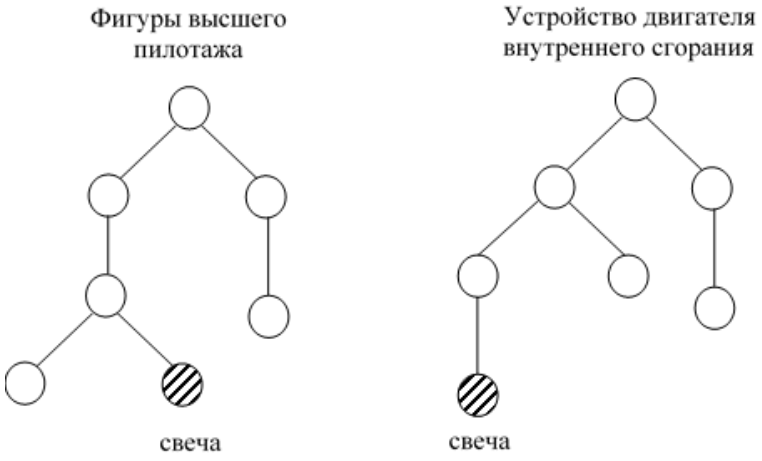


Рис. 6.29. Сужение контекста поискового запроса

Методом определения контекста пользователь сам выбирает интересующие его понятия в базе знаний, на основе которых система определит

ближайшие концепты, связанные с запросом пользователя. Такой метод может обеспечить наиболее эффективный поиск информации в базе знаний экспертной системы.

Одним из необходимых компонентов базы знаний экспертной системы является механизм вывода, который одержит правила для решения конкретной задачи. Механизм вывода ссылается на информацию из базы знаний и выбирает факты и правила, которые будут применяться при попытке ответить на запрос пользователя. Данный механизм обеспечивает аргументацию информации в базе знаний и позволяет человеку получать информацию из базы знаний в виде предложений, построенных на естественном языке. Для вывода информации из экспертной системы необходима модель, которая позволила бы учитывать и структурировать взаимосвязи между понятиями в определенной предметной области. Одним из наиболее удачных механизмов, позволяющих моделировать смысл высказываний на естественном языке является механизм, основанный на теории падежных грамматик Ч. Филлмора [294]. В падежной грамматике семантика предложения рассматривается как система семантических валентностей. Валентность определяет число актантов (участников ситуации), которые может присоединять глагол. Глагол является центром предложения и диктует в силу своего значения набор ролей (глубинных падежей), реализующихся в предложении посредством именных форм. Иными словами, падежная грамматика показывает связь существительного или местоимения с другими словами в предложении. Обрамление падежами подвержено определённым ограничениям, например, каждый семантический падеж может встречаться в предложении только один раз. Некоторые падежи являются обязательными, другие - необязательными. Обязательные падежи нельзя удалять, рискуя получить грамматически неправильные предложения.

Для описания события в каждом фрейме иерархической структуры в первую очередь выделяется действие, которое обычно описывается глаголом. Далее определяются:

- объект, который действует - агенс;

- объект, над которым это действие выполняется - пациенс.

Количество падежей варьирует в трудах отечественных и зарубежных ученых (В.В. Богданов [295], Р.Джекендофф [296] и др.), однако набор универсальных глубинных падежей в основном совпадает в различных трактовках данной теории. В концепции Ч. Филлмора число и названия семантических падежей так же неоднократно изменялись. Обратимся к 6 универсальным падежам, выделенным Ч. Филлмором и взятым за основу во многих последующих теориях (таблица 6.3).

Таблица 6.3. Семантические падежи Ч. Филлмора

Падеж	**Значение**
Агентив	Объект, который производит действие; актант, участник ситуации.
Объектив	Наиболее размытый падеж; к нему обычно относят все те существительные, не относящиеся к другим падежам.
Датив	Одушевленное существо, затронутое выражаемым глаголом действием или ситуацией.
Инструменталис	Предмет, посредством которого совершается действие.
Фактив	Предмет, возникший в результате действия, прекративший существование или подвергшийся изменению.
Локатив	Место действия, выражаемое глаголом.

На каждое отношение накладывается множество ограничений, например, структура глагола «пилотировать» может включать следующие семантические падежи:

1) Агентив – ***Летчик*** *пилотировал самолет*;

2) Пациентив - *Летчик пилотировал* ***самолет*;**

3) Инструменталис – *Летчик пилотировал самолет* ***боковой ручкой управления*;**

4) Локатив – *Летчик впервые пилотировал самолет* ***на аэродроме*** *с плохим покрытием*.

Возможно включение с структуру данного глагола и других семантических падежей, а также атрибутивных отношений, например, «*пилотировать каким образом?*» - «*умело*», «*профессионально*» и так далее.

Такие ограничения необходимо накладывать для того, чтобы система могла строить правильные семантико-синтаксические структуры, что обеспечит связность и внутреннюю интерпретируемость знаний в экспертной системе.

Падежные отношения, а также соответствие между синтагматикой и линейной упорядоченностью слов в предложении можно представить в виде дерева непосредственно составляющих. Глубинная структура пропозиционального компонента любого простого предложения представляет собой конструкцию, состоящую из глагольной и именной группы, которые находятся в специальных помеченных отношениях (семантических падежах) ко всему предложению. Исходя из этого, модель семантической целостности предложения (*C*) будет выглядеть следующим образом:

$$C =< V^I, N^I >,$$

где *V*– глагольная составляющая,

N– именная составляющая,

I – семантический падеж.

Для вывода информации из базы знаний требуются также атрибутивные отношения для описания формы, цвета, размера и других характеристик объектов, а также дополнительные падежи цели, условия, времени и др., т.е. в зависимости от целей экспертной системы перечень, содержащий 6 основных падежей, будет расширен.

Добавим к структуре рассматриваемого предложения дополнительные семантические падежи и применим теорию семантических падежей для описания семантической структуры знания о событии. Допустим, системе нужно вывести сообщение «*Грузовой космический корабль пристыковался к МКС в 15:00 мск в штатном режиме. Груз доставлен на станцию*». Так будет выглядеть семантическая сеть, описывающая событие доставки груза на МКС (рис.6.30):

Рис.6.30. Пример семантической сети, описывающей событие доставки груза на МКС

В процессе поиска информации в базе знаний, основанной на семантической сети, происходит сопоставление общей сети с сетью запроса для поиска нужного фрагмента информации. При выводе на семантических сетях также применяется метод перекрестного поиска, в процессе которого сопоставляются не узлы семантической сети, а дуги – отношения между объектами.

6.5. Выводы по главе

Показано применение разработанных моделей и методов обработки англо- и русскоязычных научно-технических текстов для решения прикладных задач терминографии, учебной лексикографии, сертификации программного обеспечения, выявления тенденций развития научных направлений.

Получены следующие результаты:

1. Разработан и программно реализован лексический тренажер по дисциплине «Иностранный язык», который позволяет собирать из учебных пособий конкретный лексический материал, а на основе параллельного корпуса подбирать переводные эквиваленты, а также примеры с контекстами их употребления. Также предусмотрена возможность автоматической генерации

упражнений для овладения лексикой из учебного пособия и отработки навыков ее практического употребления в письменной речи.

2. Описан подход к выявлению тенденций развития научных направлений по результатам анализа научных публикаций на примере предметной области компримированного природного газа. Показаны способы определения направленности и закономерностей научных публикаций по популярности отдельных тематик внутри одной предметной области. Материалом исследования послужили научные публикации по компримированному природному газу, размещенные в наукометрических базах данных РИНЦ, Scopus и Web of Science и данные единой государственной информационной системы учета.

3. Разработан метод формирования нормативных профилей при сертификации программного обеспечения, в основу которого положены модели текстов стандартов, а комплексный анализ лингвистических особенностей показал, что поиск конкретных требований из текстов стандартов целесообразно проводить в два этапа: поиск стандартов или их фрагментов, наиболее подходящих запросов, а затем поиск конкретных требований в отобранных фрагментах.

4. Разработана модель представления знаний для описания конструкций летательных аппаратов, обеспечивающая информационно-структурную надежность базы знаний экспертной системы. Предложен метод разрешения лексической многозначности поискового запроса в базе знаний. Описан механизм вывода информации из базы знаний на основе падежной грамматики Ч. Филлмора, позволяющей определить семантико-синтаксическую структуру выводимого предложения. Показано, что посредством расстановки ограничений на число участников ситуации, которые может присоединять глагол, можно обеспечить эффективный вывод информации из базы знаний.

Основные результаты к разделу опубликованы в работах [9, 10, 12, 17-26, 37-45, 297, 298].

ЗАКЛЮЧЕНИЕ

В диссертации решена крупная научная проблема, имеющая важное прикладное 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.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. 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.

2. 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.

3. 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.

4. Бутенко Ю. И., Строганов Ю. В., Сапожков А. М. Метод извлечения русскоязычных многокомпонентных терминов из научно-технических текстов // Прикладная информатика. 2021. Т. 16, № 6(96). С. 21-27. DOI 10.37791/2687-0649-2021-16-6-21-27.

5. 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

6. 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

7. 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

8. 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.

9. 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.

10. 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.

11. 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.

12. 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

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

25. 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

26. 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.

27. 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.

28. 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.

29. 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.

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

35. 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.

36. 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.

37. 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.

38. 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.

39. 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.

40. 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.

41. 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. 2021. Vol. 012037.- P. 1 – 7. DOI: 10.1088/1757-899X/1061/1/012037.

42. 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.

43. 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.

44. 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.

45. 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.

46. Кружков М. Г. Информационные ресурсы контрастивных

лингвистических исследований: электронные корпуса текстов // Системы и средства информатики. 2015. Т. 25, № 2. С. 140-159.

47. Захаров В. П., Богданова С. Ю. Корпусная лингвистика: учебник. 3-е изд., перераб. — СПб.: Изд-во С.-Петерб. ун-та. 2020. 234 с.

48. Захаров В. П. Корпуса русского языка // Труды института русского языка им. В.В. Виноградова. 2015. № 6. С. 20-65.

49. Захаров В. П. Корпусно-ориентированный подход к построению тезаурусов и онтологий // Структурная и прикладная лингвистика. 2015. № 11. С. 123-141.

50. Бунтман Н. В., Зализняк А. А., Зацман И. М. и др. Информационные технологии корпусных исследований: принципы построения кросслингвистических баз данных // Информатика и ее применения. 2014. Т. 8, № 2. С. 98-110. DOI 10.14357/19922264140210.

51. Козеренко Е. Б. Лингвистические фильтры в статистических моделях машинного перевода // Информатика и ее применения. 2010. Т. 4, № 2. С. 83-92.

52. Козеренко Е. Б., Лунева Н. В., Морозова Ю. И., Ермаков И. В. Проектирование многоязычного лингвистического ресурса для систем машинного перевода и обработки знаний // Системы и средства информатики. 2009. Т. 19, № 1. С. 119-141.

53. Волченкова К. Н. Параллельный корпус как справочная база данных в работе переводчика // Проблемы и перспективы развития образования в России. 2015. № 33. С. 32-35.

54. Кокорева А. А. Методические условия обучения студентов профессионально-ориентированной лексике на основе корпуса параллельных текстов // Вестник Тамбовского университета. Серия: Гуманитарные науки. 2013. № 1(117). С. 142-146.

55. Butenko Yu. I., Kochetkova E. L. Analysis of Automation Tools for Translation // Наука, технологии и бизнес: сборник материалов II межвузовской заочной конференции аспирантов, соискателей и молодых ученых, Москва, 28–29 апреля 2020 года. – М.: МГТУ им. Н.Э. Баумана (национальный

исследовательский университет), 2020. С. 25-29.

56. Бутенко Ю. И., Авагян Н. А. Анализ качества перевода нормативных документов на основе параллельного корпуса научно-технических текстов (на примере модальных глаголов) // Языки и культуры в эпоху глобализации: особенности функционирования, перспективы развития и взаимодействия. Сборник научных статей. М.: РУДН, 2021. С. 87-96.

57. Добровольский Д. О. Корпус параллельных текстов и сопоставительная лексикология // Труды института русского языка им. В.В. Виноградова. 2015. № 6. С. 413-449.

58. Сичинава Д. В. Параллельные тексты в составе национального корпуса русского языка: новые направления развития и результаты // Труды института русского языка им. В.В. Виноградова. 2015. № 6. С. 194-235.

59. Struktura Českého národního korpusu. URL: https://wiki.korpus.cz. (accessed 10.12.2021)

60. Куратчик М. Параллельные корпуса русского и польского языков и их использование в сопоставительной лингвистике и лингводидактике // Русский язык и литература в пространстве мировой культуры: Материалы XIII Конгресса МАПРЯЛ: В 15 т., Гранада, Испания, 13–20 сентября 2015 года / Составители: Н. М. Марусенко, М. С. Шишков. Том 11. – Гранада, Испания: Международное некоммерческое партнерство преподавателей русского языка и литературы "МАПРЯЛ", 2015. С. 152-157.

61. Тао Ю. Создание и использование параллельного корпуса русского и китайского языков // Вестник МГПУ. Серия: Филология. Теория языка. Языковое образование. 2015. № 3(19). С. 76-82.

62. Чэнь С., Кукушкина О. В. О параллельных корпусах русских и китайских текстов // Вестник Московского университета. Серия 9: Филология. 2018. № 2. С. 170-197.

63. Мухин М. Ю., Ян И. Проект создания китайско-русского параллельного корпуса официально-деловых текстов с дискурсивно-структурной разметкой // Вестник Южно-Уральского государственного

университета. Серия: Лингвистика. 2016. Т. 13, № 4. С. 23-31. DOI 10.14529/ling160404.

64. Хайрова Н., Колесник А., Мамырбаев О., Мухсина К. Выровненный казахско-русский параллельный корпус, оринтированный на криминальную тематику // Вестник Алматинского университета энергетики и связи. 2020. №1(48). С. 84-92.

65. Ziemski M., Junczys-Dowmunt M., Pouliquen B. The United Nations Parallel Corpus v1.0. // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. P. 3530-3534. DOI: 10.13140/RG.2.1.1816.2801

66. Маландина А. С. Особенности русско-английского параллельного корпуса экономических текстов // Наука сегодня: вызовы, перспективы и возможности : материалы международной научно-практической конференции, Вологда, 12 декабря 2018 года / Научный центр «Диспут». – Вологда: ООО «Маркер», 2018. С. 112-114.

67. Салчак А. Я., Ондар В. С. Создание русско-тувинского параллельного подкорпуса электронного корпуса тувинского языка: первые итоги // Новые исследования Тувы. 2020. № 1. С. 6. DOI 10.25178/nit.2020.1.6.

68. Тимирбаева Г. Р. Параллельные корпуса научно-технических текстов: принципы составления и возможности применения // Казанская наука. 2019. № 12. С. 131-133.

69. Бутенко Ю. И., Киселева А. Д. Анализ современных корпусов параллельных текстов // Актуальные проблемы лингвистики и лингводидактики в неязыковом вузе: 4-я Международная научно-практическая конференция : сборник материалов конференции : в 2 т., Москва, 16 декабря 2020 года / МГТУ им. Н. Э. Баумана, Ассоциация технических университетов России и Китая, Евразийское общество прикладной лингвистики. Том 1. – Москва: МГТУ им. Н.Э. Баумана. 2021. С. 238-242.

70. Сичинава, Д. И. Параллельные корпуса восточнославянских языков: отражение исторической специфики текста и перевода / Д. И. Сичинава //

Информационные технологии и письменное наследие: Материалы IV международной научной конференции El'Manuscript–2012, Петрозаводск, 03–08 сентября 2012 года / Ответственные редакторы: Баранов Виктор Аркадьевич, Варфоломеев Алексей Геннадьевич. – Петрозаводск, 2012. – С. 247-250. – EDN PXZGKP.

71. Butenko Iu. I., Garazha V. V. BMSTU Corpus of Scientific and Technical Texts: Conceptual Framework // Applied Linguistics Research Journal. 2021. Vol 5(3). P. 76-81. – DOI: 10.14744/alrj.2021.15579.

72. Захаров В. П., Азарова И. В., Митрофанова О. А., Попов А. М., Хохлова М. В. Моделирование в корпусной лингвистике: специализированные корпусы русского языка; отв. ред. В.П.Захаров. СПб.: Изд-во С.-Петерб. ун-та, 2019. 208 с.

73. 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.

74. Butenko Iu. I., Margaryan T. D., Bolotova E. E. Scientific and Technical Text Corpus as the Basis for Aerospace Terminology Standardization // Applied Linguistics Research Journal. 2021. Vol. 5(3). P. 113-119. DOI: 10.14744/alrj.2021.72677

75. Бутенко Ю. И.,Семенова Е. Л., Сидняев Н. И. Математические аспекты в современной языковедческой теории и практике // Alma Mater (Вестник высшей школы). 2018. № 4. С. 73-78. DOI 10.20339/AM.04-18.073.

76. Полицын С. А., Полициына Е. В. Применение корпуса текстов для автоматической классификации в комплексе инструментов автоматизированного анализа текстов // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2018. № 2. С. 162-167.

77. Книжный рынок России. Состояние, тенденции и перспективы

развития. Отраслевой доклад / Под общ. ред. В. В. Григорьева. – М.: Министерство цифрового развития, связи и массовых коммуникаций Российской Федерации, 2023. 106 с.

78. Информационный портал по международной стандартизации Федерального агентва по техническому регуированию и метрологии. URL: http://iso.gost.ru// (дата доступа: 20 августа 2020).

79. Scott M. WordSmith Tools (Version 5.0). URL: http://www. lexically.net/software/index.htm (accessed: August 18, 2018).

80. Barlow M. MonoConc Pro (Version 2.2). URL: http://www.athel.com/ mono.html (accessed: August 18, 2018).

81. Anthony L. AntConc (Version 3.3.5). URL: http://www.antlab.sci.waseda.ac.jp/ (accessed: August 18, 2018).

82. Davies M. BYU corpora. URL: http://corpus.byu.edu (дата обращения: 18.08.2018).

83. Hardie A. CQPweb. URL: http://cwb.sourceforge.net/ cqpweb.php (accessed: 18.08.2018).

84. Kilgarriff A. SketchEngine. URL: http://www. sketchengine.co.uk/ (accessed: 18.08.2018).

85. Rayson P. Wmatrix. URL: http://ucrel.lancs.ac.uk/ wmatrix/ (accessed: 18.08.2018)

86. Шереметьева С. О., Бабина О. И. Платформа для концептуального аннотирования многоязычных текстов // Вестник Южно-Уральского государственного университета. Серия: Лингвистика. 2020. Т. 17, № 4. С. 53-60. DOI 10.14529/ling200409.

87. Шереметьева С. О., Бабина О. И., Зиновьева А. Ю., Неручева Е. Д. Об использовании метода кейс-стади для создания универсальных ресурсов концептуального аннотирования многоязычных текстов // Вестник Южно-Уральского государственного университета. Серия: Лингвистика. 2020. Т. 17, № 4. С. 46-52. DOI 10.14529/ling200408.

88. Сулейманов Д. Ш., Мухамедшин Д. Р. Система корпус-менеджер:

архитектура и модели корпусных данных // Программные продукты и системы. 2018. № 4. С. 653-658.

89. Барахнин В. Б., Кожемякина О. Ю., Мухамедиев Р. И. и др. Проектирование структуры программной системы обработки корпусов текстовых документов // Бизнес-информатика. 2019. Т. 13, № 4. С. 60-72. DOI 10.17323/1998-0663.2019.4.60.72.

90. Белозеров А. А., Вахлаков Д. В., Мельников С. Ю. и др. Технологические аспекты построения системы сбора и предобработки корпусов новостных текстов для создания моделей языка // Известия ЮФУ. Технические науки. 2016. № 12(185). С. 29-42. DOI 10.18522/2311-3103-2016-12-2942.

91. Носов А. В. Лингвистическая разметка корпусов переводных текстов // Индустрия перевода. 2017. Т. 1. С. 68-72.

92. Потемкин С. Б. Проблемы разработки параллельного корпуса переводов русской классики // Армия и общество. 2012. №2(30). С.138-146.

93. Козеренко Е. Б. Стратегии выравнивания параллельных текстов: семантические аспекты // Информатика и ее применения. 2013. Т. 7, № 1. С. 82-89.

94. Морозова Ю. И., Козеренко Е. Б., Шарнин М. М. Методика извлечения пословных переводных соответствий из параллельных текстов с применением моделей дистрибутивной семантики // Системы и средства информатики. 2014. Т. 24, № 2. С. 131-142. DOI 10.14357/08696527140209.

95. Лесников С. В. Виды разметок текстовых корпусов русского языка // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2019. № 9. С. 27-30. DOI 10.36535/0548-0027-2019-09-4.

96. Захаров В. П., Азарова И. В. Параметризация специальных корпусов текстов // Структурная и прикладная лингвистика. 2012. № 9. С. 176-184.

97. Steinberger R., Ebrahim M., Poulis A., Carrasco-Benitez M., Schlüter P., Przybyszewski M., Gilbro S. An overview of the European Union’s highly multilingual parallel corpora // Language resources and evaluation. 2014. V.48. pp.679-707.

98. Aulamo M., Sulubacak U., Virpioja S., Tiedemann J. OpusTools and Parallel

Corpus Diagnostics // Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020. pp. 3782–3789.

99. Scherrer Y. TaPaCo: a corpus of sentential paraphrases for 73 languages // Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association (ELRA). Marseille, 11–16 May 2020. pp. 6868–6873.

100. Gezmu A.M., Seyoum B.E., Gasser M. Nürnberger A. Contemporary Amharic Corpus: Automatically Morpho-Syntactically Tagged Amharic Corpus // Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing. 2018. pp. 65-70.

101. Costa-Jussa M. R., Fonollosa J. A. , Marino J. B. et al. A large Spanish-Catalan parallel corpus release for machine translation // Computing and Informatics. 2014. V. 33 (4). pp. 907-920.

102. Toral A., Rubino R., Ramírez-Sánchez G. Re-assessing the Impact of SMT Techniques with Human Evaluation: a Case Study on English-Croatian // Proceedings of the 19th Annual Conference of the European Association for Machine Translation. 2016. pp. 368-375.

103. Vastl M., Zeman D., Rosa R. Predicting Typological Features in WALS using Language Embeddings and Conditional Probabilities: ÚFAL Submission to the SIGTYP 2020 Shared Task // Proceedings of the Second Workshop on Computational Research in Linguistic Typology, 2020. pp. 29–35,

104. Bojar O., Dušek O., Kocmi T., Libovický J., Novák M., Popel M., Sudarikov R., Variš D. Czeng 1.6: enlarged czech-english parallel corpus with processing tools dockered // Text, Speech, and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings 19. 2016. pp. 231-238.

105. Galuščáková P., Bojar O. Czech-Slovak Parallel Corpora for MT between Closely Related Languages // Natural Language Processing, Multilinguality, p.65.

106. Bojar O., Dušek O., Kocmi T., Libovický J., Novák M., Popel M., Sudarikov R., Variš, D. Czeng 1.6: enlarged czech-english parallel corpus with

processing tools dockered // Text, Speech, and Dialogue: 19th International Conference Proceedings, TSD 2016, Brno, Czech Republic, September 12-16, 2016, pp. 231-238.

107. Štromajerová A., Baisa V., Blahuš M. Between comparable and parallel: English-czech corpus from Wikipedia // The 10th Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2016. Karlova Studánka, Czech Republic, December 2–4, 2016. p.3-8.

108. Pilevar M. T., Faili H., Pilevar A. H. Tep: Tehran english-persian parallel corpus // Computational Linguistics and Intelligent Text Processing: 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part II 12. pp. 68-79.

109. Ngo Q.H., Winiwarter W. November. Building an English-Vietnamese bilingual corpus for machine translation // 2012 International Conference on Asian Language Processing. 2012. pp. 157-160.

110. Ljubešić N., Esplà-Gomis M., Ortiz Rojas S., Klubička F., Toral A. Finnish-English parallel corpus fienWaC 1.0, Slovenian language resource repository. URL: http://hdl.handle.net/11356/1060. (accessed 10.09.2018).

111. Bojar O., Diatka V., Rychlý P., Stranák P., Suchomel V., Tamchyna A., Zeman, D. HindEnCorp-Hindi-English and Hindi-only Corpus for Machine Translation // Conference: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014). Reykjavik, Iceland. May 26-31, 2014. pp. 3550-3555.

112. Barkarson S., Steingrímsson S. Compiling and filtering ParIce: an English-icelandic parallel corpus // Proceedings of the 22nd Nordic Conference on Computational Linguistics. 2019. pp. 140-145.

113. Andelkovic J. Aligned Parallel Corpus for the Domain of Management: Preparation and Potential Applications // Infotheca. 2018. Vol. 18. #2. pp. 7-28.

114. Duwal S., Bal B.K. December. Efforts in the development of an aug-mented english–nepali parallel corpus // Proceedings of the 1st International Conference on Language Technologies for All. Paris, Franc. 2019. pp. 375-378.

115. Hareide L., Hofland K. Compiling a Norwegian-Spanish parallel corpus:

Methods and challenges // Quantitative methods in corpus-based translation studies: A practical guide to descriptive translation research. 2012. pp. 75-114.

116. Toral A., Esplà-Gomis M., Klubička F., Ljubešić N., Papavassiliou V., Prokopidis P., Rubino R., Way A. Tourism English-Croatian Parallel Corpus 2.0. Slovenian language resource repository. URL: http://hdl.handle.net/11356/1049. (accessed 05.08.2020).

117. Schäfer U., Read J., Oepen S. Towards an ACL anthology corpus with logical document structure // An overview of the ACL 2012 Contributed Task. In Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries. 2012. pp. 88-97.

118. Kunstmann P. Corpus of Old French literary texts // Corpus-Based Perspectives in Linguistics. 2008. pp. 85-90.

119. Usonienė A., Grigaliūnienė J., Ryvitytė B., Būtėnas L., Jasionytė E. Lietuvių mokslo kalbos tekstynas // Baltistica. 2011. Vol. 43(1), pp.101-114.

120. Erjavec T., Fišer D., Ljubešić N. The KAS corpus of Slovenian academic writing // Language Resources and Evaluation. 2021. V.55. pp.551-583.

121. Hennoste T., Koit M., Roosmaa T., Saluveer M. Structure and usage of the Tartu University corpus of written Estonian // International Journal of Corpus Linguistics. 1998. V. 3(2). pp.279-304.

122. Striletchi C., Chitez M., Csürös K. Building Roger: Technical Challenges While Developing a Bilingual Corpus Management and Query Platform // Proceedings of the 17th International Conference on Software Technologies (ICSOFT 2022). Lisbon, Portugal, July 11-13, 2022. pp. 390-398.

123. Sugimoto G. Examining Web User Flows and Behaviours in CLARIN Ecosystem // CLARIN Annual Conference. 2017. pp. 46-60.

124. Kim J.D., Ohta T., Tateisi Y., Tsujii J.I. GENIA corpus – a semantically annotated corpus for bio-textmining // Bioinformatics. 2003. V. 19(1). pp.i180-i182.

125. Nikiforos M.N., Voutos Y., Drougani A., Mylonas P., Kermanidis K.L.The modern Greek language on the social web: a survey of data sets and mining applications // Data. 2021. V. 6(5). p.52.

126. Widdows D., Dorow B., Chan Ch. Using Parallel Corpora to enrich Multilingual Lexical Resources // Third International Conference on Language Resources and Evaluation, ELRA, Las Palmas, May 2002, Pages 240-245.

127. De Jong F.M.G., Maegaard B., De Smedt K., Fišer D., Van Uytvanck D. 2018. CLARIN: Towards FAIR and responsible data science using language resources // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018. pp. 3259-3264.

128. Römer U., Wulff, S. Applying corpus methods to written academic texts: Explorations of MICUSP // Journal of Writing research. 2010. Vol. 2(2), pp. 99-127.

129. Cavalla C., Loiseau, M. Scientext comme corpus pour l'enseignement // L'écrit scientifique: Du Lexique au discours. Autour de scientext. 2013. pp.163-182.

130. Hyland K. Corpora and academic discourse // Corpus applications in applied linguistics. 2012. pp.30-46.

131. Горбань О. А., Косова М. В., Шептухина Е. М. Структурная разметка деловых документов в диахроническом лингвистическом корпусе: проблемы и решения // Вестник Волгоградского государственного университета. Серия 2: Языкознание. 2021. Т. 20, № 4. С. 5-18. DOI 10.15688/jvolsu2.2021.4.1.

132. Ринчинов О. С. Структурная разметка бурятских старописьменных сочинений для диахронического корпуса бурятского языка // Культура Центральной Азии: письменные источники. 2019. № 12. С. 106-117. DOI 10.30792/2304-1838-2019-106-117.

133. Горбунов А. Ю., Долбунова Л. А. Структура и языковые особенности англоязычных текстов технической документации // Огарёв-Online. 2015. № 12(53). С. 1.

134. Новиков А. И. Структура содержания текста и возможности ее формализации (на материале научно-технических текстов) : специальность 10.02.19 "Теория языка" : диссертация на соискание ученой степени доктора филологических наук / Новиков Анатолий Иванович. М., 1983. 355 с.

135. Крупнов В. В. В творческой лаборатории переводчика. – М.: Международные отношения, 1976. 161 с.

136. Бизюкова Н. Ю., Тарасова О. А., Рудик А. В. и др. Автоматическое распознавание названий химических соединений в текстах научных публикаций // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2020. № 11. С. 36-46. DOI 10.36535/0548-0027-2020-11-5.

137. Зацман И. М. Логико-семантические модели полнотекстовых научных документов // Научно-техническая информация. Серия 2: Информационные процессы и системы. 1999. № 5. С. 13-22.

138. Гусенков А. М. Интеллектуальный поиск сложных объектов в массивах больших данных // Электронные библиотеки. 2016. Т. 19, № 1. С. 40-76.

139. Сухомлинова М. А. Особенности композиционной структуры текста англоязычной академической лекции // Вестник Северного (Арктического) федерального университета. Серия: Гуманитарные и социальные науки. 2021. Т. 21, № 4. С. 83-92. DOI 10.37482/2687-1505-V119.

140. Акаева Э. В. Композиция англоязычной научной медицинской статьи // Colloquium-Journal. 2019. № 10-5(34). С. 59-60.

141. Шамара И. Ф. Аннотация научной медицинской статьи: от анализа дискурсивной структуры к созданию собственного текста на английском языке // Теория языка и межкультурная коммуникация. 2019. № 2(33). С. 185-193.

142. Гришечкина Г. Ю. Композиционная структура французских научно-популярных лингвистических текстов // Вопросы когнитивной лингвистики. 2012. № 4(33). С. 103-107.

143. Мохов А. С. Метод классификации библиографической информации на основе комбинированных профилей классов с учетом структуры документов: специальность 05.13.01 «Системный анализ, управление и обработка информации (по отраслям)» : диссертация на соискание ученой степени кандидата технических наук / Мохов Андрей Сергеевич, 2017. 180 с.

144. Груздо И. В. Модель рубрицированного объекта в задачах машинного анализа текстов, учитывающая значимость структурных частей // Системы обработки информации. 2013. №2. С.125-131.

145. Бутенко Ю. И. Модель учебно-научного текста для разметки корпуса научно-технических текстов // Экономика. Информатика. 2021. Т. 48, № 1. С. 123-129. DOI 10.52575/2687-0932-2021-48-1-123-129.

146. Астраханцев Н. А., Федоренко Д. Г., Турдаков Д. Ю. Методы автоматического извлечения терминов из коллекции тектов предметной области // Программирование. 2015. № 6. С. 33-52.

147. Drouin P., Morel J.B., L'Homme M.C. Automatic term extraction from newspaper corpora: Making the most of specificity and common features // Proceedings of the 6th International Workshop on Computational Terminology. 2020. pp. 1-7.

148. Korkontzelos I., Klapaftis I.P., Manandhar S. Reviewing and evaluating automatic term recognition techniques // Advances in Natural Language Processing: 6th International Conference, GoTAL 2008 Gothenburg, Sweden, August 25-27, 2008 Proceedings. 2008. pp. 248-259.

149. Nugumanova A., Akhmed-Zaki D., Mansurova M., Baiburin Y., Maulit A. NMF-based approach to automatic term extraction // Expert Systems with Applications. 2022. V.199. p.117179.

150. Кузнецов И. О. Автоматическое извлечение двусловных терминовпо тематике "Нанотехнологии в медицине"на основе корпусных данных// Научно-техническая информация. Серия 2: Информационные процессы и системы. 2013. № 5. С. 25-33.

151. Simon N. I., Kešelj V. Automatic term extraction in technical domain using part-of-speech and common-word features // Proceedings of the ACM Symposium on Document Engineering 2018, DocEng 2018: 18, Halifax, NS, 28–31 August 2018. – Halifax, NS, 2018. P. a51. DOI 10.1145/3209280.3229100.

152. Наместников А. М., Филиппов А. А., Шигабутдинов И. М. Подход к извлечению многословных терминов из текстов на естественном языке с применением синтаксических шаблонов // Автоматизация процессов управления. 2021. № 3(65). С. 87-95. DOI 10.35752/1991-2927-2021-3-65-87-95.

153. Loukachevitch N., Dobrov B. Ontological Resources for Representing

Security Domain in Information-Analytical System // Открытые семантические технологии проектирования интеллектуальных систем. 2018. No. 8. С. 185-191.

154. Морев Н. А. К проблеме лингвистического анализа терминологии в области нанотехнологий (о необходимости разработки исследовательского корпуса терминологических единиц) // Вестник Московского государственного лингвистического университета. 2012. № 646. С. 115-124.

155. Кочеткова Н. А. Метод извлечения технических терминов с использованием усовершенствованной меры странности // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2015. № 5. С. 25-32.

156. Клышинский Э. С., Кочеткова Н. А., Карпик О. В. Метод выделения коллокаций с использованием степенного показателя в распределении Ципфа // Новые информационные технологии в автоматизированных системах. 2018. № 21. С. 220-225.

157. Бессмертный И. А., Нугуманова А. Б., Мансурова М. Е., Байбурин Е. М. Метод контрастного извлечения редких терминов из текстов на естественном языке // Научно-технический вестник информационных технологий, механики и оптики. 2017. Т. 17, № 1. С. 81-91. DOI 10.17586/2226-1494-2017-17-1-81-91.

158. Большакова Е. И., Лукашевич Н. В., Нокель М. А. Извлечение однословных терминов из текстовых коллекций на основе методов машинного обучения // Информационные технологии. 2013. № 7. С. 31-37.

159. Астраханцев Н. А. Автоматическое извлечение терминов из коллекции текстов предметной области с помощью Википедии // Труды Института системного программирования РАН. 2014. Т. 26, № 4. С. 7-20. DOI 10.15514/ISPRAS-2014-26(4)-1.

160. Гринева М., Гринев М. Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов / М. Гринева, М. Гринев // Труды Института системного программирования РАН. 2009. Т. 16. С. 155-165.

161. Кононенко И. С., Ахмадеева И. Р., Сидорова Е. А., Шестаков В. К.

Проблемы извлечения терминологического ядра предметной области из электронных энциклопедических словарей // Системная информатика. 2018. № 13. С. 49-76. DOI 10.31144/si.2307-6410.2018.n13.p49-76.

162. Лукашевич Н. В. Модели и методы автоматической обработки неструктурированной информации на основе базы знаний онтологического типа : специальность 05.25.05 "Информационные системы и процессы" : диссертация на соискание ученой степени доктора технических наук / Лукашевич Наталья Валентиновна. М., 2014. 312 с.

163. Захаров В. П., Хохлова М. В. Автоматическое выявление терминологических словосочетаний // Структурная и прикладная лингвистика. 2014. № 10. С. 182-200.

164. Кочеткова Н. А., Ермаков П. Д. Метод извлечения однословных терминов на основе статистического распределения слов внутри контекста // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2017. № 1. С. 23-28.

165. Terryn A. T., Hoste V., Lefever E. In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora // Language Resources and Evaluation. 2020. V.54.2. pp. 385-418.

166. Петров А. С., Шульга Т. Э. Математическая модель русскоязычного текстового документа для решения задачи автоматического извлечения терминов из текста // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2017. № 3. С. 195-203.

167. Клышинский Э. С., Кочеткова Н. А. Метод извлечения технических терминов с использованием меры странности // Новые информационные технологии в автоматизированных системах. 2014. № 17. С. 365-370.

168. Nugumanova A., Akhmed-Zaki D., Mansurova M., Baiburin Y., Maulit A. NMF-based approach to automatic term extraction // Expert Systems with Applications. 2022. V.199. p.117179.

169. Lopes L., Fernandes P., Vieira R. Estimating term domain relevance through term frequency, disjoint corpora frequency-tf-dcf // Knowledge-Based

Systems. 2016. Vol. 97. pp.237-249. doi.org/10.1016/j.knosys.2015.12.015

170. De Handschutter P., Gillis N., Siebert, X., 2021. A survey on deep matrix factorizations // Computer Science Review. 2021. Vol. 42. p.100423.

171. Ефремова Наталья Эрнестовна. Методы и программные средства извлечения терминологической информации из научно-технических текстов : диссертация ... кандидата физико-математических наук: 05.13.11 / Ефремова Наталья Эрнестовна. – М., 2013. – 135 с.

172. Sterckx L., Demeester T., Deleu J., Develder C. Topical word importance for fast keyphrase extraction // Proceedings of the 24th International Conference on World Wide Web. 2015. pp. 121-122. DOI.org/10.1145/2740908.2742730

173. Teneva N., Cheng W. Salience rank: Efficient keyphrase extraction with topic modeling // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. Vol.2. pp. 530-535.

174. Süzek T.O. Using latent semantic analysis for automated keyword extraction from large document corpora // Turkish Journal of Electrical Engineering & Computer Sciences. 2017. V.25(3). pp. 1784–1794. https://doi.org/10.3906/ELK-1511-203

175. Abuzayed A., Al-Khalifa H. BERT for arabic topic modeling: an experimental study on BERTopic technique // Procedia Computer Science. 2021. V.189. pp. 191–194.

176. Cram D., Daille B. Terminology extraction with term variant detection // Proceedings of ACL-2016 system demonstrations. 2016. pp. 13-18.

177. Гринева М., Гринев М. Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов // Труды Института системного программирования РАН. 2009. Т. 16. С. 155-165.

178. Lang C., Wachowiak L., Heinisch B., Gromann, D., 2021, August. Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains // Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021. pp. 3607-3620.

179. Lossio-Ventura J.A., Jonquet C., Roche M., Teisseire M. Biomedical term

extraction: overview and a new methodology // Information Retrieval Journal. 2016. Vol.19. pp.59-99. DOI.org/10.1007/s10791-015-9262-2

180. Козловская Н. В., Янурик С. ИИ-композиты как объект неологии и неографии XXI века // Филологические науки. Научные доклады высшей школы. 2021. № 2. С. 23-30. DOI 10.20339/PhS.2-21.023.

181. Шмелева О. Ю. Терминологические процессы в синхронии и диахронии (на материале английского языка). СПб.: Изд-во СПбГУЭФ, 2010. 120 с.

182. Цисун Е., Шелов С. Д. О классификации номенов и номенклатурных наименований (на материале наименований товаров) // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2015. № 6. С. 37-44.

183. Лейчик В.М. Терминоведение: Предмет, методы, структура. Изд. 4-е. – М.: Книжный дом ≪ЛИБРОКОМ≫, 2009. 256 с.

184. Dugan L., Ippolito D., Kirubarajan A., Shi S., Callison-Burch C. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text // Proceedings of the AAAI Conference on Artificial Intelligence. 2023. Vol. 37, No. 11, pp. 12763-12771.

185. Jawahar G., Mageed M.A., Laks Lakshmanan V.S. Automatic Detection of Machine Generated Text: A Critical Survey // Proceedings of the 28th International Conference on Computational Linguistics. 2020. pp. 2296-2309.

186. Грицай Г. М., Грабовой А. В., Кильдяков А. С., Чехович Ю. В. Поиск искусственно сгенерированных текстовых фрагментов в научных документах // Доклады Российской академии наук. Математика, информатика, процессы управления. 2023. Т. 514, № 2. С. 308-317. DOI 10.31857/S2686954323601677.

187. Чехович, Ю. В. Модели генеративного искусственного интеллекта с полным их разоблачением / Ю. В. Чехович, А. Грабовой, Г. Грицай // Университетская книга. – 2024. – № 5. – С. 58-65.

188. Николаев В. В., Рахконен М. Е. Применение различных инструментов и использование чат-бота "chatgpt" при написании научных работ, проверяемых

в программе «Антиплагиат» // Профессиональное юридическое образование и наука. 2023. № 1(9). С. 78-81.

189. Wei F., Nguyen U.T. Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings // 2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). 2019. pp. 101–109.

190. Sneha K., Ferrara E. Deep neural networks for bot detection // Information Sciences. 2018. V.467. pp. 312-322.

191. Dukić D., Keča D., Stipić D. Are you human? detecting bots on twitter using bert // 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). 2020. pp. 631–636.

192. Gromov V., Dang Q.N. Spot the Bot: Distinguishing Human-Written and Bot-Generated Texts Using Clustering and Information Theory Techniques // Pattern Recognition and Machine Intelligence. PReMI 2023. Lecture Notes in Computer Science. V. 14301. 2023. https://doi.org/10.1007/978-3-031-45170-6_3

193. Черкасова М. Н., Тактарова А. В. Особенности определения сгенерированного искусственным интеллектом академического текста: прагмалингвистический анализ // Вестник научных исследований. 2024. № 3 (45). С. 30–40.

194. Бусел Т. В. Разработка автоматизированного метода порождения деловых документов на основе лингвистических правил // Актуальные проблемы современной прикладной лингвистики: сб. науч. ст., посвящ. 80-летию д-ра филол. наук, проф., акад. Междунар. акад. информатизации А. В. Зубова. – Минск: МГЛУ. 2017. С. 126–133.

195. Селиванова Е. А. Лингвистическая энциклопедия. Полтава, Довкиля - К, 2010.

196. Сухомлинова, М. А. Особенности композиционно-смысловой организации академических текстов (на материале английского языка) // Научная мысль Кавказа. 2018. № 4(96). С. 102-109. DOI 10.18522/2072-0181-2018-96-4-102-109.

197. Бутенко Ю. И. Модель текста стандарта при информационном поиске в коллекции документов нормативной базы // Вестник компьютерных и информационных технологий. 2020. Т. 17, № 11(197). С. 23-32. DOI 10.14489/vkit.2020.11.pp.023-032.

198. Попова Т. Г. Структура испанской научно-технической статьи как первичного жанра научного дискурса // Вестник Российского университета дружбы народов. Серия: Русский и иностранные языки и методика их преподавания. 2004. № 1. С. 108-115.

199. Романов Д. А. Кратко о структуре экспериментальной научной статьи на английском языке // Вестник Казанского технологического университета. 2014. Т. 17, № 6. С. 325-327.

200. Раицкая Л. К. Структура научной статьи по политологии и международным отношениям в контексте качества научной информации // Полис. Политические исследования. 2019. № 1. С. 167-181. DOI 10.17976/jpps/2019.01.12.

201. Попов, Н. Г. Введение к научной статье на английском языке: структура и композиция // Высшее образование в России. 2015. № 6. С. 52-58.

202. 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.

203. Иванов В. П. Как написать научную статью (структура материала и организация работы) // Вестник Полоцкого государственного университета. Серия В. Промышленность. Прикладные науки. 2016. № 3. С. 195.

204. Тюрина Л. Г. Особенности текста учебной книги // Известия высших учебных заведений. Проблемы полиграфии и издательского дела. 2007. № 3. С. 70-73.

205. Тюрина Л. Г. Состав и структура учебной книги как педагогической системы // Известия высших учебных заведений. Проблемы полиграфии и издательского дела. 2005. № 4. С. 78-88.

206. Рыбакова Г. Р. О категории «учебный текст» в научной литературе // Научное обозрение. Серия 2: Гуманитарные науки. 2011. № 6. С. 64-73.

207. Лыков М. Н. Оглавление как структурный элемент вузовского учебника (на примере учебника по истории отечества для высшей школы) // Альманах современной науки и образования. 2008. № 10-1. С. 102-105.

208. Лупачев В.Г., Павлюк С.К. Методические основы и принципы разработки учебной литературы: методическое пособие для слушателей курсов повышения квалификации и переподготовки кадров; под ред. В.А. Сидорова. Минск. БНТУ: 2011. 63 с.

209. ПНСТ 118-2016 Атомные станции. Контроль и управление, важные для безопасности. Использование программируемых интегральных схем для применения в системах, выполняющих функции категории А. М.: Стандартинформ, 2016. – 69 с.

210. НП 306.5.02/3.035-2000. Требования по ядерной и радиационной безопасности к информационным и управляющим системам, важным для безопасности атомных станций. М.: Стандартинформ, 2000. – 59 с.

211. МЭК 60880. Атомные электростанции. Системы контроля и управления, важные для безопасности. Программное обеспечение компьютерных систем, выполняющих функции категории А. М.: Стандартинформ, 2011. – 90 с.

212. ЧП 306.5.02/3.035-2000. Требования по ядерной и радиационной безопасности к информационным и управляющим системам, важным для безопасности атомных станций. М.: Стандартинформ, 2000. – 88 с.

213. NUREG/CR-6303. Method for Performing Diversity and Defense-in-Depth Analyses of Reactor Protection Systems, U.S. Nuclear Regulatory Commission, December 1994.

214. NS-G-1.1. Software for computer-based systems important to safety in nuclear power plants. - IAEA Safety standards series. Safety Guide. Ed. International Atomic Energy Agency, Vienna, 2000

215. Гальперин И.Р. Текст как объект лингвистического исследования. М.,

Наука, 1981. 139 с.

216. 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.

217. Ястребенецкий М. А. Управление старением критических систем // Радиоэлектронные и компьютерные системы. 2008. №6. С. 114-121.

218. Бадмаева Л. Д. Бурятско-русские параллельные тексты: проблемы асимметрии // Томский журнал лингвистических и антропологических исследований. 2018. № 3(21). С. 19-30. DOI 10.23951/2307-6119-2018-3-19-30.

219. Циткина Ф. А. Терминология и перевод. – Львов: Высшая школа, 1988. – 157 с.

220. Сорокина Э. А. Проблемы анализа неоднословных терминов // Вестник Московского университета. Серия 22: Теория перевода. 2018. № 4. С. 150-158.

221. 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.

222. Терминология сварки металлов / под ред. С. А. Чаплыгина, Д. С. Лотте. – Москва: Изд-во Акад. Наук СССР, 1937. -31 с.

223. Лотте Д. С. Основы построения научно-технической терминологии / Д.С. Лотте. – М.: Изд-во АН СССР, 1961. – 158 с.

224. Золотых В.Т. Англо-русский словарь по сварочному производству / под ред. А.А. Ерохина. – изд. 2-е, перераб. и доп. – М.: Сов. Энциклопедия, 1967. – 376

225. Кулик Т.А. Словарь-справочник по сварке / под ред. К.К. Хренова. – Киев: Наукова думка, 1974. – 196 с.

226. ГОСТ 2601-84. Сварка металлов. Термины и определения основных

понятий. – Взамен ГОСТ 2601-74; введ. 01.07.85. – М.: Госстандарт СССР: Изд-во стандартов, 1984. – 51 с.

227. Шуфан С., Шелов С. Д. Номенклатурные наименования как элемент китайской научной лексики (на материале языкознания и литературоведения) // Вестник Санкт-Петербургского университета. Востоковедение и африканистика. 2014. № 3. С. 5-16.

228. Лейчик В. М. О языковом субстрате термина // Вопросы языкознания. – 1986. – № 5. – С. 87-97.

229. Раренко, М. Б. Учение об актуальном членении предложения и его значение для развития теории и практики перевода на современном этапе // Вестник Московского государственного областного университета. Серия: Лингвистика. 2022. № 3-2. С. 22-33. DOI 10.18384/2310-712X-2022-3-2-22-33.

230. Крылова, О. А. Структурные схемы и актуальное членение предложения // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2012. № 2. С. 6-14.

231. Кутилин, Д. С., Быкадорова Е. С., Чусовлянова С. В. Функции переводческих трансформаций при переводе текстов научно-технической литературы // Русский лингвистический бюллетень. 2022. № 8(36). С.22. DOI 10.18454/RULB.2022.36.16.

232. Weller M., Gojun A., Heid U., Daille B., Harastani, R. Simple methods for dealing with term variation and term alignment // 9th International Conference on Terminology and Artificial Intelligence (TIA 2011). 2011. pp. 87-93.

233. Repar A., Martinc M., Ulcar M., Pollak S. Word-embedding based bilingual terminology alignment // Electronic lexicography in the 21st century (eLex 2021). 2021. p.408-417.

234. Chen Y., Liu Y., Chen G. et al. Accurate word alignment induction from neural machine translation // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing and the 10th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2020. P. 566-576.

235. Qader W.A., Ameen, M.M., Ahmed B.I. An overview of bag of words;

importance, implementation, applications, and challenges // 2019 International Engineering Conference (IEC). 2019. pp. 200-204.

236. Бутенко Ю. И. Использование базы данных моделей структурных переводческих трансформаций для извлечения многокомпонентных терминологических единиц // Системы и средства информатики. 2023. Т. 33, № 1. С. 35-44. DOI 10.14357/08696527230104.

237. 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.

238. 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.

239. Sabet M. J., Dufter P., Yvon F., Schütze H. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings // EMNLP 2020. 2020. pp. 1627-1643.

240. Рецкер, Я.И. Теория перевода и переводческая практика. Очерки лингвистической теории перевода. - 4-е издание. – М.: Валент, 2010. 244 с.

241. Комиссаров, В.Н. Теория перевода. – М.: Высш. шк., 1990. 254 с.

242. Лагутин М. Б. Наглядная математическая статистика :учеб.пособие для вузов / Лагутин М. Б. - 7-е изд. - М. : БИНОМ. Лаборатория знаний, 2019. 472 с.

243. Сидняев Н.И. Теория вероятностей и математическая статистика: учебное пособие. М.: Юрайт, 2011. 310 с.

244. Пугачёв В. С.Теория вероятностей и математическая статистика : учебник / Пугачёв В. С. - М. : Транспортная компания, 2019. - 496 с.

245. Сидняев Н.И. Статистический анализ и теория планирования эксперимента: учебное пособие/Н.И. Сидняев. –Москва: Издательство МГТУ им. Н.Э. Баумана, 2017. 195 с.

246. Сидняев Н.И. ,Вилисова Н.Т. Введение в теорию планирования

эксперимента: учебное пособие. М.: Изд-во МГТУ им. Н.Э.Баумна, 2011 г. 399 с.

247. 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.

248. Soto A., Olivas J. A., Prieto M. E. Fuzzy approach of synonymy and polysemy for information retrieval // Granular computing: At the junction of rough sets and fuzzy sets. – Springer, Berlin, Heidelberg, 2008. pp. 179-198.

249. Марчук Ю. Н. Лексические проблемы новых информационных технологий // Современный ученый. 2017. № 5. С. 56-62.

250. Люгер Дж. Искусственный интеллект: Стратегии и методы решения слож. проблем [Пер. с англ. Н. И. Галагана и др.]. - 4. изд. - М.: Вильямс, 2003. 863 с.

251. Палкова А. В. Основные понятия электронной лексикографии // Вестник Тверского государственного университета. Серия «Филология». 2015. № 4. С. 88–93.

252. Мезит А. Э. Концепция «Словаря специальной лексики русской гидроэнергетической отрасли» // Вопросы лексикографии. 2019. №16. С. 138–152. DOI: 10.17223/22274200/16/8.

253. Орлова Е. В. Электронный учебный словарь коллокаций для специалистов МЧС России как средство развития учебной иноязычной лексической компетенциити // Пожарная и аварийная безопасность. 2019. №3(14). С. 32–35.

254. Ятаева Е. В. Электронный учебный словарь как средство развития учебной иноязычно-лексической компетенции // Вестник Челябинского государственного педагогического университета. 2016. №10. С. 135–140.

255. Калугина Л. В., Лосева О. М. Английский язык в эпоху цифровых технологий. Книга 1 = English in the Digital Age: мультимедийное учебное пособие. М.: Изд-во МГТУ им. Н. Э. Баумана. 2018. 108 с.

256.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), 189–200. DOI 10.52575/2687-0932-2024-51-1-189-200

257. Makri A. Pakistan and Egypt had highest rises in research output in 2018 // Nature. 2018. p. 21.

258. Сушенцова Н. В., Чекалина Т. А. Научные электронные библиотеки открытого доступа // Образование. Карьера. Общество. 2013. № 4-1(40). С. 31-34.

259. Birkle C., Pendlebury D.A., Schnell J., Adams, J. Web of Science as a data source for research on scientific and scholarly activity // Quantitative science studies. 2020. 1(1). pp.363-376.

260. Mongeon P., Paul-Hus A. The journal coverage of Web of Science and Scopus: a comparative analysis //Scientometrics. 2016. V. 106, №. 1. pp. 213-228.

261. Савельева Ю.В., Хоперсков А.В. Научные журналы и эффективность научной работы: поисковые системы и базы данных // Управление большими системами: сборник трудов. 2013. № 44. С. 381-407.

262. Van Eck N. J., Waltman L. VOSviewer manual // Leiden: Univeristeit Leiden. 2013. V. 1, № 1. pp. 1-53.

263. Мельников А.К., Ронжин А.Ф. Обобщенный статистический метод анализа текстов, основанный на расчете распределении вероятностей значений статистик // Информатика и ее применения. 2016. №4(10). С. 89-95. DOI: 10.14357/19922264160409

264. Волков А.В. Особенности компьютерной обработки научного текста // Управление инновациями: теория, методология, практика. 2013. № 5. С. 144-151.

265. Яцко В. А. Алгоритмы и программы автоматической обработки текста // Вестник Иркутского государственного лингвистического университета. 2012. № 1(17). С. 150-160.

266. Tollefson J. World’s carbon emissions set to spike by 2% in 2017 // Nature News. 2017. V.551, № 7680. p. 283.

267. Randles B.M., Pasquetto I.V., Golshan M.S., Borgman, C.L. Using the Jupyter notebook as a tool for open science: An empirical study // 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL). 2017. pp. 1-2.

268. Odarushchenko O., Strjuk O., Bulba Y., Leontiiev K., Ivasyuk A. Kharchenko V. Fault insertion software and hardware testing for safety PLC-based system SIL certification // 9th International Conference on Dependable Systems, Services and Technologies (DESSERT). 2018. pp. 202-206. DOI**:** 10.1109/DESSERT.2018.8409128

269. Тарасюк О. М. Методы и инструментальные средства метрико-вероятностной оценки качества программного обеспечения информационно-управляющих систем критического применения: дис. ... канд. тех. наук: 05.13.06 / Тарасюк Ольга Михайловна. Харьков, 2004. 201 с.

270. Vilkomir S.A., Khasrchenko V.S. The Formalized Models of an Evaluation of a Verification Process of Critical Software // Proceedins PSAM5, (November 27 – December 1, 2000). Osaka, Japan. V.4. p. 2383-2388.

271. Babeshko Eu., Yasko A., Kharchenko V. FMEDA-based NPP I&C systems safety assessment: toward to minimization of experts' decisions uncertainty // Proceedings of the 24th International Conference on Nuclear Engineering (ICONE24), Volume 5, June 26-30, 2016, Charlotte, North Carolina, USA, Paper ID: ICONE24-60377.

272. Kharchenko V., Gordieiev O., Fedoseeva A. Profiling of Software Requirements for the Pharmaceutical Enterprise Manufacturing Execution System. // Applications of Computational Intelligence in Biomedical Technology. Springer, Cham, 2016, pp. 67-92.

273. Андрашов А. А. Таксономические модели профилирования требований информационно-управляющих систем критического применения // Радиоэлектронные и компьютерные системы. 2010. №7 (48). С. 104–108.

274. Volochiy B., Mulyak O., Ozirkovskyi L., Kharchenko V. Automation of quantitative requirements determination to software reliability of safety critical NPP I&C systems // 2016 Second International Symposium on Stochastic Models in

Reliability Engineering, Life Science and Operations Management (SMRLO), 2016. pp. 337-346.

275. Loukachevitch N., Dobrov B. RuThes Thesaurus for Natural Language Processing // The Palgrave Handbook of Digital Russia Studies. 2021. pp. 319-334.

276. Manning C. Understanding human language: Can NLP and deep learning help? // Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1-1. 2016.

277. Tata S., Potti N., Wendt J. B., Costa L. B., Najork M., Gunel B. Glean: structured extractions from templatic documents // Proc. VLDB Endow. 14, 6 (February 2021), pp. 997–1005. https://doi.org/10.14778/3447689.3447703

278. Скатов Д. С., Ерехинская Т. Н., Окатьев В. В Модели и методы анализа иерархически структурированных текстов // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М.: РГГУ, 2009. – С. 458-464.

279. Hovorushchenko T. and Pomorova O. Information Technology of Evaluating the Sufficiency of Information on Quality in the Software Requirements Specifications // ICTERI Workshops, 2018. pp. 555-570.

280. Липаев В. В. Надежность и функциональная безопасность комплексов программ реального времени: Монография. – М: Институт системного программирования РАН, 2013. 207 с.

281. Hovorushchenko T., Pavlova O. September. Evaluating the software requirements specifications using ontology-based intelligent agent // 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). 2018. Vol. 1, pp. 215-218.

282. Wang Y., Yin F., Liu J., Tosato M. Automatic construction of domain sentiment lexicon for semantic disambiguation // Multimedia Tools and Applications. 2020. V.79. pp. 22355-22373.

283. Loukachevitch N., Dobrov B. Ontologies for Natural Language Processing: Case of Russian // Third International Conference Computational Linguistics in

Bulgaria. 2018. p. 93.

284. Gavrilova T. A., Leshcheva, I. A. Ontology design and individual cognitive peculiarities: A pilot study // Expert system with Applications, 2015. pp. 3883-3892.

285. Globa L., Kovalskyi M., Stryzhak O. Increasing web services discovery relevancy in the multi-ontological environment // The series «Advances in Intelligent and Soft Computing» (AISC), Springer, 2015. pp. 335-344.

286. Smirnov A., Levashova T., Shilov N. Patterns for Context-based Knowledge Fusion in Decision Support // Information Fusion. 2015. Vol. 21. pp. 114–129.

287. Бутенко Ю.И., Сидняев Н.И., Казанцева Е.С. Оптимальные и адаптивные самонастраивающиеся системы в динамических структурах управления // Физические основы приборостроения, 2022, №1(43), С.38-43.

288. Сигов А. С., Нечаев В. В., Кошкарёв М. И. Архитектура предметно-ориентированной базы знаний интеллектуальной системы // International Journal of Open Information Technologies. 2014. №12. С. 1-6.

289. Helbig H. Knowledge Representation and the Semantics of Natural Language. – Berlin, Heidelberg, New York, 2006. 655 p.

290. Елисеев Д. В. Модель представления знаний при создании адаптивной информационной системы // Наука и образование: научное издание МГТУ им. Н.Э. Баумана. 2010. №03. С. 1-6.

291. Ломов П.А., Шишаев М. Г. Формирование когнитивных фреймов на основе онтологических паттернов для визуализации онтологий // Информационные системы и технологии. 2015. Т. 92. №. 6. С. 12-22.

292. Даниленко В. П. Русская терминология: опыт лингвистического описания. М.: Наука, 1977. 246 с.

293. Шарафутдинова Н. С. Немецко-русский синонимический словарь авиационных терминов.Ульяновск: УлГТУ, 2016. 196 с.

294. Fillmore Ch. J. The Case for Case // Universals in Linguistic Theory. London: Holt, Rinehart and Winston, 1968. pp. 1-25.

295. Богданов В. В. Структурно-семантическая организация предложения.

Л.: Изд-во ЛГУ, 1977. 205 с.

296. Jackendoff R. S. The Status of Thematic Relations in Linguistic Theory // Linguistic Inquiry. 1987. Vol. 16. pp. 369-411.

297. 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.

298. 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.