

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Институт лингвистики и международных коммуникаций
Кафедра лингвистики и перевода

ДОПУСТИТЬ К ЗАЩИТЕ
Заведующий кафедрой,
д.филол.н., доцент
_____ /Т.Н. Хомутова/

АВТОМАТИЧЕСКАЯ ОБРАБОТКА ТЕКСТА (ПРОБЛЕМА ЛЕММАТИЗАЦИИ)

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ЮУрГУ – 45.03.03.2018. 286.ВКР

Руководитель, к.филол.н., доцент
_____ /Е.В. Кравцова/
« ____ » _____ 2018 г.

Автор,
студент группы ЛМ-437
_____ /О. И. Сухих/
« ____ » _____ 2018 г.

Нормоконтролер,
к.филол.н., доцент
_____ /О.И. Бабина/
« ____ » _____ 2018 г.

Работа защищена с оценкой

« ____ » _____ 2018 г.

Челябинск

2018

ОГЛАВЛЕНИЕ

Введение.....	3
Глава 1 Системы автоматической обработки текста.....	6
1.1 Компьютерная лингвистика, ее цели и задачи.....	6
1.2 Лингвистический автомат	12
1.3 Системы АОТ и их уровневое построение.....	14
1.4 Современные зарубежные методы АОТ.....	16
1.5 Анализ электронных текстов	17
Выводы по главе 1.....	27
Глава 2 Корпусная лингвистика. Проблема исходной формы слова	29
2.1 корпуса текстов: принципы построения	29
2.2 Текст. Общие положения	34
2.3 Электронный текст как основа корпуса	399
2.4 Лемматизация и нормализация. Проблема исходной (словарной) формы слова.....	41
2.5 Правила приведения словоформ к исходной (словарной) форме слова ...	42
2.6 Программное обеспечение	44
Выводы по главе 2.....	47
Заключение	49
Библиографический список	51

ВВЕДЕНИЕ

В настоящее время основными проблемами лингвистики являются изучение лексики и семантики, а также быстрый автоматизированный перевод. К середине XX века, непрерывный рост объемов производимой информации сделал крайне актуальными задачи поиска информации в огромных объемах данных, ее выбора и упорядочения по тем или иным признакам. В данных исследованиях невозможно обойтись без работы со словарями, энциклопедиями, архивами. Но, к сожалению, у учёных не всегда существует возможность доступа к необходимым информационным ресурсам. Помочь в этом современным лингвистам может такая отрасль науки, как компьютерная, прикладная лингвистика, которая занимается созданием разнообразных систем по обработке естественного языка. Но эта обработка невозможна без наличия лингвистических информационных ресурсов.

Появление вычислительной техники способствовало в 1960-е гг. созданию различных теорий в области лингвистики и представления знаний (Ю.Д. Апресян, М. Мински, Д.А. Поспелов, Р. Шенк, И. Уилкс, В.А. Звягинцев, Т. Виноград, А.К. Жолковский, Ч. Филмор и др.), развитию методов автоматической обработки текста.

В последние десятилетия появилось множество систем автоматической обработки текста, предназначенных для решения отдельных или небольшого набора задач. В связи с вышеизложенным, наиболее **актуальным** вопросом в современной прикладной лингвистике являются методы и анализ автоматической обработки текстов.

Объектом исследования является процесс автоматической обработки корпуса текстов зарубежных научно-фантастических произведений.

Предметом нашего исследования являются автоматическая лемматизация текста и его автоматическая обработка, метод морфологического анализа словоформ.

Целью данной дипломной работы является создание автоматизированного лемматизатора текста.

Достижение поставленной цели предполагает решение следующих **задач**:

1. Проанализировать современные зарубежные методы автоматической обработки текста (АОТ).
2. Рассмотреть понятия «электронный текст» и «корпус текстов».
3. Создать корпус текстов научно-фантастических зарубежных произведений на английском языке.
4. Выявить правила приведения словоформ к исходной (словарной) форме слова.
5. Создать автоматизированный лемматизатор и провести лемматизацию корпуса научно-фантастических зарубежных произведений.

Материалом исследования послужили научно-фантастические произведения, а именно: роман английского писателя Герберта Уэллса «Машина времени»; роман «Грядущая раса» английского писателя Эдварда Бульвера-Литона.

Теоретико-методологической базой для дипломной работы послужили труды Е.И. Большаковой, Э.С. Клышинского, Д.В. Ландэ, А.А. Носкова, О.В. Песковой, Е.В. Ягуновой, И.С. Николаева, О.В. Митрениной, Т.М. Ландо, А.В. Луканина, Р.Г. Пиотровского и других.

В работе использовались такие **методы и приемы анализа** как логический, дискурсивный, метод корпусной лингвистики.

Научная новизна работы определяется в том, что впервые создан лемматизатор, работу которого мы проверяли на основе корпуса текстов зарубежных научно-фантастических произведений.

Теоретическая значимость работы состоит в том, что полученные в ходе исследования выводы вносят определённый вклад в развитие компьютерной и корпусной лингвистики.

Практическая значимость данной работы состоит в том, что данную программу можно использовать для любого корпуса текстов на английском

языке. Также она состоит в возможности использования его результатов в вузовских курсах по новым информационным технологиям, компьютерной лингвистике, автоматической обработке текста, лексикологии, лексикографии.

Цель и задачи исследования определили **структуру и объем работы**, которая состоит из 2 глав, заключения и списка литературы, состоящего из 31 источника.

ГЛАВА 1 СИСТЕМЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТА

1.1 Компьютерная лингвистика, ее цели и задачи

Компьютерная лингвистика (КЛ) появилась на пересечении таких наук, как лингвистика, математика, информатика (Computer Science) и искусственный интеллект. Истоками компьютерной лингвистики являются исследования известного американского ученого Н. Хомского в области формализации структуры естественного языка (ЕЯ); развитие КЛ происходит на основе результатов в области общей лингвистики (языкознания). Языкознание изучает общие законы естественного языка – его структуру и функционирование, и включает такие области как:

1. фонология – изучает звуки речи и правила их соединения при формировании речи;

2. морфология – занимается внутренней структурой и внешней формой слов речи, включая части речи и их категории;

3. синтаксис – изучает структуру предложений, правила сочетаемости и порядка следования слов в предложении, а также общие его свойства как единицы языка;

4. семантика и прагматика – данные области очень тесно взаимосвязаны: семантика занимается смыслом слов, предложений и других единиц речи, а прагматика – особенностями выражения этого смысла в связи с конкретными целями общения;

5. лексикография описывает лексикон конкретного ЕЯ – его отдельные слова и их грамматические свойства, а также методы создания словарей [Большакова 2011, с. 90].

Результаты Н. Хомского, которые были выявлены на стыке лингвистики и математики, заложили основу для теории формальных языков и грамматик (часто называемых генеративными, или порождающими грамматиками). Эта теория относится к математической лингвистике и используется для обработки не столько ЕЯ, но искусственных языков, а главным образом –

языков программирования. По общему складу данной дисциплины она вполне математическая.

Термин компьютерная лингвистика (КЛ) в последнее время намного чаще встречается в связи с разработкой различных прикладных программных систем, в их числе – коммерческие программные продукты. Связано это с достаточно бурным ростом в обществе текстовой информации, так же и в сети Интернет этот рост не является исключением, и необходимостью автоматической обработки текстов на естественном языке (ЕЯ). Отмеченное обстоятельство стимулирует развитие компьютерной лингвистики как области науки и разработку новых информационных и лингвистических технологий. В рамках компьютерной лингвистики, которая зародилась уже более 50 лет назад (и также известной под названиями машинная лингвистика, автоматическая обработка текстов на ЕЯ) предложено много перспективных методов и идей, но далеко не все они еще нашли свое выражение в программных продуктах, используемых на практике [Большакова 2011, с. 91].

К математической лингвистике учёные относят также лингвистику называемую квантитативной, которая изучает частотные характеристики языка – слов, их комбинаций, синтаксических конструкций и тому подобные, при этом используются математические методы статистики, поэтому данный раздел науки можно назвать статистической лингвистикой.

Компьютерная лингвистика находится в тесной связи с такой междисциплинарной научной областью, как искусственный интеллект (ИИ), в рамках которого разрабатываются компьютерные модели отдельных интеллектуальных функций. Одна из первых программ в области ИИ и КЛ, при чём не просто разработанных, а работающих – это всеобщая известная программа Т. Винограда, которая понимала простейшие приказы человека по изменению мира кубиков, сформулированные на ограниченном подмножестве ЕЯ. Отметим, что несмотря на очевидное пересечение исследований в области КЛ и ИИ (так как владение языком относится к

интеллектуальным функциям), ИИ не поглощает всю компьютерную лингвистику, поскольку она имеет свой теоретический базис и методологию. Компьютерное моделирование как основной метод и итоговая цель исследований являются общим для данных указанных наук. Таким образом, разработка компьютерных программ для автоматической обработки текстов на ЕЯ является более полной и точной формулировкой задачи компьютерной лингвистики.

Перед компьютерной лингвистикой стоят **задачи** лингвистического обеспечения процессов сбора, накопления, обработки и поиска информации. Самыми выделяющимися и важными из них являются:

1. Автоматизация составления и лингвистической обработки машинных словарей.

2. Автоматизация процессов обнаружения и исправления ошибок при вводе текстов в ЭВМ.

3. Автоматическое индексирование документов и информационных запросов. Создание поискового образа документа предполагает индексирование его текста, т.е. выделение в нем ключевых слов.

4. Автоматическая классификация и реферирование документов. Компрессия текста (реферирование и аннотирование). Решение этой задачи состоит из двух этапов:

1) сегментация на высказывания (части высказываний);

2) выбор наиболее значимых (синтез).

5. Лингвистическое обеспечение процессов поиска информации в одноязычных и многоязычных базах данных

6. Машинный перевод текстов с одних естественных языков на другие. В настоящее время существует целый спектр компьютерных систем перевода разного качества, но, несмотря на то, что данное направление развивается уже на протяжении многих десятилетий, в целом задача машинного перевода еще не нашла своего решения и даже можно сказать, что она далека от полного решения этой задачи.

7. Построение лингвистических процессоров, благодаря которым пользователи имеют возможность общения с автоматизированными интеллектуальными информационными системами (также в частности, с экспертными системами) на естественном языке, или на том языке, который является достаточно близким к естественному;

8. Извлечение фактографической информации из неформализованных текстов. Извлечение фактов и знаний (Information Extraction). Извлечение информации из текстов часто требуется при решении задач экономической и производственной аналитики. Для произведения всех этих действий осуществляется выделение каких-то определенных объектов в тексте ЕЯ — именованных сущностей (имен, персоналий, географических названий), их отношений и связанных с ними событий. Как правило, данные действия реализуются на основе неполного синтаксического анализа текста, который даёт возможность выполнять такие действия как, например, обработку потоков новостей от информационных агентств [Большакова 2011, с. 90].

Машинные словари являются незаменимой частью любой системы автоматической обработки текстовой информации. Они могут представлять собой словари слов и/или словари словосочетаний, выражающих устойчивые научно-технические понятия. В процессе составления словарей необходимо так тщательно их анализировать, чтобы в результате они в максимальной степени отражали лексический состав текстов. Поэтому их нужно составлять на основе текстов достаточно большого объема (как минимум, по текстам объемом в несколько десятков миллионов лексических единиц). А такая работа является очень трудоемкой и может быть выполнена в разумные сроки только, если в процессе будет использоваться широкое применение средств автоматизации.

В словарях словосочетаний научно-технических понятий словосочетания не выделяются в тексте формально, их границы определяются сознанием пользователя, поэтому их автоматизированное составление характеризуется большей сложностью по сравнению с задачей составления словарей слов.

Однако эксперименты показывают, что границы именных словосочетаний можно определить с достаточной точностью посредством несложных процедур синтаксического анализа. Возникшие при этом ошибки можно устранить посредством применения статистических методов и редактирования уже составленного словаря.

9. Распознавание звучащей речи и синтез речи по тексту. Прибор для распознавания речи был изобретен в начале 50-х годов прошлого века. Устройство обладало способностью узнавать сказанные человеком цифры. На данный момент есть несколько способов распознать речь. В первую очередь, это распознавание отдельных команд. В процессе распознавания отдельных речевых команд из заранее определенного словаря достигается самая высокая достоверность. В качестве примера можно привести голосовую навигацию по сайтам. Распознавание фраз, которые соответствуют заданным правилам, широко используется в системах голосового самообслуживания. В ходе поиска слов-ключей в потоке слитной речи речь преобразуется в текст не в полной мере. Автоматически в речи определяются участки, содержащие заданные фразы или слова. Такой поиск преобладает в системах поиска и мониторинга речи. Технология распознавания речи на большом словаре является наиболее близкой к человеческим мечтам о взаимодействии с автоматами. В этом случае все, что говорится людьми, преобразуется в текстовую информацию. Иногда технологию называют *speech to text (STT)*. Однако данная задача в полной мере пока не решена, но, следует отметить достаточно высокий уровень достоверности распознавания.

10. Поддержка ввода текста на электронные носители. Программы автоматического переноса слов и орфографической проверки текста (спеллеры) – это первые приложения в данном направлении. С помощью программ обеспечивается коррекция на синтаксическом и лексико-морфологическом уровне. Вход сличается с перечнем допустимых структур, так называемое распознавание с дискретным входом. В случае неудачи

выполняется поиск ближайшего соответствия. Близкими этим задачам по смыслу можно считать автозавершение и распознавание рукописного и печатного текста.

11. Классификация текстов. Она необходима при выполнении операций создания объемных коллекций документации, когда актуальны задачи кластеризации и классификации с целью выявления классов, которые наиболее близки теме документов.

- Анализ нормативных текстов. Аналитика текстов законодательных актов, постановлений, указов, планов работ делается с целью выявления логических пропусков и противоречий. В пример можно привести анализ одного из региональных законодательных актов, выявивший неполноту описания. Описание действий субъектов, которые должны выполнять закон, а также алгоритм его выполнения, было сделано всего на 30%. В результате такой закон не способен эффективно функционировать.

- Вопросно-ответные системы (Question Answering). Решение данной задачи осуществляется посредством выявления типа вопроса, поиска текстов, которые в потенциале содержат ответ на вопрос, извлечения ответа из текстов. За решение подобных проблем отвечает технология Text Mining (интеллектуальный анализ текста). В перечень входит выделение понятий и феноменов, реферирование, кластеризация и классификация, тематическое индексирование, ответы на вопросы, поиск по ключевым словам.

- Диалог с компьютерными системами на естественном языке. Данная задача имеет отношение к специализированным базам данных. Здесь приходится сталкиваться с ограниченным языком запросов в лексическом и грамматическом плане. Такое положение дел дает возможность применять упрощенные языковые модели. Запросы к БД, которые сформулированы на естественном языке, переводят на формальный язык. После этого осуществляется поиск необходимой информации и построение соответствующей фразы ответа [Большакова 2011, с. 91].

1.2 Лингвистический автомат

История АОТ (автоматическая обработка текста) в период с 60 по 80-е годы 20-го века характеризуется как эпоха романтических проектов. В это время разработчики заняты идеей создания искусственного разума и машинного перевода высокого качества, который по своей эффективности мог бы составить конкуренцию высококвалифицированным переводчикам. Во второй половине 80-х годов прошлого века произошла вторая когнитивная революция. Она окончательно поставила точку на логистико-романтических подходах к решению проблем АОТ, разъяснив несостоятельность исходных положений и провальность подобных проектов.

Сейчас реально функционирующие АОТ – это достаточно грубые аналоги некоторых аспектов речемыслительной человеческой деятельности (РМД), которые объединяются концепцией лингвистического аппарата (ЛА) [Пиотровский 2008, с.34].

Лингвистический аппарат рассматривается в качестве сбалансированного комплекса программных, аппаратных, лингвистических, лингводидактических средств, которые взаимодействуют с внушительной базой лингвистических данных и знаний (БДЗ). В идеальном варианте БДЗ должны обладать такими характеристиками, как:

- полифункциональность. БДЗ должны быть способны выполнять различные виды обработки больших потоков информационных текстов;
- минимизация информационных потерь и ослабление эффекта отторжения ЕЯ языком ЛА;
- живучесть. Под живучестью понимают способность лингвистического аппарата сохранять наиболее важные качества после воздействия негативных факторов, например, после выхода из строя участков оперативной памяти, искажения фрагментов текста и других;
- способность к дальнейшему развитию и совершенствованию, продиктованных необходимостью адаптировать лингвистический аппарат к

коммуникативно-информационной эволюции общества к прагматике отдельных потребителей;

- возможность подключиться к разным каналам связи, включая Интернет.

Лингвистический аппарат имеет следующие блоки:

- ввода, распознавания и первичной обработки печатного текста;
- озвучивающий печатный текст;
- распознавания устной речи, который преобразует звуковой сигнал в текст;
- машинного перевода;
- БДЗ;
- индексирования и аннотирования текста;
- вывода печатного текста;
- управления и средства человеко-машинной синергетики [Луканин 2011, с. 3-4].

Следует упомянуть также о распространенном термине АПТ, который расшифровывается как автоматическая переработка текста. В работах Р.Г. Пиотровского используется именно АПТ, учитывая тот факт, что текст изменяется или дополняется служебной информацией в процессе переработки.

Обработку текста в лингвистический аппарат в общем виде можно представить так. Текст, принятый блоком ввода и коррекции, отправляется на вход лексико-морфологического анализатора. Лексико-морфологический анализатор посредством подблока БДЗ строит подстрочник. Далее подстрочник следует в подблоки семантического и синтаксического анализа. Так происходит разбивка подстрочника на глагольные и именные группы. Каждая группа подвергается анализу и обработке с помощью графов переходов. Результат описанных действий поступает в подблок порождения выходного текста, например, аннотации, перевода, индекса, и уже после этого отправляется потребителю. Ввод, обработку и выдачу текста

обеспечивает многофункциональная оболочка системы, оформленная в виде АРМ (автоматизированное рабочее место).

В рамках лингвистического аппарата создается интегрированная среда. Пользователю не нужно выходить из нее, чтобы создать и оформить качественную обработку текстовой информации. Интегрированная среда позволяет оперативно выполнять технологию обработки, начиная с ввода документа в ПК, постредактирования и завершая работу готовым переводом, аннотацией или индексом текста на нужном языке и необходимом уровне.

Современные технологии дают возможность в полном объеме использовать возможности, предоставляемые программами проверки орфографии, главным и карманным сканерами, адаптером для связи с ПК и другими считывающими устройствами [Луканин 2011, с. 4].

1.3 Системы АОТ и их уровневое построение

Многообразие систем автоматической обработки неструктурированных текстов сегодня вызывает необходимость их систематизации и классификации с целью упрощения выбора решения, наиболее адекватного для конкретной задачи [Луканин 2011, с. 5].

Обработка текстов, которые представляют собой неструктурированную информацию, например, патенты, истории болезней, диссертации, преследует задачи, которые условно делятся на категории:

1. Распространенные пользовательские задачи, с которыми потребители сталкиваются постоянно. Здесь можно отметить фильтрацию спама, проверку орфографии, автоперевод небольших текстовых фрагментов.

2. Обработка внушительных массивов текста, например, полноценный автоперевод целостных текстов, поиск релевантных ответов на вопросы, построение рекомендательных систем, которые работают с большим объемом неструктурированной информации, аналитика отзывов и мнений [Луканин 2011, с.5].

Отличительной особенностью данных задач является отсутствие формализации и сложность. В реально работающих современных системах данные проблемы не решены. Вместо полноценного набора решений используются вспомогательные методы, например, такие, как:

- выделение ключевых словосочетаний и слов;
- классификация текстов;
- суммаризация (автоматическое реферирование).

Здесь большое внимание уделяется технологиям визуализации больших объемов текстовой информации.

Неотъемлемой частью многих систем обработки текстов являются корпуса. Слова в корпусах наделены полными грамматическими характеристиками, например, часть речи, форма, синтаксическая роль. Корпусы – это входные данные для обучения в задачах классификации текстов по жанрам и темам, синтаксических программ и парсеров, которые применяются для снятия омонимии и допуска анафоры. Для обучения машинных переводчиков применяются параллельные корпуса, которые состоят из одинаковых текстов на разных языках. Сбор корпусов осуществляется десятилетиями. Это очень трудоемкое исследование с участием больших групп научных специалистов. В качестве примера можно привести проект под названием «Национальный корпус русского языка». Данный проект реализуется уже тринадцать лет при поддержке компании «Яндекс» [Козлова 2013, с. 14].

Морфологические словари являются важным типом входных данных любой АОТ. Здесь можно упомянуть библиотеку «АОТ», которая применяется во многих коммерческих и исследовательских проектах. Библиотека представляет собой словарь Зализняка в цифровом варианте.

Еще одним распространенным типом входных данных являются семантические сети (тезаурусы). WordNet – самый известный тезаурус. WordNet – ресурс связанных между собой слов. Связь между словами осуществляется по типу семантических отношений. Например, гипонимия

(обобщение – частное), синонимия, гиперонимия (частное – обобщение), меронимия (часть – целое). WordNet эффективен при решении задач классификации текста, машинного перевода, генерации текстов. Стоит отметить, что пока, к сожалению, русский аналог WordNet, не разработан.

Развития АОТ-систем, уже в наши дни представляющих коммерческий интерес и использующихся при решении следующих прикладных задач:

1. Machine Translation and Translation Aids – машинный перевод;
2. Text Generation – генерация текста;
3. Localization and Internationalization – локализация и интернационализация;
4. Controlled Language – работа на ограниченном языке;
5. Word Processing and Spelling Correction – создание текстовых документов (ввод, редактирование, исправление ошибок);
6. Information Retrieval – информационный поиск и связанные с ним задачи.

Нужно отметить, что это деление достаточно условное, и в реальных системах часто встречается объединение функций. Так, для машинного перевода требуется генерация текста, а при исправлении ошибок приходится заниматься поиском вариантов словоформы и т.д.

1.4 Современные зарубежные методы АОТ

На современном отрезке времени существует острая необходимость в оперативном создании приложений (прикладных программных сетей) для автоматизированной или автоматической обработки текстов на ЕЯ. Это обусловлено активным ростом объемов текстовой информации. В качестве примеров подобной обработки можно привести фильтрацию и сбор данных из разных, которые находятся в разных источниках, реферирование, извлечение знаний, аннотирование. При разработке приложений часто возникают такие сложности, как интеграция огромного числа программных компонентов, которые выполняют алгоритмы текстов на естественном языке,

работают на разных уровнях текста, например, обработка, абзацев, слов, предложений.

Для решения задач АОТ необходимо выполнить аналитику текста на различных уровнях представления. Виды анализа:

1. Графематический, в ходе которого из массива данных выделяются предложения и слова (токены).

2. Морфологический, в ходе которого выделяется грамматическая основа, определяется часть речи, слова приводятся к словарной форме (лемматизация).

3. Синтаксический, в ходе которого выявляются синтаксические связи между словами в предложениях, строится синтаксическая структура предложений.

4. Семантический, в ходе которого выявляются семантические связи между синтаксическими группами и словами, извлекаются семантические отношения.

Каждый из описанных выше анализов представляет собой самостоятельную задачу. Она не имеет своего практического применения, но активно используется в качестве составной части более глобальных задач.

1.5 Анализ электронных текстов

В связи с развитием современных средств коммуникации и все более широким распространением нового типа текста – электронного – в лингвистике намечается отход от традиционного понятия текста как объединенной смысловой связью (линейной) последовательности знаковых единиц, основными свойствами которой являются связность и цельность. Представляется, что эволюция электронных текстов достигла того этапа, когда стало необходимо выделить отдельную отрасль внутри текстовой лингвистики, которая бы занималась изучением именно их особенностей. Электронный дискурс исследуется как зарубежными, так и отечественными учеными, однако большинство работ посвящено анализу отдельных типов

электронных текстов (социальных сетей, блогов, электронных периодических изданий, чатов, форумов и т. п.).

В Интернете как в глобальном информационном социокультурном пространстве существует и размещается информация, включая электронные тексты. Электронный текст в качестве нового феномена современной культуры появился и начал изучаться благодаря активному созданию электронных библиотек, под которыми понимаются коллекции электронных текстов. Электронный текст, наделенный качествами традиционного текста в плане содержания и информативности, характеризуется рядом особенностей, которые связаны со спецификой сетевой среды, а также отличительной формой представления информации, а также ее поиска с помощью системы ссылок и гипертекста. Понятие «электронный текст» трактуется как текст, существующий в электронной среде, сети или локальном доступе. Традиционный текст зачастую соотносят с конкретным изданием или книгой, а электронный может представлять различные издания возможного жанра. Традиционная форма включает такие типологии и классификации, как газеты, журналы, энциклопедии, альбому и так далее. Называя данные издания, мы подразумеваем текст в виде книги. Для электронной формы жанровость имеет второстепенное значение. Электронный текст представляет собой любую информацию, позволяет находить и выделять факты и цитаты, а система отсылок и гиперссылок дает возможность активно читать электронные тексты. Также употребляется формулировка «электронный документ». Электронный документ представляет собой информацию, которая фиксируется в виде набора символов, изображения или звукозаписи на материальном носителе и предназначена для передачи в пространстве и времени посредством электросвязи и средств вычислительной техники для сохранения и общественного применения. Информация в электронном документе представлена электронно-цифровом виде.

В кратком описании особенностей прикладного подхода к пониманию текста будем ориентироваться на книгу Н.Н. Леонтьевой, одного из признанных авторитетов в этой области. Прежде всего, отметим, что автоматическое понимание текстов является необходимой частью разнообразных прикладных задач. Вполне очевидно, что, например, задачи машинного перевода и автоматического аннотирования (или реферирования) суть разные задачи, предполагающие разный результат автоматического понимания текста. Путь учета реальности таких разных подходов Н.Н. Леонтьева видит в последовательном применении идеи «мягкого» понимания текста. «Мягкое» понимание можно трактовать как подстройку работы автомата под разные коммуникативные цели. Процедуры автоматического понимания, в отличие от естественного понимания, подразумевают разделение ролей. Это означает, что определение цели и оценка результата выполняется человеком. Задача автомата заключается в понимании текста в соответствии с установленной человеком целью и оценивании результат на промежуточных уровнях. Результат понимания реализован в форме конкретной семантической структуры. По мнению Н.Н. Леонтьевой существует перечень структур, в который входит:

- Семантическая сеть целого текста (глобальное размытое понимание).
- Информационная структура целого текста (глобальное обобщенное понимание).
- «Лингвистическая структура предложений текста (локальное понимание).
- Структура баз данных и знаний (выборочное специальное понимание)» [Белоногов 2004, с. 138].

К построению семантических сетей целого текста прибегают авторы многих современных работ. Такие сети представляют собой размытую структуру понимания глобального типа. Посредством данной глобальной сети выполняется реализация шага «Смысл \Leftrightarrow Текст». В ее рамки включен глубинно-семантический компонент. Под «смыслом текста» понимается

результат перевода семантико-синтаксических представлений предложений текста на язык элементарных единиц. Для выхода в данную сеть необходимо осуществить ввод коммуникативных (информационных) отношений между предложениями и внутри них. В данном случае идет речь об установлении рема-рематических (внутриструктурных) и референтных (между соседними предложениями) связей. Существует целый ряд случаев, когда в дополнительном глубинно-семантическом компоненте выполняется объединение языкового и энциклопедического представления.

Задача информационных структур целого текста (потоков текстов) состоит в фиксации в качестве результата обобщенное понимание текста. Данное понимание фиксируется в единицах терминологии определенной предметной области по классификаторам, рубрикаторам, тезаурусам и так далее. Такие структуры применяются в ИПС (информационно-поисковых системах). ИПС функционируют на основе материала произвольных текстов и практически не имеют ограничений в плане тематической области и текстовой структуры. Результат работы систем – поисковый образ документа. Безусловным плюсом для оценки эффективности такого подхода автоматического понимания является масштабность разработок и высокая востребованность различных информационно-поисковых систем. Данный подход ограничивается, главным образом, небольшим смысловым потенциалом, имеется в виду обобщенное понимание текстов в огромных масштабах работы систем. В лучшем случае результат подобного рода можно соотнести с итогом классификации текстов по тематической области в соответствии с очень грубой ситуативной моделью. Данная классификация является усеченным вариантом понимания текста. Однако если смотреть с другой стороны, описанный подход может соответствовать проблематике изучения работы ключевых слов в рамках ИПС. Здесь речь идет об определении тематической области текста на основе ключевых слов, выделенных из текста. Особенно это имеет смысл, если ключи являются терминологическими элементами.

Лингвистическая структура предложений текста осуществляет фиксацию результата локального понимания, которое ограничено рамками каждого предложения. К самым известным лингвистическим структурам относят структуры на основе модели «Смысл \Leftrightarrow Текст». Основа семантико-семантического представления – синтаксическое дерево предложения, которое имеет семантические узлы или связи. Функционирование автомата на базе рассматриваемой модели выполняется с опорой на объемные и сложные словари, являющиеся компонентами работы автомата. В результате работы представляется информация о связях и единицах в границах предложения.

Преимуществом и одновременно недостатком такого подхода является формализованность представления. Главное достоинство подобных структур заключается в детальности аналитики, которая отражается в форме дерева. Данное дерево представляет собой семантическое и синтаксическое представление структуры предложения. Автомат выполняет построение правильной синтаксической структуры (поверхностной, а затем структурной) в условиях наличия словарных статей для всех слов предложения, а также в условиях правильности структуры в плане законов входного языка. Семантическую структуру можно получить, если выполнить замену всех узлов глубинной семантической структуры соответствующими словарными толкованиями, сохранив все связи глубинной синтаксической структуры и расширив нотацию.

В рамках системы машинного перевода ЭТАП-2 реализуется единая синтаксическая структура. В ее узлах помещены слова исходной фразы. В данной структуре сохраняются подробные связи поверхностной структуры. Данные связи достаточно дифференцируемы, потому что возможен их перевод в семантический план, а соединяемые ими слова имеют семантические характеристики в комбинаторном словаре. Недостаток жестких древовидных структур заключается в невозможности выхода за

границы предложения, невозможности выборочного восприятия (схватывания наиболее важной информации).

Другим ограничением является слабая корреляция с системами представления знаний. «Пока реально достижимое СемП /семантическое представление/ целого – это последовательность СинСемП /синтактико-семантических представлений/ всех подряд предложений текста». Лингвистическая структура предложений текста может быть соотнесена с уровнем, который предшествует построению текстовой основы как системы пропозиций текста. Данные структуры соотносят в какой-то степени с поверхностными структурами, которые предполагают работу человека с эксплицитными поверхностными структурами (не более одного предложения или клаузы).

Лингвистические структуры характеризуются большим объемом. Они включают в себя и глубинное, и поверхностное представление о структуре единицы (предложения). Данное направление в работах по автопониманию текста приближены к моделированию восприятия текста, который лишен цельности и связности, адресатом.

Эксперименты подобного рода проводились на текстах, которые характеризуются:

- предложениями, которые перемешаны случайным образом;
- клаузами, которые перемешаны случайным образом.

Если принимать во внимание задачи настоящего исследования, данное направление можно соотнести с моделированием восприятия текста адресатом, если он может опираться только на структуру текущего предложения и не может пользоваться механизмами контекстной предсказуемости, которые связывают текстовые фрагменты за рамками одного предложения. Тем не менее, ранее уже говорилось о том, что смысл текста – это не сумма смыслов предложений, из которых состоит текст, и понимание его человеком не ограничивается только этим уровнем.

Структуры баз данных и знаний – это специальное, выборочное понимание, которое по максимуму учитывает экстралингвистическое представление, отображение части определенной действительности. Структуры БД (баз данных) представляют собой жестко фиксированные, формальные структуры, над которыми можно выполнить математические операции. Это может быть, к примеру, таблица с данными о кадровом составе учреждения с заранее заданными полями. Структуры баз знаний представляют собой полужесткие структуры динамического типа, например, фреймы, сценарии. Они широко распространены в системах искусственного интеллекта и отображают представление целого текста. К членению на предложения данные структуры безразличны. Р. Шенк представляет системы, в которых данные структуры ориентированы на распознавание определенного сюжета в тексте. Заданную тему текста можно рассматривать в качестве квазиденотата, поэтому такой подход часто называется денотативным. Главное ограничение подхода заключается в «ограничении на мир». В экстралингвистических моделях иллюстрируется зависимость понимания от предварительных знаний о предмете. Однако подобные модели плохо или совсем никак не способны моделировать несюжетные тексты, включая научно-технические.

Н.Н. Леонтьева предлагает идею «мягкого» понимания, а также информационно-лингвистическую модель понимания. Данные идеи направлены на примирение лингвистической и информационной анализы, на обеспечение результативного взаимодействия разных уровней обработки текста. В процессе информационного анализа потеря части информации, которая иначе называется информационным сбросом, неизбежна. Более или менее информативные составляющие текста должны определяться с опорой на лингвистические исследования. Выполнение процедуры понимания «снизу – вверх» (от поверхностных структур к денотативным представлениям) можно описать следующим образом. Их главное предназначение заключается в том, чтобы создавать контекст, который

достаточен и необходим для вычленения информативных единиц, переходящих в структуры следующего уровня, на каждом уровне. Информационный сброс, контролируемый лингвистически, дает возможность автомату работать при отсутствии идеальных условий. В этом случае появляется возможность снимать структурные ограничения на обрабатываемые тексты, например, если автомат принимает на вход неполные или синтаксически неправильные предложения, допускает работу с неполными БД или словарями. Можно предложить, что изучение модели рассматривается в качестве моделирования понимания текста искусственным носителем языка в разных условиях коммуникации для текстов различных функциональных стилей.

Компрессия представляет собой один из наиболее востребованных механизмов автоматической обработки текста. Задача компрессии состоит в получении аннотации или реферата, которые представляют собой компактную формулировку одного текста или монотематического (группы текстов) текстового массива. Степень сжатия и принципы определяют задачи системы. Реферат или аннотация – это вторичные тексты. Основной проблемный вопрос, который решается при автоматическом моделировании понимания текста и построении вторичных текстов – это связность и цельность текста. Данный вопрос невозможно решить без рассмотрения проблем референции.

Формализуемые средства обеспечения связности:

- понятия, повторяющиеся в тексте (объекты, субъекты, явления и так далее) в одном лексическом выражении;
- понятия, понятия, повторяющиеся в текстах (объекты, субъекты, явления и так далее) в разных лексических выражениях. В качестве примера можно привести однокоренные дериваты или слова одного лексико-семантического поля;
- местоименные слова и местоимения, как правило, они имеют отношение к средствам выражения понятий, повторяющихся в тексте;

- «слова-текстопостроители», которые обозначают обобщенные логикокомпозиционные связи между компонентами. Это разные уровни текстовых составляющих. В качестве примера можно привести такие слова, как все же, особо подчеркнем, резюмируя, итак, однако и так далее;

- союзные слова, которые занимают промежуточные положение и характеризуют в основном связи между клаузами. Промежуточное положение заключается в том, что с одной стороны союзы передают синтаксические отношения, с другой стороны, их значение соотносят со значением некоторого полнозначного знаменательного слова (повторение понятия).

И.П. Севбо в своих трудах для повторения одних и тех же понятий независимо от их лексического выражения ввел понятие нанизывания. Компрессионный текст можно получить посредством его предварительного развертывания. Схемы нанизывания строятся с помощью канонических кустов, в которых выполняется восстановление всех связей. И.П. Севбо в своем исследовании под названием «Структура связного текста и автоматизация реферирования» описал результат проведенного им эксперимента по составлению аннотации текстов различных функциональных жанров на базе особенностей нанизывания. Речь идет о синтаксической структуре упрощенных нормализованных предложений и сведениях о повторяемости слов и понятий в тексте [Севбо 1969, с. 46].

Как правило, автоматическое реферирование осуществляется посредством следующих способов или их комбинации:

1. Отбор из текста наиболее существенных предложений на основании статистического алгоритма. Далее с помощью синтаксического анализа из этих предложений выделение самых значимых фрагментов.

2. Использование на первом этапе алгоритма синтаксического анализа. За счет этого выделяются наиболее важные части предложений. На следующем этапе статистический анализ проводится только в отношении наиболее важных частей предложений текста.

3. Определение «веса слова» посредством синтаксического и статистического анализа. Вес одного и того же существительного будет меняться в зависимости от синтаксической роли. Например, существительное, являющееся подлежащим, имеет большее значение по сравнению с тем же существительным, которое является составной частью предложно-падежной конструкции.

В идеальном варианте, вышеназванные способы автоматического реферирования должны:

- выделить слова, наиболее важные для понимания текста;
- характеризовать распределение самых значимых единиц в структуре текста и в структуре высказываний как составляющих текста.

Таким образом, существует необходимость соотнесения исследований в сфере автоматического реферирования и моделирования поверхностного восприятия и понимания текста пользователем. Имеется в виду наличие условий ограничений на базу знаний адресата.

Существует несколько алгоритмов реферирования, которые эффективно работают с информационно-аналитическими и научно-техническими текстами. В этом плане можно выделить монографию Н.В. Лукашевич «Тезаурусы в задачах информационного поиска». В ней содержатся главы, которые посвящены описанию связности текста и созданию на основе их результатов моделей автоматического реферирования. Наибольший интерес представляют экстрактивные аннотации, которые используют фрагменты исходного текста (система анализа текста) для порождения текста аннотации (вторичного текста). В монографии указываются лингвистические признаки, лежащие в основе определения веса – уровня значимости фрагмента от слова до предложения. Имеется в виду частотность слов, позиция в тексте, название сущности и так далее [Лукашевич 2011, с. 167].

Создание аннотации на основе многих текстов – один из новых и наиболее актуальных вопросов, во всяком случае, в лингвистике. В качестве таких наборов документов могут выступать тексты, кластеры/сюжеты,

организованные в циклы, а может быть, более сложные в лингвистическом плане объекты. В процессе составления таких аннотаций (обзорных рефератов) нужно решить следующие вопросы:

- идентифицировать важные различия между документами;
- бороться с избыточностью информации;
- обеспечить тематическую связность текста.

Работа усложняется тем, что предложения могут быть взяты из разных источников. Проблема аннотации находится на стыке не только разных параграфов данной главы, но и разных глав, предыдущей и следующей. Анализ композиционной структуры текста (анализ риторических отношений в терминах теории риторических структур) является лингвистически значимым. Даже в научных текстах выделяют разные типы композиционных (риторических) структур, которые разнятся по количеству и весу. Это снова побуждает вернуться к проблеме однородности кластера или коллекции не только в плане тематики, но и в плане стиля. Композиционная структура рассматривается в качестве одной из стилевых характеристик текста или коллекции. Часть стилевых характеристик можно предугадать уже на уровне задачи описания исходных свойств выбора коллекции, например, события или череды похожих событий, интервью, аналитика и так далее.

Выводы по главе 1

Учитывая все вышесказанное, можно сделать вывод о том, что КЛ (компьютерная лингвистика) появилась на стыке математики, лингвистики, информатики и искусственного интеллекта. Компьютерная лингвистика существует уже более полувека и известна также под названиями «машинная лингвистика», «автоматическая обработка текстов на естественном языке». В рамках КЛ исследователями и разработчиками предложено множество решений, которые являются достаточно перспективными. Стоит также отметить, что далеко не все эти решения были воплощены в жизнь в виде программных продуктов. Несмотря на этот недостаток, компьютерная

лингвистика показывает вполне реальные результаты. Это видно по различным приложениям по автоматической обработке текстов на естественном языке. Дальнейшее развитие КЛ зависит от разработки новых приложений, различных языковых моделей, в которых пока не решены многие задачи.

Неотъемлемой частью многих систем обработки текстов являются корпуса. Все слова в корпусах имеют исчерпывающие грамматические характеристики, в перечень которых входит часть речи, форма слова, синтаксическая роль. Корпусы представляют собой входные данные, которые служат для обучения в задачах классификации текстов по жанрам и темам. Кроме того, они применяются для обучения синтаксических программ и парсеров, применяемых в процессе снятия омонимии и разрешения анафоры.

На современном этапе наиболее разработаны модели морфологического синтеза и анализа. Необходимо уточнить, что модели синтаксиса пока не являются устойчиво и эффективно функционирующими моделями. Они не смогли достичь этого уровня, несмотря на присутствие большого количества предложенных методов и формализмов. Модели семантики и прагматики исследованы и формализованы еще меньше, но нужно учитывать, что автоматической обработки дискурса уже требует ряд приложений. Решение существующих проблем могут активизировать используемые инструменты КЛ, а также применение корпусов текстов и машинного обучения.

ГЛАВА 2 КОРПУСНАЯ ЛИНГВИСТИКА. ПРОБЛЕМА ИСХОДНОЙ ФОРМЫ СЛОВА

2.1 Корпусы текстов: принципы построения

Понятие «корпус текстов» не может быть раскрыто без объяснения понятия «корпус данных». Корпус данных – это сформированная по заданным правилам выборка данных из области языковой системы. Данная выборка включает феномены, которые подлежат лингвистическому описанию. Корпус данных характеризуется единичным измерением. Он может измеряться только речью, потому что не имеет возможности производить свои составляющие. Тем не менее, это не значит, корпусы данных не могут применяться в целях реконструкции языка как системы. Напротив, это является одной из первостепенных задач лингвистического изучения корпуса. Таким образом, мы видим одно из глобальных противоречий, которое присуще продуктам языковой системы (от звука до текста). Выводы о функциональности языка как системы лингвист вынужден делать по отдельным результатам его деятельности.

Корпус текстов представляет собой разновидность корпуса данных. Его единицы – это тексты или достаточные по объему фрагменты текста, которые включают какие-либо отрывки текстов проблемной области.

Корпусом (от лат. *corpus* – “body”) по большому счету может быть назван любой набор, в котором присутствует более одного текста. Зачастую можно столкнуться с тем, что отдельные тексты применяются для лингвистического и литературного анализа разных типов. Несмотря на сказанное, понятие «корпус в качестве основы для электронной лингвистики» отличается от проверки единичных текстов.

В соответствии с мнением В.В. Рыкова корпусами текстов являются некоторые собрания текстов. Основе данных собраний заложена логическая идея, замысел, который объединяет эти тексты. Логическая идея находит воплощение в правилах, по которым осуществляется объединение текстов, а также программе и алгоритме анализа корпусов, методологии и идеологии,

связанной с этим. В.В. Рыков полагает, что корпус принадлежит к текстам на машинном носителе – четвертой фактуре речи.

В зависимости от поставленной цели выделяют несколько типов корпусов текстов:

1. По форме хранения:

- в звуковой форме;
- письменные;
- смешанные;

2. По языку представления текстов:

- одноязычные;
- многоязычные;

3. По жанровой принадлежности:

- литературные;
- диалектные;
- разговорные;
- публицистические;
- смешанные;

4. По способам доступа:

- свободно доступные;
- коммерческие;
- закрытые;

5. По назначению:

- исследовательские;
- иллюстративные;

6. По динамичности:

- динамические (мониторные);
- статические;

7. По наличию дополнительной информации:

- аннотированные (размеченные);
- неразмеченные [Захаров 2011, с. 89].

В.В. Рыков приводит несколько иную классификацию корпусов текстов:

1. По степени организации и структурированности:

– электронный архив – это тексты на электронном носителе, но их форма представленная на машинном носителе не стандартизирована и не унифицирована;

– электронная библиотека – тексты здесь представлены однородным и стандартизированным образом;

– корпус текстов – форма стандартизирована и унифицирована, тексты предназначены для отражения части лингвистической реальности;

– субкорпус – это некоторая автономная часть корпуса.

2. По хронологическому признаку:

– синхронический;

– мониторный (отслеживает текущее состояние языка);

– диахронический.

3. По индексации:

– простой;

– аннотированный.

4. По языку:

– одноязычный;

– двуязычный;

– многоязычный.

5. По способу применения и использования корпуса:

– исследовательский;

– иллюстративный;

– параллельный.

6. По способу существования корпуса:

– динамический;

– статический [Козлова 2013, с. 43].

Конструирование и использование корпусов осуществляется не по единой методике. Это обусловлено различием языков, традиций, технологических

процессов. Однако вполне возможно выделить главные требования. По В.В. Рыкову в список первостепенных требований включает такие факторы:

1. Тип пользователя – группа, лингвистическое общество, индивид.
2. Логическая задумка, лежащая в основе.
3. Объем данных, которым придется оперировать при составлении корпуса.
4. Насколько реально и необходимо создавать корпус.
5. Какие составляющие применяются – текстовые отрывки, полные тексты, одновременно и то, и другое.
6. Как выполняется процедура отбора текстов, потому что она осуществляется по-разному для разных целей:
 - обследуется речевой материал;
 - сканируются тексты;
 - завершение формирования корпуса.
7. Как представлен корпус в стандартизированном виде на уровне отраслевых стандартов. Корпус в качестве продукта может быть представлен:
 - аннотацией всего текста в целом;
 - унифицированным представлением словесного материала текста.
8. Аннотирование, индексирование словесной информации [Козлова 2013, с. 67].

Вопрос об объеме текстов, отбираемых в корпус, является одним из самых принципиальных. Было бы замечательно, если бы изучаемое явление, даже, если оно является редким для языка, находило отражение в корпусе. Полнота – одно из важных требований, которое предъявляется к составу и структуре корпусов.

И вот здесь появляется противоречие с другим важным принципом – репрезентативностью. Задача создателей корпуса заключается в сборе наибольшего числа текстов, которые имеют отношение к выбранному языковому подмножеству, для изучения которого, собственно, и создается корпус. Получается так, что независимо от специфичности феномена, корпус

не может включать все его реализации. Таким образом, корпус представляет собой определенную выборку из проблемной области, осуществляемую на базе конкретных критериев, которые формулируют исследователи в зависимости от стоящих перед ними задач. Данная выборка обязана отразить разные параметры изучаемого языкового явления в пропорции, что и язык или некоторое языковое подмножество в целом.

В зависимости от отбора корпуса делятся на сбалансированные и мониторные. Первые состоят из текстов, которые представляют различные модули дискурса, включая устные и письменные тексты. Они могут быть совершенно разными по стилям, жанрам и тематике. В процессе создания корпуса исследователи устанавливают пропорции, в которых будут представлены разные по типу тексты. Данные корпуса характеризуются фиксированным объемом. Их пополнение выполняется только после проведения тщательного отбора новых текстов.

Мониторные корпуса характеризуются постоянным пополнением. В них регулярно добавляются новые тексты на данном языке без соблюдения баланса стилей, модусов и жанров. Разработчики мониторных полагают, что статистическая обоснованность данных достигается за счет объема, который исчисляется в миллиардах слов.

Одним из важных параметров корпуса считается объем. Объемы первых корпусов достигали миллиона слов, точнее, текстоформ и словоупотреблений. Объем современных корпусов составляет сотни миллионов и миллиарды слов. Например, объем Национального корпуса русского языка исчисляется 140 миллионами слов, объем Bank of English (корпус английского языка) насчитывает более 2,5 миллиардов слов.

Решение лингвистических задач требует дополнительной лингвистической и металингвистической информации (разметка/аннотация), которая содержалась бы в текстах и отдельных языковых единицах внутри текстов. Кроме метаразметки, которая отражает разную экстралингвистическую информацию о тексте, включая, имя автора,

жанровую принадлежность, название, современные корпуса содержат разметку, соответствующую различным уровням лингвистического описания. Это может быть морфологическая, фонетическая, синтаксическая или другая разметка.

Для решения различных лингвистических задач необходимо, чтобы тексты и отдельные языковые единицы внутри текстов содержали дополнительную лингвистическую и металингвистическую информацию - разметку (аннотацию). В современных корпусах помимо метаразметки (отражающей различную экстралингвистическую информацию о тексте, включая его название, автора, жанровую принадлежность и т.п., подробнее см. разметка корпуса), содержится разметка, соответствующая различным уровням лингвистического описания, - морфологическая, синтаксическая, фонетическая и др.

2.2 Текст. Общие положения

Основные свойства текста, имеющие важное значение для его изучения в контексте речевой коммуникации (порождения и восприятия):

- развернутость (последовательность знаковых единиц);
- связанность и цельность;
- отдельнооформленность.

Развернутость можно соотнести с вопросом об уровне и размерности иерархии единицы – текста. Структурными элементами такой единицы являются синтагмы, слова, фразы, сверхфразовые единства. Текст понимается как главная конструктивная языковая единица, основной лингвистический контекст, реализующий единицы более низкого уровня, например, слова, коллокации, синтагмы, фразы, композиционные фрагменты и сверхфразовое единство. Базовость и конструктивность может казаться очевидной, однако стоит учитывать авторитетную формулировку, которую дает В.К Касевич. В ней говорится о том, что текст, являясь целостной единицей, показывает в отношении своих структурных элементов свойство

неаддитивности в плане неполной выводимости характеристик текста из признаков его составляющих. В первую очередь, значение, которое передается текстом, несводимо к сумме значений компонентов. Отдельнооформленность характеризуется наличием сигналов начала и конца, а также представлением о фреймах. Под фреймами понимается знание носителей языка о структуре текстов различных по функциональности текстов (коммуникативной и текстовой компетенции). Выделяется внутренняя и внешняя связность. В соответствии с исследованиями И. Беллелрта связный текст определяется как последовательность высказываний S_1, \dots, S_n с семантической интерпретацией высказываний S_i (при $2 < i < n$), зависящей от интерпретации высказываний в последовательности S_1, \dots, S_{i-1} .

Таким образом, можно сделать вывод о наличии взаимосвязанности и взаимообусловленности структурных компонентов текста, которые и являются основой его связности цельности. Связность бывает пространственной при контактно расположенных структурных составляющих, логической и ассоциативной. Цельность и связность – важные характеристики текста, но формализуются они сложно. Связность (когерентность) имеет отношение к структурной организации. Связность подразделяется на тематическую (смысловую) и синтаксическую. В перечень формализуемых средств смысловой связности входят:

- связующие слова – слова с причинно-следственными, темпоральными следственными значениями, союзы;
- механизмы кореференции и референции – повторные номинации, например, повторяющиеся в тексте слова.

Синтаксическая связность текста, а также связность высказываний как структурных элементов выражается, главным образом, посредством семантико-синтаксической структурированности данных единиц. Связность текста определяется разными исследователями по-разному. В последних трудах когезия и когерентность разделяется все чаще. Когезию определяют

как связь компонентов, при которых одни элементы интерпретируются в зависимости от других. Когерентность соотносится с прагматической стороной. Она опирается на базу знаний адресата, выводя за границы текста в коммуникативную ситуацию. Когерентность в большой степени связана с реализацией ожиданий адресата и презумпцией осмысленности. Стоит отметить, что в реальных моделях понимания текста носителем языка невозможно четко отделить друг от друга эти два вида связности²⁹. Цельность и связность в процедурах речевой деятельности реализуются посредством механизма контекстной предсказуемости. В таком случае, совершенно естественно предположить, что при выборе произвольной точки в границах текста свойства ее правого, непосредственного соседа не будут случайными. Таким образом, текст можно описать как взаимодействие между более сложными единицами, метафорически понимаемыми кривыми силами связей между словами. При этом определенные позиции будут сильно воздействовать на проявления справа, другие достаточно слабо смогут предсказывать своих непосредственных соседей. Множественность кривых определяют множеством параметров и признаков, посредством которых производится связывание. Происхождение природы предсказуемостей и связей может:

- 1) быть связано с семантической и лексической несочетаемостью или сочетаемостью;
- 2) определяться синтаксическими правилами;
- 3) соотноситься с информационной значимостью;
- 4) задаваться честной или общей коммуникативной ситуацией.

Характер предсказуемости усложняется, если предсказание позиций осуществляется не свойствами ближайшего предшествующего элемента, а на основании знания о смысловой целостности или связности, которыми обладает слушающий. Силы связей между словами достаточно полно описаны в математических сетевых моделях. Тем не менее, у таких моделей есть недостаток. У них есть ограничение естественного характера в виде уже

упомянутого множества связей разнотипных по своей лингвистической природе, которые, в большинстве своем, изучены слабо. В этом случае остается надеяться только на то, что в ближайшей перспективе произойдет существенное расширение возможности данного моделирования. Новейшее моделирование должно учитывать варьирование типов контекстов и единиц, а также учитывать различные параметры и признаки. Подобную работу можно выполнить посредством подключения лингвистически сбалансированных и специально выбранных коллекций, когда каждая задача соответствует своей коллекции или набору коллекций [Большакова 2011, с.45].

В процессе коммуникативных актов люди осуществляют непрерывное планирование или программирование своей речи, а также восприятия, при этом выполняются необходимые переключения и регулировки. В этом плане следует учитывать, что каждая последующая единица сверяется и согласовывается с уже произнесенным или воспринятым на данный момент. Оценка прогнозирования дается в прикладном направлении, которое обозначается английским словом «*readability*». Данный термин объясняет, главным образом, не читабельность, а «*понимабельность*» текста. Это оценка уровня правильности извлечения смысла, даже, если имеет место быть беглое чтение, а также искажения. Минимальное окно анализа (проверки) приравнивается к одной единице, например, слову или высказыванию. Тем не менее, минимальное необходимое прогнозирование преобладает в статистике и является для нее типичным. Максимальное прогнозирование определяют текстом и общей коммуникативной ситуацией.

В соответствии с традициями когнитивных теорий текст рассматривается как реализация определенного фрейма. Основоположником данного подхода является Марвин Минский. По его определению фрейм – это структура данных, которая предназначается для представления конкретной типовой ситуации. В качестве примера можно привести фреймы деловой, бытовой, научной коммуникативной ситуации, которые дают возможность делать

прогноз развития событий в данной ситуации, включая восприятие разных по функциональности текстов. Знание слушающим конкретного фрейма можно соотнести со знанием адресатом смысла и смысловой связности текста, а текст в данном случае выступает в качестве реализации этого фрейма. Достаточно сильно противопоставление типов целей и исследовательских процедур изучения текстов:

- понимание и интерпретация текста людьми – этим занимаются в русле когнитивистского и традиционного подходов, об этом говорится в работах зарубежных авторов, которые частично рассматриваются ниже;
- понимание и интерпретация текстов в духе прикладных задач, например, автоматическое понимание текста, автоматическое извлечение информации из него, машинный перевод, автоматическое реферирование и прочее [Клышинский 2011, с. 108].

Разница в приведенных выше подходах нуждается в применении разных носителей языка, которые помещаются в центр исследования. Автомат выступает в качестве искусственного носителя языка в ходе прикладных экспериментов. Следствие такой разницы – это степень вовлеченности базы знаний, которая дает возможность построить прогноз дальнейшего развития ситуации на базе знания видов коммуникативных ситуаций (внелингвистических данных).

Совершенно ясно, что автомату затруднительно формировать некоторую макроструктуру текста, которая представляет собой результат функционирования в процедурах интерпретации и понимания не только структурных элементов, но также и выводных, фоновых знаний. Можно предположить, что степень вовлеченности выводных и фоновых знаний находится в прямой зависимости от типа фрейма и знаний коммуниканта об этом фрейме.

2.3 Электронный текст как основа корпуса

Электронные тексты, по сравнению с печатными, характеризуются принципиально новыми свойствами. Это обусловлено современным подходом к их хранению и распространению. Электронные тексты открывают перед лингвистами широкие перспективы. Это касается обработки объемных массивов информации с учетом дополнительной классификации и новых подходов к решению традиционных проблем.

В лингвистике компьютерные технологии используются давно. Р. Буса в конце 40-х годов прошлого века начинает работать над составлением конкорданса произведений Фомы Аквинского. Данная работа длилась несколько десятилетий. На современном этапе можно назвать много проектов, которые применяются в различных лингвистических областях. Эти проекты функционируют посредством ЭВМ. В качестве примера можно привести автоматическую подготовку к печати и генерированию индексов, содержательный и стилистический анализ текстов с помощью ПК, электронные издания, распознавание и генерирование речи. Данная исследовательская область собирательно обозначается как «гуманитарная информатика» (Humanities Computing).

Электронные тексты представляют собой основу любой деятельности, которая связана с компьютерной лингвистикой и филологией. Изначально электронный текст представлял собой одну из ступеней создания печатного текста. Постепенно электронные тексты трансформировались в конечный самостоятельный продукт, например, специализированные ИС, лексические БД, электронные издания. Для превращения текста в электронную форму (дигитализация) нужно выполнить перенос информации о каждом знаке, а также метаинформации на электронный носитель. Метаинформация в текстах электронного вида – это дополнительные данные, которые содержит текст. Сюда входят подчеркивания, жирный шрифт, разделение на главы, строфы, стихи и другая информация, вплоть до названия произведения, имени автора, дополнительных лингвистических данных. Процесс

(ре)дигитализации печатного текста включает оцифровку печатного оригинала и снабжение его дополнительной метаинформацией.

В процессе ввода текста может частично выполняться markup – разметка, внесение метаданных. Эти действия выполняются напрямую или опосредованно, в автоматическом режиме, основываясь на текстовых нюансах. Большинство текстов с усложненной структурой требуют дополнительной обработки профессиональным филологом. Специалист вносит замечания и дополнения, составляет критический аппарат или связывает гиперссылками отдельные фрагменты. Чтобы опубликовать электронный текст, его необходимо адаптировать под программную оболочку, посредством которой выполняется его репрезентация, навигация и поиск.

Для лингвистов, которые работают с электронным текстом, является серьезной проблемой его долгосрочное сохранение. Большая часть современных электронных изданий тесно связаны с программами, которые предназначены для репрезентации и поиска в них, и и напрямую зависят от срока службы данных программ. Text Encoding Initiative (TEI, wvuw.tei-c.org) – система филологической разметки текстов, разработанная независимой группой лингвистов, позволяет осуществить кодировку, которая обладает большей независимостью от программной оболочки и ОС. Данная кодировка по своей долговечности сопоставима с печатным текстом. Text Encoding Initiative базируется на Extensible Markup Language (XML) – универсальный язык разметки, который является международным стандартом.

Она оснащена возможностями выделять в тексте особенности, которые специфичны для лирики, прозы, драмы, может выполнять разметку словарей, транскрипции речи, терминологических БД. В разметке реализуются механизмы, которые необходимы для создания сложных гипертекстовых ссылок, кодирования любых символов. Text Encoding Initiative и XML основываются на концепции семантической разметки, отделенной от типографической информации, описывающей ее представление.

Репрезентация текста выполняется с использованием данных, которые представлены собственным языком. Эти данные необходимы для описания форматирования и учитывают особенности носителя, на который выводится информация.

Достоинство Text Encoding Initiative заключается в возможности использовать единый стандарт для долговременного сохранения информации. Данный стандарт, используемый филологами при работе с электронными текстами, больше 10 лет развивается и тестируется представителями группы независимых лингвистов из разных стран мира.

2.4 Лемматизация и нормализация. Проблема исходной (словарной) формы слова

Процесс нормализации, реализованный в грамматическом словаре, позволяет убрать из исходного текста грамматическую информацию (падежи, числа, глагольные виды и времена, залоги причастий, род и так далее), оставляя смысловую составляющую.

Нормализация текста не использует стемминг, поэтому она лишена недостатков потери релевантности из-за особенностей русского словоизменения.

Лемматизатор по своим результатам стоит намного ближе к нормализатору. Однако он применяет упрощенный анализ слов, не учитывая контекст. Это приводит к неоднозначностям при определении части речи. Например, лемматизация слов в словосочетании мы роём яму даст для второго слова два варианта лемматизации: существительное рой и глагол рыть. Эта неоднозначность не может быть разрешена без привлечения морфологического анализатора.

Лемматизация (англ. lemmatization) - это метод морфологического анализа, который сводится к приведению словоформы к ее первоначальной словарной форме (лемме).

В любом естественном языке существует некоторый процент слов, которые могут давать неоднозначные результаты в процессе лемматизации, например, словоформа "рой" может в итоге быть приведена к двум леммам - рыть (глагол) и рой (существительное).

Исходная форма слова- термин, используемый иногда в лингвостатистике в значении "основной формы слова". Является не совсем правильным, ибо под исходной формой слова в языкознании понимается простая (или производная) основа, являющаяся источником дальнейшего словообразования и формообразования.

При лемматизации мы берем слово и получаем для него лемму - нормальную (начальную, словарную) форму: словарем - словарь. Для русского языка это означает, что существительное в любой грамматической форме приводится к форме именительного падежа. Для подавляющего большинства русских существительных нормальная форма также означает единственное число, хотя для некоторых существительных, не употребляющихся в единственном числе, это может быть и форма множественного числа: санками - санки. Русский лексикон сформирован таким образом, что названия словарных статей всегда соответствуют начальной форме существительного. Поэтому лемматизация может быть побочным продуктом морфологического анализа. Однако морфологический анализ сам по себе достаточно тяжел и требует наличия очень большой словарной базы.

2.5 Правила приведения словоформ к исходной (словарной) форме слова

Задача морфологического анализа заключается в обеспечении определения нормальной формы слова, от которой образовалась словоформа, и перечня параметров, являющихся частью этой словоформы. Данный анализ способствует ориентированию в будущем только на основе нормальных

форм, а не на всех возможных словоформах, применять параметры, например, которые используются для проверки согласования слов.

Различные формы слов или словоформы были придуманы людьми для того, чтобы выделить явные различия в употреблении слов. Независимо от вида словоформы, с ее помощью обозначается одинаковое понятие. При рассмотрении понятия общепринято считать, что используется его нормальная форма, то есть, одна из словоформ, которая выделена для обозначения понятия. Например, возьмем столов «*стул*». Для него есть различные формы: *стулом*, *стула*, *стулу* и так далее. Каждая форма соответствует ряду параметров и характеристик, например, падежу, роду, числу, который характеризуют данную словоформу. Кроме того, каждое слово соответствует определенной части речи, которая показывает род оперируемого понятия. В конкретном месте речи стоит слово с заданной частью речи в определенной форме, однако в ходе машинной обработки данные интуитивные рассуждения необходимо формализовать. Стоит отметить также, что подобное разнообразие создает проблемы при проведении анализа текста. Проблема заключается в обработке всех словоформ вместо обработки единственного слова. Избежать проблемной ситуации позволяют этапы морфологического анализа и синтеза. Задача морфологического анализа заключается в определении нормальной формы по словоформе. Может выясниться, что одна словоформа сопоставима с несколькими парами. Задача морфологического синтеза обратно противоположна анализу. В его рамках словоформа определяется по нормальной форме и перечню параметров. Если давать более формальные определения, то под нормальной формой слова понимается форма слова (строка), которой принято обозначать понятие, связанное с этим словом. Принято считать, что формы слова образуются от нормальной формы, однако есть случаи, когда данная связь не прослеживается. В качестве примера можно привести слова «идти – шел». Таким образом, нормальной формой следует считать одну из форм слова, которая выделена по языковой

традиции. Под словоформой понимают форму слова (строки), которая связана с нормальной формой и указывает на особенности использования слова. Будем характеризовать словоформу пятеркой – нормальной формой, от которой была образована словоформа, частью речи, строкой словоформы. Нам необходима часть речи нормальной формы, потому что, например, удобно, когда деепричастие считается формой глагола, а не выводится в отдельную форму. В целом, перечень основных частей речи состоялся, хотя споры о составе служебных частей речи в исследовательской среде ведутся до сих пор. В процессе использования определенного морфологического словаря самое важное заключается в выборе списка, так как в последующем его изменение приведет к удорожанию операций. Выполнение практических задач удобно с участием любой из имеющихся в наличии логически обоснованных систем деления слов на части речи.

2.6 Программное обеспечение

В рамках нашей дипломной работы была разработана программа «лемматизатор», которую мы испытывали на основе корпуса текстов зарубежных научно-фантастических произведений. Данная программа проводит лемматизацию всего корпуса текста, а затем выводит частотный список всех лемм, благодаря которому мы можем отследить какие леммы являются самыми частотными и употребляемыми, а какие почти не используются. Для более точной лемматизации в программе используются базы данных с правилами образования словоформ на английском языке. К сожалению, создать идеально работающий лемматизатор, это очень сложная задача. Даже прописав многие правила словообразования, нет гарантии, что весь текст будет лемматизирован безошибочно. Существует множество исключений, и написанные правила просто могут неправильно увидеть слово и привести к неправильной словоформе.

Запуск и работа с программой:

1. На рабочем столе находим файлы с дипломной работой, затем находим папку с программой лемматизатора и базами данных для проведения лемматизации (рис. 2.1)

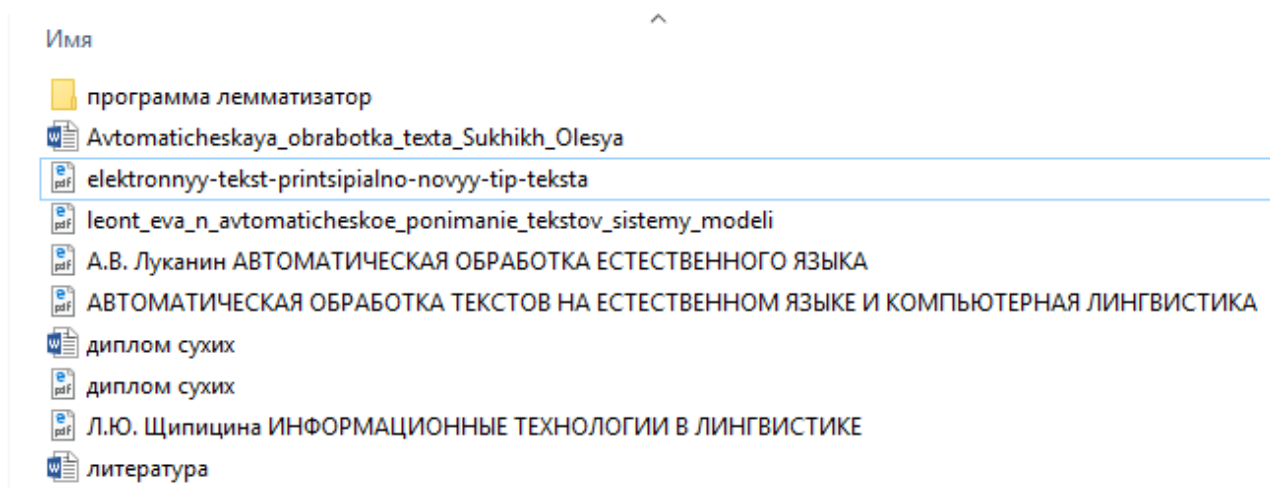


Рисунок 2.1 — Папка с дипломной работой

2. В папке с программой находим файл «lemmatizator.py» (рис. 2.2)

Имя	Дата изменения	Тип	Размер
data_inflections	12.06.2018 11:25	Data Base File	0 КБ
data_postfixes	10.06.2018 22:17	Data Base File	2 КБ
data_prefixes	10.06.2018 22:17	Data Base File	2 КБ
korpus	12.06.2018 11:20	Текстовый докум	641 КБ
lemmatizator	12.06.2018 11:20	Python File	3 КБ

Рисунок 2.2 — Папка с программой лемматизатора

3. Открываем программу, используя IDLE (рис. 2.3)

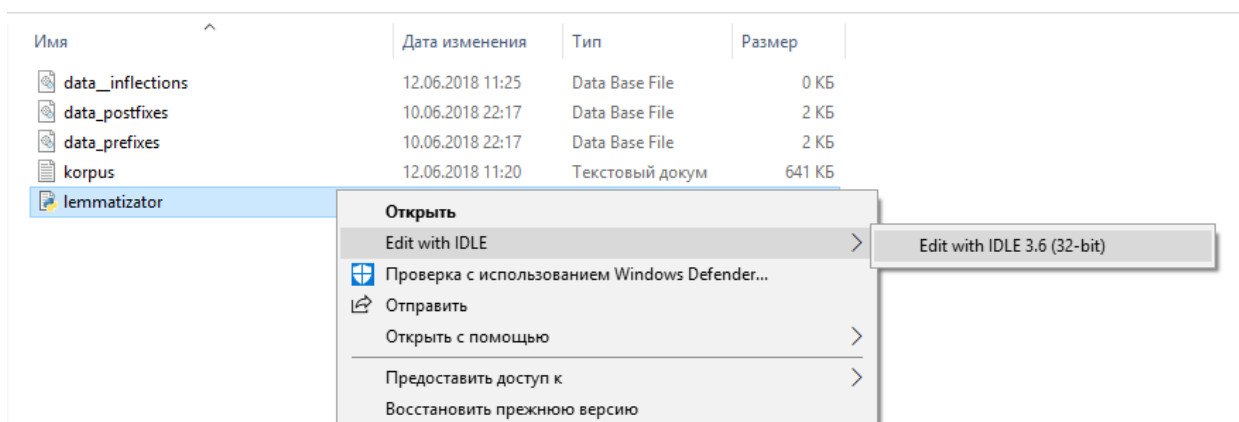


Рисунок 2.3 — Открытие программы

4. Запускаем программу кнопкой F5 или с помощью командной строки (рис. 2.4)

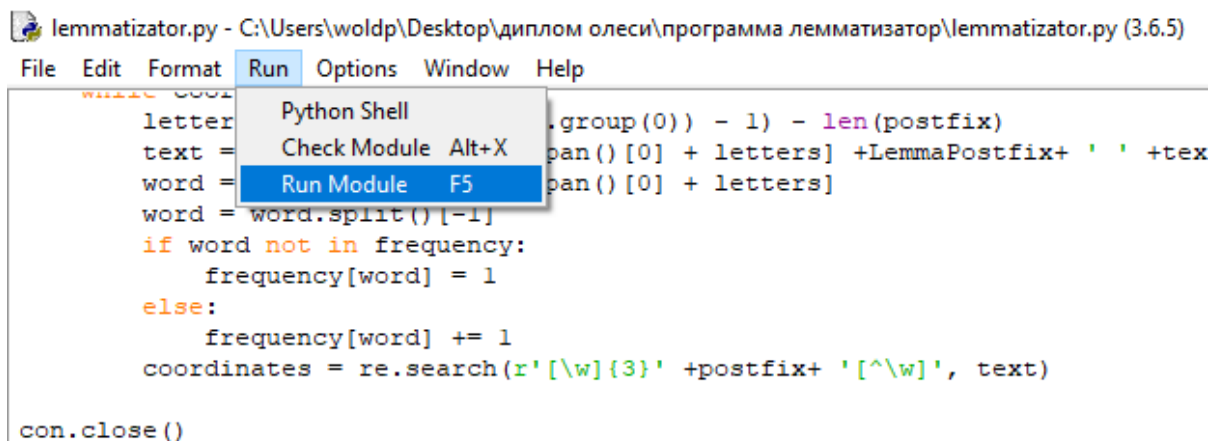


Рисунок 2.4 — Запуск лемматизатора

5. Вводим название документа, который мы хотим лемматизировать. В нашем случае вводить название документа не нужно. При запуске, программа автоматически проводит лемматизацию нашего корпуса текстов (рис. 2.5)

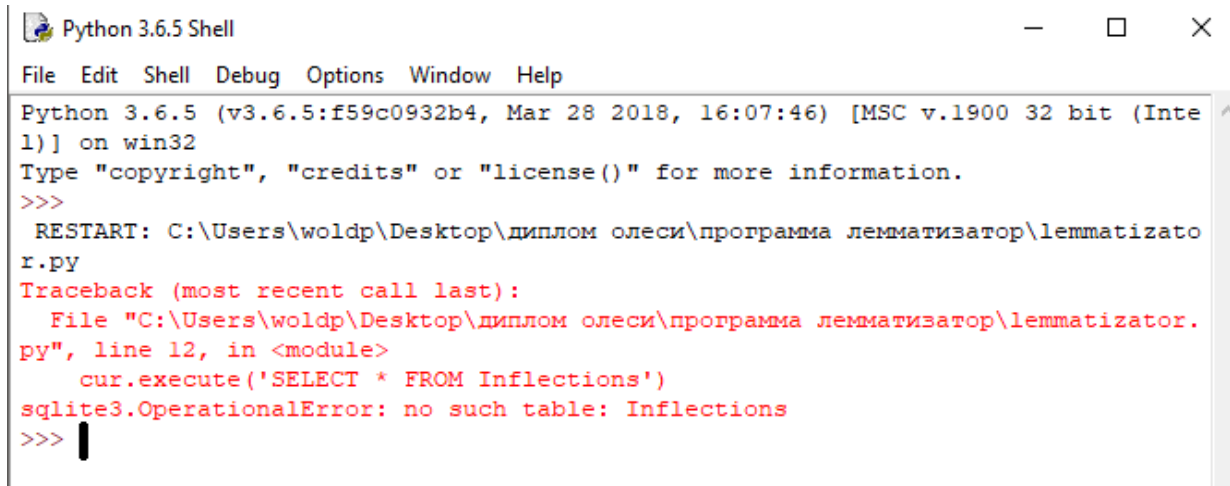
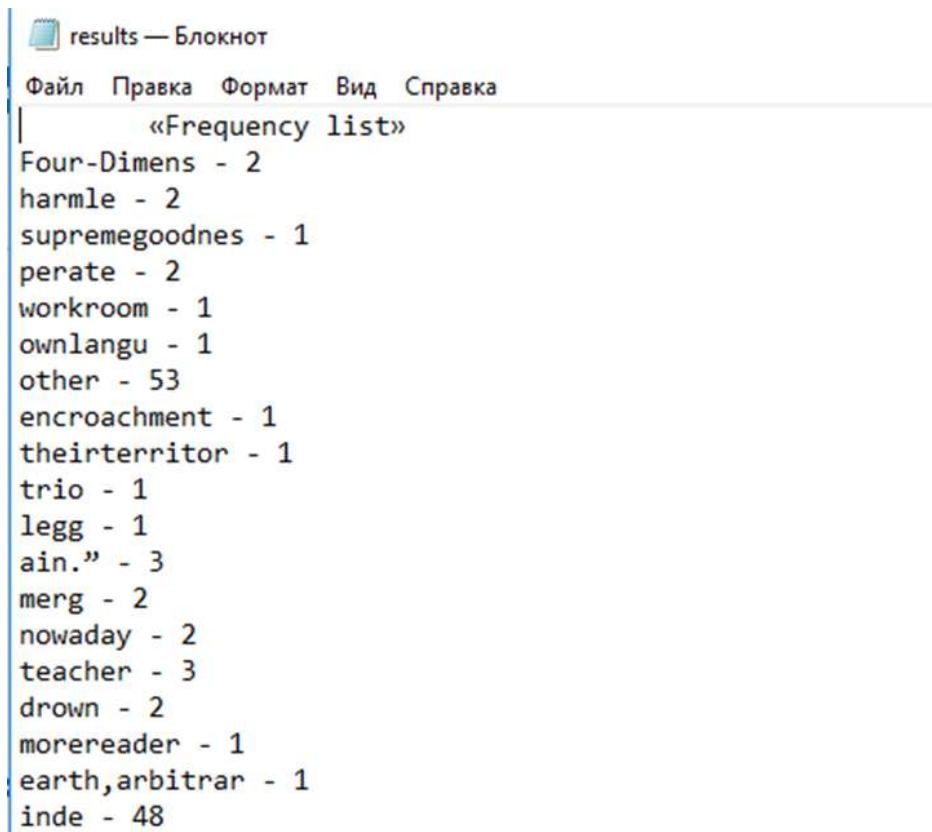


Рисунок 2.5 — Название документа

6. Обработка нашего корпуса текста занимает примерно 10-15 минут (рис. 2.6)



```
results — Блокнот
Файл  Правка  Формат  Вид  Справка
|
|     «Frequency list»
|
| Four-Dimens - 2
| harmle - 2
| supremegoodnes - 1
| perate - 2
| workroom - 1
| ownlangu - 1
| other - 53
| encroachment - 1
| theirterritor - 1
| trio - 1
| legg - 1
| ain.” - 3
| merg - 2
| nowadays - 2
| teacher - 3
| drown - 2
| morereader - 1
| earth,arbitrar - 1
| inde - 48
```

Рисунок 2.6 — Результаты

В результате проведённой лемматизации мы получаем частотный список всех лемм корпуса текстов английских научно-фантастических произведений и лемматизированный корпус. По результатам можно сказать, что есть очевидные недочёты и проблема лемматизации не так проста в решении.

Выводы по главе 2

Любой сборник, состоящий больше чем из одного текста может назваться корпусом (от лат. corpus – “body”). Нередко отдельные тексты берутся для использования в различных видах литературного и лингвистического анализов. Однако, понятие корпуса, взятого за основу для электронной лингвистики, достаточно сильно отличается от проверки единичных текстов.

Объём является очень важным параметром корпуса. Если практически все первые корпуса достигали миллиона слов (точнее, словоупотреблений или текстоформ), то объем современных корпусов насчитывает сотни миллионов (например, объем Национальный Корпус Русского языка на данный момент составляет около 140 млн. слов) или миллиардов слов (например, объем

корпуса английского языка «Bank of English» превышает 2,5 миллиарда слов).

Лемматизация (англ. lemmatization) - это метод морфологического анализа, который сводится к приведению словоформы к ее первоначальной словарной форме (лемме).

Лемматизатор применяет упрощенный анализ слов, не учитывая контекст. Это приводит к многим неоднозначностям, когда мы пытаемся определить часть речи. Например, лемматизация слов в словосочетании «*мы роём яму*» даст для второго слова не один, а два варианта лемматизации: существительное «рой» и глагол «рыть». Эта неоднозначность при определении части речи не может быть решена без использования морфологического анализатора.

Созданная нами программа лемматизатора была апробирована на основе корпуса зарубежных научно-фантастических произведений. По результатам можно сказать, что наш лемматизатор работает корректно, но есть ещё недочёты, для исправления которых потребуется много времени.

ЗАКЛЮЧЕНИЕ

Компьютерная лингвистика (КЛ) появилась на стыке таких наук, как лингвистика, математика, информатика и искусственный интеллект. В данном направлении разрабатывается множество программ, методов автоматической обработки текстов. Несмотря на достаточно длительное существование компьютерной лингвистики, на данном этапе её развития ещё многие идеи не нашли своё применение в программных продуктах. Сами же инструменты данной науки могут существенно помочь в решении многих проблем при создании программ.

В процессе написания данной дипломной работы были теоретически осмыслены такие понятия как корпус текста, текст, компьютерная лингвистика, лемматизатор, лемматизация и другие.

В данной дипломной работе мы создали автоматизированный лемматизатор, который может анализировать и приводить к начальной словоформе целый корпус текстов на английском языке.

Лемматизация широко используется в алгоритмах поисковых систем. Так, она позволяет найти большее количество результатов, а не только результаты по запросу слова только в той форме, в которой оно было введено. Так же лемматизация применяется при проверке уникальности текста, веб-разработке, программировании и составлении семантического ядра.

Для более глубокого понимания проблемы лемматизации нами была написана программа, которая позволила провести анализ двух английских научно-фантастических произведений, которые были собраны в единый корпус. Для работы лемматизатора понадобилось создать базы данных, в которых прописаны правила словообразования в английском языке, благодаря им программа может лемматизировать весь корпус.

Как результат работы программы, мы можем увидеть частотный список лемм, которые употребляются в используемом корпусе. А также лемматизированный текст.

Данную программу можно использовать для любого корпуса текстов на английском языке. Чтобы провести лемматизацию корпуса, например, на русском языке, нужно создать базы данных с правилами словообразования в русском языке.

Важно отметить, что лемматизатор не работает идеально. Это происходит из-за неполноты баз данных с правилами словообразования. Для того, чтобы лемматизация корпуса выдавала верные результаты, нужно проделать очень трудоемкую работу, которая может занять не один месяц. Для этого нужно учесть все правила словообразования и все исключения. Например, при лемматизации на английском языке в базу данных нужно вносить все сведения о неправильных глаголах, для существительных, прилагательных и глаголов создать разные базы данных, в которых будут прописаны правила словообразования именно этих частей речи. Приводя слово к лемме иногда возникает проблема определения того, к какой части речи она относится, к существительному или же глаголу. Абсолютного решения данной проблемы пока не существует, так как для этого нужно учитывать все особенности словообразования.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ, А.А. Носков, О.В. Пескова, Е.В. Ягунова. – М.: МИЭМ, 2011. – 272 с.
2. Апресян, Ю.Д. Идеи и методы современной структурной лингвистики / Ю.Д. Апресян. – М.: Просвещение, 1966. – 301 с.
3. Баранов, А. Н. Введение в прикладную лингвистику: учебное пособие / А.Н. Баранов. – М.: Эдиториал УРСС, 2001. – 360 с.
4. Белоногов, Г.Г. Компьютерная лингвистика и перспективные информационные технологии / Г.Г. Белоногов. – М.: Русский мир, 2004. – 248 с. – ISBN 5-85810-077-9.
5. Беляева, Л.Н. Лингвистические автоматы в современных гуманитарных технологиях: учеб. пособие / Л.Н. Беляева. – СПб.: Книжный Дом, 2007. – 192 с.
6. Беляева, Л.Н. Автоматический (машинный) перевод / Л.Н. Беляева, М.И. Откупщикова // Прикладное языкознание: учебник. – СПб.: Изд-во С.-Петербург. ун-та, 1996. – С. 360–388.
7. Захаров, В.П. Корпусная лингвистика: учебник для студентов гуманитарных вузов. / В.П. Захаров, С.Ю. Богданова. – Иркутск. ИГЛУ, 2011. – 161 с.
8. Захаров, В.П. Информационно-поисковые системы: учеб. – метод, пособие / В.П. Захаров. – СПб.: СПбГУ, 2005. – 48 с.
9. Зубов, А.В. Информационные технологии в лингвистике: учеб. пособие / А.В. Зубов, И.И. Зубова. – М.: Академия, 2004. – 208 с.
10. Козлова, Н. В. Лингвистические корпуса: определение основных понятий и типология / Н. В. Козлова // Вестник НГУ. Сер. Лингвистика. – Новосибирск, 2013. – 95 с.
11. Клышинский, Э.С. Начальные этапы анализа текста / Э.С. Клышинский // Автоматическая обработка текстов на естественном

языке и компьютерная лингвистика: учеб. пособие. – М.: МИЭМ, 2011. – С. 106–140.

12. Луканин, А.В. Автоматическая обработка естественного языка: учебное пособие / А.В. Луканин. – Челябинск: Издательский центр ЮУрГУ, 2011. – 70 с.

13. Луканин, А.В. Инструментарий прикладного лингвиста / А.В. Луканин // Современные направления прикладной лингвистики: материалы I Студенческой научно-практической конференции. – Челябинск: Международный студенческий научный вестник, 2008. – 34 с.

14. Лукашевич, Н.В. Тезаурусы в задачах информационного поиска / Н.В. Лукашевич. – М.: Издательство МГУ, 2011. – 512 с.

15. Марчук, Ю.Н. Компьютерная лингвистика: учеб. пособие / Ю.Н. Марчук. – М.: АСТ Восток – Запад, 2007. – 317 с.

16. Мельчук, Н.А. Автоматический синтез / Н.А. Мельчук // Большая советская энциклопедия. – М.: Советская энциклопедия. 1969 – 1978. – URL: <http://dic.academic.ru/dic.nsf/bse/61319/Автоматический> (дата обращения: 28.04.2018).

17. Нагель, О. В. Корпусная лингвистика и ее использование в компьютеризированном языковом обучении / О. В. Нагель // Язык и культура. – М., 2008. – №4. – 118 с.

18. Пиотровский, Р.Г. Лингвистический автомат (в исследовании и непрерывном обучении) / Р.Г. Пиотровский. — СПб.: Изд-во РГПУ им. А.И. Герцена, 2008. – 256 с.

19. Севбо, И.П. Структура связного текста и автоматизация реферирования / И.П. Севбо. – М.: Наука, 1969. – с. 136

20. Селезнев, К. Лингвистика и обработка текстов / К. Селезнев, А. Владимиров // Открытые системы. – 2013. – № 04. – С. 46–49.

21. Щипицина, Л.Ю. Информационные технологии в лингвистике: учеб. пособие / Л.Ю. Щипицина. – М.: ФЛИНТА: Наука, 2013. – 128 с.

22. Kuznetsov, S.O. Fitting Pattern Structures to Knowledge Discovery in Big Data / S.O. Kuznetsov. – ICFCA 2013. – 318 p.
23. Manning, C.D. Foundations of Statistical Natural Processing / C.D. Manning, H. Schuetze. – MIT Press, 1999. – 620 p.
24. Mirkin, B. Core Concepts in Data Analysis: Summarization, Correlation and Visualisation / B. Mirkin. – Springer, 2011. – 390 p.
25. Torrejón, E. Controlled Translation: A New Teaching Scenario Tailor-made for the Translation Industry / E. Torrejón, C. Rico // 6th EAMT Workshop Teaching Machine Translation. – Manchester, 2002. – 254 p.

Словари и энциклопедии

26. БЭС – Языкознание. Большой энциклопедический словарь / гл. ред. В. Н. Ярцева. – 2-е изд. – М.: Большая Российская энциклопедия, 1998. – 685 с.
27. ЛЭС – Лингвистический энциклопедический словарь / под ред. В.Н. Ярцевой. – М.: Сов. Энциклопедия, 1990. – 685 с.
28. Нелюбин, Л.Л. Толковый переводоведческий словарь. / Л.Л. Нелюбин. – 3-е изд., перераб. – М.: Флинта: Наука, 2003. – 320 с.
29. Толдова, С.Ю. Корпусная лингвистика / С.Ю. Толдова., А.В. Архипов, Е.А. Логинова, Д.П. Попова. – М.: Филология, 2011. – URL: www.lomonosov-fund.ru/enc/ra/encyclopedia:01210:article (дата обращения: 28.04.2018).
30. Толдова, С.Ю. Автоматический морфологический анализ. / С.Ю. Толдова, А.А. Бонч-Осмоловская. – М.: Филология, 2011. – URL: www.lomonosov-fund.ru/enc/ru/encyclopedia:0127430 (дата обращения: 28.04.2018).
31. Что такое корпусная лингвистика? – URL: <http://fb.ru/article/325516/chto-takoe-korpusnaya-lingvistika.htm> (дата обращения: 25.10.2017). – Загл. с экрана.