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/

ЛИНГВИСТИЧЕСКИ ОБУСЛОВЛЕННЫЙ АНАЛИЗ ТОНАЛЬНОСТИ НАУЧНЫХ ТЕКСТОВ НА АНГЛИЙСКОМ ЯЗЫКЕ

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
ЮУрГУ – 45.03.02.2018.264.ВКР

Руководитель, к.филол.н., доцент

_____ /О.И. Бабина/
«____» _____ 2018 г.

Автор
студент группы ЛМ-431
_____ /Н.Э. Лебедев /
«____» _____ 2018 г.

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

Работа защищена с оценкой

_____
«____» _____ 2018 г.

Челябинск
2018

Оглавление

ВВЕДЕНИЕ

Объектом является лексические единицы корпуса научных рецензий на английском языке.

Предмет- семантические, структурные и функциональные особенности лексических единиц, отражающие тональность текста.

Цели:

1) Выявить семантические и структурные особенности, влияющие на тональность текста

2) Провести лингвистически обусловленный анализ на разных уровнях языка

Задачи:

1) Изучить влияние тональности слов по отношению к их принадлежности к группе ключевых слов

2) Создать удобную структурированную систему для работы с корпусом текстов

3) Найти необходимый алгоритм по работе с корпусом текстов

4) Разработать легкий и быстрый доступ к необходимой для исследования и обучения информации

Актуальность:

1) Как подтверждают результаты исследований в данной области, а так же проведенный нами анализ корпуса текстов проблема остается до конца не изучена.

2) В теории и практике имеются предпосылки, чтобы говорить о не достаточном развитии компьютерных программ работающих с тональностью текста.

Новизна:

Определены лексические единицы-маркеры, влияющие на тональность текста.

Выявлены семантические и морфологические особенности.

Составлена классификация маркеров, влияющих на определение тональности.

Впервые маркеры тональности выявляются на материале научных текстов.

Теоретическая значимость определятся тем, что проведенное нами исследование вносит определенный вклад в развитие корпусной лингвистике в плане изучения отдельных единиц языка, содержащих в себе определенные характеристики.

Практическая значимость: Собранный материал и результаты, полученные в ходе исследования характеристик слов-маркеров, являются вкладом в развитие и разработку методов извлечения слов-маркеров и описания их характеристик.

Теоретико-методологической базой работы являются труды, посвященные стилистике и лексикологии английского языка, таких авторов как Антрушина Г.Б. и Арнольд И.В., а также труды, посвященные семантическому полю и проблемам синонимии Шкуропацкой М.Г.

В ходе работы использовались методы сплошной выборки, метод компонентного анализа, метод структурного анализа, дистрибутивный анализ.

В данной работе будет рассмотрена проблема извлечения информации из научных источников на английском языке. На сегодняшний день данный вопрос является актуальным и остается не изученным до конца. На данный момент проблема также является актуальной в связи с ростом технологий. Появляется много программ для извлечения информации из текстов. Вопрос также состоит в том, как правильно происходит процесс извлечения информации и нахождения ключевых слов в корпусе текстов. Различные научные разработки предлагают разные решения данной проблемы. Они состоят в разработке специального программного обеспечения для автоматического извлечения слов-маркеров и информации.

Также перед тем как приступать к извлечению слов-маркеров необходимо произвести структуризацию корпуса текстов или отдельного документа. Это полезно делать для создания методов и алгоритмов извлечения информации, а значит и правильной работы программного обеспечения, направленного на это. Современный этап научного познания характеризуется значительными темпами увеличения научного знания. Согласно данным, утвержденным Высшей аттестационной комиссией (ВАК), количество диссертаций на соискание ученой степени в последнее время значительно выросло [1].

Из этого следует, что в работе будут рассмотрены три ключевых понятия: корпус текстов, ключевое слово (лексический состав), значение(семантика).

ГЛАВА 1 КОРПУСНАЯ ЛИНГВИСТИКА

1.1 Корпус текстов

Под текстовым корпусом в современной лингвистике понимается ограниченный в размере набор текстов, пригодный для машинной обработки и отобранный так, чтобы наилучшим образом представлять языковое множество. Следовательно, представленное документально научное знание можно считать текстовым корпусом. Возникает проблема поиска и анализа научной информации в корпусе неструктурированных текстов.

Корпус текстов содержит большое количество информации. Для извлечения более важной и нужной информации на сегодняшний день используется большое количество программ, методов и алгоритмов. Для человека в большинстве случаев это задача становится непосильной, именно поэтому сейчас ведется масштабная разработка способов нахождения ключевой информации, а именно слов-маркеров. В большинстве своем научное сообщество предлагает несколько путей выделения главных моментов.

Корпус документов может содержать тексты одного языка (одноязычные корпусы) или нескольких языков (многоязычные корпусы).

Многоязычные корпусы, создаются специально для сопоставления. Корпусы данного типа называют сопоставительными.

Для приобретения корпусом более простой и понятной формы используется разметка текста, так называемая аннотация. На дальнейших этапах работы с корпусом текстов это упрощает поставленную задачу. Примером этого может быть морфологическая разметка, которая производится с помощью специальных программ автоматического морфологического анализа. На сегодняшний день идет глобальная разработка таких программ. Разные исследователи и ученые предлагают свою методы и разработки. К примеру, в научных статьях Квятковской и

Седовой предлагается метод латентного анализа. Данному методу я уделю особое внимание далее.

К некоторым корпусам применяются дальнейшие структурные уровни анализа. Например, некоторые маленькие корпусы документов могут в большинстве своем быть полностью синтаксически размечены. Данные корпусы часто называют глубоко аннотированными или синтаксическими, а сама при этом синтаксическая структура, обычно называется древом зависимости. Сложность разметки целого корпуса часто заключается в том, что такие корпусы чаще всего небольшие и содержат в среднем от одного до трёх миллионов слов. Допускается использование и других уровней лингвистического структурного анализа, включая аннотацию морфологии, семантики и прагматики.

Корпус - основное понятие и база данных корпусной лингвистики. Темой большинства работ в компьютерной лингвистике является анализ корпуса документов и его последующее деление, выделение главное информации, ранжирование, распознавания речи и машинного перевода, в которых корпусы часто применяются при создании скрытых Марковских моделей для маркирования частей речи и других задач. В процессе обучения иностранному языку корпусы и частотные словари могут сыграть немаловажную роль и могут оказаться очень полезными.

1.2 Анализ текстовых документов

Автоматизированный интеллектуальный анализ текстовых документов включает в себя:

• Выбор определенного количества документов для анализа.

•Первоначальную подготовку документов для анализа. Производится доморфологический анализ, из документа удаляются все неинформативные слова, все слова приводятся к нормальной форме.

• Извлечение информации и применение методов Text Mining. На этом этапе решается достаточно широкий круг задач. Соответственно, существует

определенный набор методов для решения каждой конкретной задачи. В данной работе принципиально важно решение следующих задач: извлечение ключевых понятий и отношений между ними, извлечение частоты выявленных терминов.

• Интерпретацию результатов. Результаты анализа предоставляются аналитику в графическом и текстовом виде.

Разработано достаточное количество алгоритмов для извлечения из заданного корпуса текстов концептов, отношений между ними, а так же, характеристик выявленных концептов и отношений между ними.

Для качественного автоматизированного наполнения семантической модели содержательного компонента необходимо использовать следующие объекты:

• Корпус текстов в заданной предметной области. При анализе корпуса необходимо опираться на закономерности присущие научно-техническим текстам.

• Наиболее полный словарь определений понятий заданной предметной области.

• Набор когнитивных признаков, формирующий заданный концепт. Когнитивные признаки можно получить в результате экспериментов с носителями языка. Подобные эксперименты необходимо проводить, опираясь на принципы, описанные в работе.

1.3 Семантические связи в корпусе текстов

Рассмотрим требования к представлению системы семантических и, соответственно, лексических связей в тексте:

• Семантические связи, существующие между терминами должны быть представлены в явном виде;

•Семантические отношения и отношения между терминами в корпусе должны быть представлены однозначно;

• Система семантических связей должна быть представлена между любыми двумя терминами в корпусе текстов

• Представление должно обеспечивать определение количественные параметров, характеризующих некоторые семантические свойства отдельной единицы в структуре текста.

• Характер и направление также должны быть определенны представлением наравне с семантическими связями.

Когда происходит переоценка и изменение знаний в какой-либо предметной области, оно сказывается в первую очередь на частоте использования элементов, выражающих смысл, несущих смысловую нагрузку (в последующем используется определение “термин”), начинают появляться новые отношения между терминами, сюда же можно отнести и в исчезновение старых отношений между терминами. Вывод состоит в том, что модели необходимо обладать правом добавлять новые понятия и отношения между ними, не испортив и не нарушив общую структуру модели.

Система семантических связей обладает следующими основными свойствами:

1) является сложным образование, состоящим из более мелких частиц, существует в иерархической системе;

2) составляющие могут быть двух типов: связи и элементы

3) между каждой парой элементов существует прямая либо опосредованная связь

Данные свойства присущи не только семантической системе, но и в тоже время графу. Из этого можно сделать вывод, что граф можно использовать как модель для системы семантических связей.

Граф (англ. graph) - количество непустого множества вершин и наборов пар вершин (связей между вершинами); основной объект изучения математической теории графов. В графе объекты изображаются в качестве вершин, или узлов графа, в то время как связи - как дуги, или рёбра [2].

Главное преимущество семантической сети перед другими моделями представления знаний заключается в том, что она более других соответствует современным представлениям об организации долговременной памяти человека.

Важно понимать, что формализованное описание системы семантических связей в лексике без аппарата определения семантических параметров лексики не является моделью.

1.3.1 Семантика элементов текста

Семантика-раздел лингвистики, изучающий смысловое значение единиц языка. Инструментом изучение служит семантический анализ.

Семантическое значение слов является одним из важнейших компонентов для процедуры извлечения слов-маркеров из корпуса текстов или документа. Оно объединяет слова-маркеры в определенную группу и упрощает поиск таких групп. Что, следовательно, упрощает выделение значимой информации из корпуса научных текстов.

Главной проблемой в декодировании информации текста является незнание вариантов языкового кода. Помимо знания сочетания морфем, слов, предложений, то есть знания общеязыкового кода, необходимо быть ознакомленным и с вариантами данного кода, которые определяют правила использования языковых средств в каком бы то ни было тексте. Вторичные коды также используют знаки, который являются не менее важными. Данный знаки представляют из себя сложную структуру и выполняют необычные, специфические функции. В процессе разговора(коммуникации) говорящий осуществляет отбор языковых проявлений в зависимости от контекста и среды обитания. Существенной проблемой является и то, какие именно единицы текста являются главными для понимания всего текста. Значение отдельных элементов текста для выражения общего смысла неодинаково, и наряду с центральными элементами имеются также и второстепенные элементы текста. При изменении центральных элементов смысл текста

меняется. Если изменить второстепенные единицы текста, то общий смысл может сохраниться. Актуальным на настоящий момент является вопрос о составлении объективной методики выделения слов-маркеров в научном тексте.

Для структуры научного текста характерны развернутые синтаксические конструкции. Для ее упрощения и стереотипизации, прибегают к использованию параллельных конструкций. Важно отметить, что для научного текста характерно не только использование терминов, но и других емких по смыслу слов (например, имен собственных, географических названий, обстоятельств времени и места). С помощью синтаксического параллелизма в тексте приводятся перечисления, уточнения, также возможно появление синонимичных повторов и замен. Их задача заключается в концентрирование внимания на лексической единице языка, которая несет в себе основной смысл научного текста. Семантическая структура текста представлена в виде особой организованной трехуровневой структуры, состоящей из содержательного, смыслового и уровня замысла, которые сопоставимы с тремя уровнями понимания текста. Данными уровнями являются поверхностный, глубинный, а также, уровень понимания концепта.

1.3.2 Теория семантических полей

По сути дела, теория полей затрагивает большое количество точек зрения, представляющих собой весьма весомые варианты общего концепта - концепт семантических связей слов друг с другом в языке. Теория поля является продуктивной так как в понятии “поле” ученым лингвистам удалось воплотить идею о наличии некой структурной величины, которая объединяет лексику в лексико-семантическое поле, где каждая лексема обнаруживает эту величину как доминантную сему лексического значения.

Проведенный анализ понятия семантического поля определяет, что в роли критерия взаимосвязи лексических единиц и их участия в той или иной

группе выступают “лексические значения в целом”, “смысловой признак”, “семантический признак”, различные значения какого-либо слова или варианты его значения, компоненты значения и другое. В роли такого общего элемента могут также выступать понятие, тема, некоторая ситуация.

Признаки, которые используются в качестве образующего звена семантического поля, делятся на две основополагающие группы. В первую входят признаки, которые так или иначе связанны с лексическим значением; лингвистические признаки. Во вторую группу входят признаки ориентированные на понятийную, предметно-тематическую сферу и другие сферы; их можно назвать экстралингвистическими.

Вместе с этим выявляют два основных подхода к исследованию семантических полей: лингвистический и экстралингвистический. При этом экстралингвистический подход, основоположником которого считается немецкий ученый Й. Трир, был разработан раньше лингвистического.

1.4 Роль и функции лексических единиц в корпусе текстов

До настоящего времени не разработан последовательный алгоритм выделения слов-маркеров из научных текстов человеком. Из-за этого появляется проблема нахождения и разработки программного обеспечения и методов извлечения слов-маркеров для вычислительной техники. В современных алгоритмах извлечения слов-маркеров выделяют три последовательных этапа:

1. Предобработка. На данном этапе могут применяться такие вспомогательные процедуры как графематический анализ, морфологический разбор, лексическая нормализация (также согласование синонимов), удаление стоп-слов (служебной лексики), частеречная разметка, лемматизация (стемминг)

2. Распознавание. Программа выносит справку о принадлежности того или иного слова-кандидата к группе ключевых слов.

3. Постобработка. На этом этапе может осуществляться усечение списка, приведение текста в порядок и его ранжирование, визуализация методами когнитивной графики т. п.

Набор методов автоматического извлечения слов-маркеров классифицируют по нескольким признакам:

1. Присутствию элементов обучения и подходов к их реализации;

2. Типу математического аппарата системы распознавания, обусловленного формой информации представления признаков ключевых слов;

3. Списку методов, используемых для лингвистических операций.

В результате анализа лингвистической литературы (работы А.Р. Лурии, Л.В. Сахарного, Л.Н. Мурзина, Н.А. Шехтмана, Ю.Н. Караулова, Geoff Thompson и др.) я вывел следующее понятие слов-маркеров в корпусе текстов: Слова-маркеры - это “семантические доминанты”, сигнализирующие о тональности и характеристиках, которые вложены в текст автором, одинаково понимаемые членами одного социума, облегчающие диалог между ними, позволяющие проникнуть в то, что находится за текстом.

Слова-маркеры представляют набор отличительных признаков, позволяющих определять и извлекать их из корпуса научных текстов. Семантические поля слова-маркера в корпусе текстов помогают определить является оно маркером или нет. Результатом экономии служит тот факт, что наиболее тяжелые слова по своему смыслу становятся в дальнейшем терминами. Термин часто характеризуется информационной насыщенностью, которая определяется числом непосредственных и опосредованных семантических компонентов термина. В реферате (как вторичном научном тексте) также может находиться эксплицитно выраженное указание на круг читателей, и оно также может не содержать терминов данной отрасли знания.

1) Минимальный набор слов-маркеров направлен на отображение в наиболее чистом виде тональности и характеристик заложенных в тексте;

2) Компонентный анализ слов-маркеров текста позволяет извлечь целостную характеристику тональности текста, которая является относительно неизменной и независимо от интерпретатора;

3) Слова-маркеры позволяют читателю быстро сориентироваться в том какую характеристику автор вкладывает в текст; Логически упорядоченные слова-маркеры представляют полное представление о характеристике, которую вкладывает автор текста, вводят ограничения на направление ассоциаций и импликаций читателя.

4) Коммуникативная функция выполняется за счет извлечения слов-маркеров корпуса научных текстов [4].

Следовательно, исследование позволило выделить основные функции ключевых слов в научном тексте. Результат исследования указывает на сложность и многоаспектность понятия слов-маркеров, их значимость для проникновения в смысл текста, подчеркивает необходимость дальнейших разработок методики выделения слов-маркеров в научном тексте для определения тональности текста.

Итак, можно сделать вывод, что к настоящему времени назрела необходимость обобщения, систематизации и интеграции всех существующих точек зрения на феномен слов-маркеров для уточнения терминологического статуса этого понятия выявления его сущности, определения функций и роли слов-маркеров как особых лексико-семантических единиц научного текста.

Выводы по главе 1

Из первой главы работы можно сделать вывод что, основными понятиями в работе с корпусом текстов являются слова-маркеры и семантика текста. Существует ряд способов извлечения ключе из корпуса текстов. Одним из главных является метод графа. Его суть заключается в представлении множества слов-маркеров в виде графика, который указывает, а их частотность в корпусе текстов.

Так же одним из основных моментов в работе с корпусом текстов и при его разборе являются семантические связи, которые содержаться в текстах. Они помогают более структурировано извлекать информацию, в приемлемом для работе виде. Слова-маркеры обладают множеством признаков, которые помогают находить и извлекать их из корпуса текстов. Семантическая связь слов в тексте так же является одним из таких признаков. Для более детального рассмотрения и изучения семантических связей между составляющими предложения необходимо учитывать его семантическую структуру, которая выражается в представлении его не только как единого целого, но и как содержащего более мелкие уровни.

В первой главе приведены понятия и основные функции слов-маркеров и семантических связей в корпусе текстов. Так же, представлены положительные моменты, которые могут улучшить и упростить работу с информацией в сборе текстов.

ГЛАВА 2 АНАЛИЗ ТОНАЛЬНОСТИ КАК ОДНО ИЗ НАПРАВЛЕНИЙ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ

2.1 Извлечение информации

Извлечение информации — это задача автоматического извлечения (построения) структурированных данных из неструктурированных или слабоструктурированных машиночитаемых документов [7].

Автор рассказывает, что неотъемлемой частью любой информационно-поисковой системы является индексация документов, которая в автоматическом режиме создает из текста индекс, т. е. переводит текст в записи таблицы базы данных. В данной работе используется структура индекса.

В ходе прочтения данной статьи я узнал, что для каждого термина сохраняется его личный код, частота употребления в данном документе, строковое представление отклонение от начальной формы, начальная форма, значение измерений C-value, код документа, в котором встретился данный термин, значение TF * IDF.

Также для каждого вхождения термина в документ хранится его код, код термина, номер предложения в документе, в котором встретился данный термин.

Для каждого документа хранится его код, путь к документу, дата индексации, количество слов в документе, количество терминов в документе, количество уникальных терминов в документе, размер файла, автор документа, название документа, дата защиты диссертации.

Для выделения терминов используется мера C-value. Автор с помощью графиков и таблиц описывает эксперимент, показывающий, что именно эта мера лучше всего позволяет выделить термины в текстах на русском языке.

2.2 Анализ тональности текста

Под анализом тональности текста, принято понимать группу методов, которые исследуют содержание текста, предназначенных для автоматизированного извлечения из текста или корпуса текстов, лексики, которую автор эмоционально окрашивает, либо же в которую вкладывает свое собственное мнение по отношению к человеку или его работе (т.е. объектам), о котором идет речь.

В целом тональность текста определяется как функция текста или корпуса текстов. Тональность всего текста состоит из тональности всех его составляющих частей и их взаимосвязей в тексте. Тональность – эмоционально окрашенное отношение автора к предмету его работы или темы, которую он изучает.

Эмоциональный компонент либо же составляющая выражена на уровне лексем или коммуникативных фрагментов, который в свою очередь является лексической тональностью. Через данную функцию автор проявляет свое отношение в целом к миру, его объектам, предметам и целям своего собственного исследования.

2.3 Методы анализа тональности документов

Существует два типа методов выделения слов-маркеров из корпуса текстов: статические и гибридные. Они могут требовать или же совсем не нуждаться в наличии корпуса текстов. Гибридные методы не требующие корпуса текстов описывал Кен Беркер и другие.

Существует различие между двумя гибридными методами. Различия заключаются в способе отбора кандидатов на присутствие в вершинах графа. Еще одним параметром является непосредственная близость лексем кандидатов по отношению к друг другу. В большинстве своем данные методы основаны на морфологическом, синтаксическом или даже семантическом анализе. Семантический анализ, например, описан в трудах Гриневой и связан с работой с Википедией.

Методы на основе машинного обучения так же можно отнести к числа гибридных методов. Задача извлечения слов-маркеров в таких методах рассматривается в качестве классификации. Машинные методы как правильно используют корпус текстов. Им необходим корпус с размеченными словами-маркерами.

В данной системе существует положительные и отрицательные примеры. Слова-маркеры могут быть как положительными, так и отрицательными. После этого идет следующий шаг, высчитывается релевантность каждого отдельного слова в пробном, тренировочном тексте путем сопоставления его с вектором значений различных параметров. К примеру, длинные слова словосочетания, длинные слова в абзацах и заголовках. Затем записывает значение различий между векторами для слов-маркеров и остальных слов. После этого считается примерная вероятность попадания слова в группу слов-маркеров и задается порог. Другими словами, можно сказать что модель обучается.

Анализ слов-маркеров, находящихся в научных текстах происходит путем подсчета показателя релевантности этих слов и существующей вероятности отнесения слов кандидатов в группу маркеров. Это происходит в соответствии с построенной по алгоритму моделью.

Шереметьева С.О. и Осминин П.Г. рассматривают экстрактор слов-маркеров из корпуса научных текстов. Данная программа призвана быть достаточно универсальной. В ходе работы программы должна осуществляться настройку на извлечение различных лексических групп и текстов различных предметных областей.

Авторы рассказывают каким требованиям должна отвечать программа. Во-первых, программа должна обеспечить лингвистические правильные и корректные результаты. Во-вторых, одной из главных задач, научить программу работать без корпуса текстов. Как правило на практике таких корпусов текстов часто не существует, поэтому создание таких текстов часто является нетривиальной задачей. В-третьих, авторы ставят перед программой

задачу обладать выслительно-приемлимыми свойствами и обеспечивать высокую скорость обработки данных документа. В-четвертых, обеспечить экстракцию не только высокочастотных слов в документе, но и низкочастотных. В-пятых, создать и обеспечить достаточно быстрое создание программного инструмента с помощью повторного использования.

Программа работает по следующему алгоритму вычисления n-грамм:

1.вычесление n-грамм (n=1,2,3) существующего исходного документа

2. Извлечение n-грамм которые не могут состоять в лексических группах требуемого вида

3. Привидение форм слова к одной лексеме (процесс нормализации)

Методы извлечения слов-маркеров для анализа тональности текста.

Лингвистическая информация из корпуса извлекается при помощи специальных компьютерных программ. Существует несколько основных источников разработки подобных программ.

Первый, это лингвистические отделы больших коммерческих проектов, в основном, связанных с публикацией словарей. Например, Cobuild Project. Часто это закрытое программное обеспечение, стоящее больших денег.

Второй источник разработки — компьютерная лингвистика и учёные, которые ей занимаются. В её рамках было создано немало программ, осуществляющих автоматический анализ грамматики и семантики, анализ и синтез текста, автоматический перевод и другие приложения для компьютерной обработки естественного языка. Конечно, не был обойдён стороной и анализ корпусов, в том числе, средства автоматической грамматической и синтаксической разметки — вероятностные (probabilistic), либо на основе правил (rule-based).

Латентный семантический анализ.

Латентный семантический анализ был основан в 1988 году. Сперва латентный семантический анализ был разработан и использовался в целях автоматизированной индексации корпуса текстов, выявлений семантических структур текста и получения не существующих документов

(псевдодокументов). В процессе развития данный метод успешно использовался для работы с базами знаний и построения когнитивных моделей.

В последнее время данный метод обычно применяется для поиска информации (присвоения индекса документам), разбития документов по классам, моделям понимания и во многих других областях, где необходимо выявление важных факторов из большого количества впускных данных или корпуса данных.

Латентно семантический анализ часто сравнивается с несложным видом нейросети, которая состоит из трех компонентов: первый компонент содержит множество слов (термов), второй – некое множество документов, соответствующих определенным ситуациям, а третий, средний, скрытый слой представляет собой множество узлов с различными весовыми коэффициентами, связывающих первый и второй слои.

Существуют три основных разновидности решения задачи методом ЛСА:

-сравнение двух термов между собой;

-сравнение двух документов между собой;

-сравнение терма и документа.

Достоинства метода:

1) метод является наилучшим для нахождения латентных зависимостей внутри множества документов;

2) метод подходит в использовании как с обучением, так и без обучения (например, для кластеризации);

3) используются значения матрицы близости, основанной на частотных характеристиках документов и лексических единиц;

4) в некоторых случаях возможно избежать полисемию и омонимию.

Недостатки метода:

Одним из главных недостатков метода является существенное снижение скорости вычисления при увеличении объема входных данных.

Чаще всего вероятностная модель метода не соответствует реальности. Принято считать, что слова и документы находятся в нормальном распределение, тогда как ближе к реальности в так называемом распределении Паусона. В свете этого недостатка для практических применений лучше подходит Вероятностный латентно-семантический анализ, который основывается на мультиномиальном распределении.

В работах Я. А. Седовой, И. Ю. Квятковской также преобладает метод латентного семантического анализа. Он описывает преимущества и необходимость использования именно этого алгоритма при выделении информации, а именно ключевых слов из научных статей (корпуса текстов). (LSA – latent semantic analysis) используется как метод распознаванья сходства значения слов и документов выполненный на основе статистических вычислений над большим текстовым корпусом. Он использован, т.к. требуется дополнительной информации, такой как построенные вручную словари, семантические сети или базы знаний.

В основе метода LSA заложена гипотеза о том, что между словами контекстом, в котором они используются и употребляются, существуют неявные (латентные) взаимосвязи. Авторы предполагают, что семантическое значение документа может быть показано и представлено как сумма значений, использующихся в нем слов: значение(документ) = значение (слово_)1 + значение (слово_)2 +...+ значение (слово_ m).

Метод позволяет вычислить корреляции между парой терминов, между парой документов и между термином и документом.

Каждая строка исходной матрицы C – вектор, соответствующий термину и показывающий его связь с каждым из документов корпуса**:**

Метод C-Value.

Словосочетания, схожие по своей структуре с терминами, выделяются из корпуса текстов или документа с помощью метода C-value [3]. Описаны семантическая модель корпуса документов и алгоритмы, позволяющие представить его в форме графа для последующего анализа. Разработан

алгоритм поиска в корпусе документов с помощью созданной модели. Предлагается подход к обработке текстов авторефератов кандидатских и докторских диссертаций.

Формальным критерием наличия семантической связи можно считать наличие синтаксической связи. Но система синтаксических связей не эквивалентна системе семантических связей. Поэтому выбор предложенного формального критерия наличия семантической связи между терминами влечет за собой некоторые потери.

Анализ семантики языковых единиц, представляющих номинативное поле позволит работать с семантическим содержанием концепта. Такой анализ позволит работать лишь с частью концепта, но той, которая нашла отражение в языке, а значит наиболее релевантной. Очевидно, что от уровня наполнения номинативного поля зависит качество определения содержания и структуры концепта.

Автоматизированное наполнение и использование семантической модели содержательного компонента потребует следующих алгоритмов:

1. Алгоритм, позволяющий составить частотный словарь по заданному корпусу текстов.

2. Алгоритм, позволяющий преобразовать словарь терминов в файл пригодный для автоматической обработки при определении связей между концептами.

4. Алгоритм, позволяющий преобразовать исходный корпус текстов в файл пригодный для автоматической обработки при определении частоты со встречаемости концептов.

5. Алгоритм, позволяющий определять связи между терминами, а также веса концептов и связей между ними.

6. Алгоритм, позволяющий определять степень соответствия одной модели другой в процентном соотношении

Поиск документов при наличие распределенного знания.

Анализ прочитанного материала дал понимание, как правильно производить поиск документов при наличии распределенного знания. В статье Я. А. Седовой, И. Ю. Квятковской “System analysis of the scientific documentation corpus” предлагается подход и метод решения этой задачи. Для проведения информационного поиска при наличии распределенного знания предлагается использовать агентный подход.

В своей статье автор говорит, что: “Агент – это программный модуль, который осуществляет обход веб-ресурсов по списку и загружает с них файлы определенного типа. Для кандидатских диссертаций этими веб-ресурсами являются сайты тех организаций, в которых действует соответствующий диссертационный совет, для докторских – сайт ВАК. Например, авторефераты представлены в виде файлов в форматах *.doc, *.pdf”.

Представленные в статьях методы совпадают друг с другом в некоторых моментах, но всякие направлен на достижение цели по своему конкретному пути. В связи с ростом количества научных статей публикаций, каждый новый метод не может быть идеален. В большинстве своем данные методы используются только теоретически, воплощение их в жизнь не всегда дает ожидаемый результат.

Russian Distributional Thesaurus (сокр. RDT) — проект создания открытого дистрибутивного тезауруса русского языка. На данный момент ресурс содержит несколько компонентов: вектора слов (word embeddings), граф подобия слов (дистрибутивный тезаурус), множество гиперонимов и инвентарь смыслов слов.

Все ресурсы были построены автоматически на основании корпуса текстов книг на русском языке. В следующих версиях ресурса планируется добавление и векторов смыслов слов для русского языка, которые были получены на основании того же корпуса текстов.

RDT представляет собой первый свободно доступный дистрибутивный тезаурус русского языка. Данный лингвистический ресурс покрывает около

миллиона наиболее частотных слов русского языка и представляет значительный интерес для задач автоматической обработки текстов.

Для более четкого извлечения информации из корпуса текстов следует структурированно производить все исследования. Во-первых, необходимо создать таблицу слов, которые несут главную тему и суть нужной для извлечения информации. Таблица слов также может называться wordlist.

Таблица просто представляет собой список слов, которые будут нести ключевое значение для вашей темы. На данный момент разработано большое количество компьютерных программ, которые позволяют произвести этот этап работы с текстом гораздо быстрее и эффективнее. Одной из таких программ является Russian Distributional Thesaurus.

Тезаурус представляет собой большое количество источников информации русского языка. В него входят произведения литературы, статьи из журналов, научно-популярные статьи и многое другое. Работа многих программ в основном заключается на поиске ключевых слов и сементических связей между ними. При работе с текстом поиск и извлечение нужной информации может обеспечить более глубокое понимание темы и более тщательное ее рассмотрение.

Чтобы предлагать новые решение для более четкого и быстрого извлечения информации из научных текстов, для начала нужно досконально понять, что из себя представляют понятия корпуса текстов, ключевых слов и семантических связей.

Структурирование и рефератирование текстов намного упростить взаимодействие человека с большими источниками текстов. Создание такого рода программ производится для более близкого ознакомления человека с темой работы и позволит сблизится человеку с научной сферой. Облегчение доступа к информации повысит уровень знаний.

То есть в конечном итоге создание программ для извлечения информации из корпуса научных текстов повысит глобальную образованность человечества. Поэтому эта тема привлекает многих ученных в наше время. За

последние годы был совершен большой прорыв в этой области. Ежегодно выпускается большое количество методов для упрощения работы человека с текстами.

По сути создается искусственный интеллект, который заменит человека в научном процессе. Это хорошая тенденция, которая позволит продвинуть науку на новый уровень во многих отраслях.

Статистика корпуса

Данный корпус текстов содержит 12.9 млрд словоупотреблений (150 Гб текста), извлеченных из коллекции книг на русском языке в формате FB2, очищенных от метаданных. Корпус был использован для обучения векторных представлений слов, на основании которых был построен дистрибутивный тезаурус русского языка RDT.

Русский тезаурус активно поддерживается сервисом NLPub. За счет совместных усилий продуктивность тезауруса удалось поднять на новый уровень. В разработке принимали участие не только русские ученые, но и зарубежные коллеги. Проект разрабатывается усилиями представителей УрФУ, МГУ им. Ломоносова, Университета Гамбурга. В прошлом в проект внесли свой вклад исследователи из Южно-Уральского государственного университета, Дармштадского технического университета, Волверхемтонского университета и Университета Тренто.

Каталог NLPub организован по вики-принципу и содержит сведения об инструментах, ресурсах, методах и алгоритмах, необходимых для успешного построения систем автоматической обработки русского языка. Основные разделы каталога приведены в боковом меню и продублированы в этой таблице.

Внесение новых материалов и сведений о собственных разработках в материалы NLPub горячо приветствуется. Раздел TODO включает возможные темы для новых статей [6].

Программа word2vec.

Инструмент анализа семантики естественных языков, представляющий собой технологию, которая основана на дистрибутивной семантике и векторном представлении слов word2vec является еще одним решением проблемы извлечения информации из корпуса научных текстов.

Работа данного инструмента производится следующим образом: word2vec принимает большой текстовый корпус в качестве входных данных и представляет каждое слово в виде вектора, в итоге изображая координаты слов в виде векторов. Прежде всего он создает словарь, «обучаясь» на предложенных текстовых данных, а затем вычисляет векторное представление слов. Векторное представление основывается на контекстной близости: слова, встречающиеся в тексте рядом с одинаковыми словами (а, следовательно, имеющие схожее значение), в векторном представлении будут иметь близкие координаты векторов-слов. Полученные векторы-слова могут быть использованы для обработки естественного языка и машинного обучения.

Сжатые векторные представления слов

1. Полезны сами по себе, например, для поиска синонимов или опечаток в поисковых запросах.

2. Используются в качестве признаков для решения самых различных задач:

1) выявление именованных сущностей

2) эгирование частей речи

3) машинный перевод

4) кластеризация документов

5) ранжирование документов

6) анализ тональности текста

Алгоритм программы Word2Vec.

Word2Vec включает в себя набор алгоритмов для расчета векторных представлений слов, предполагая, что слова, используемые в похожих контекстах, значат похожие вещи, т.е. семантически близки.

В числителе — близость слов контекста и целевого слова.

В знаменателе — близость всех других контекстов и целевого слова.

Технология Word2Vec использует два разных метода:

Skip-gram – предсказание близлежащих слов на основании одного слова.

При расчете используются искусственные нейронные сети.

Во время обучения алгоритм формирует оптимальный вектор для каждого слова с помощью CBOW или skip-gram.

CBOW - предсказание слова на основании близлежащих слов.

Метод представления слов в виде векторов используется для кластеризации слов и выявления их семантической близости, т.е. разделяет несвязанные слова и соединяет связанные, что помогает в задачах кластеризации и классификации текстов.

Преимущества использования RDT перед самостоятельным обучением word2vec:

1. Нет необходимости подбирать мета-параметры для русского языка. Модель RDT содержит высококачественные предобученные векторные представления слов. Были исследованы различные комбинации мета-параметров модели, такие как тип модели (SkipGram/CBOW) и размер контекстного окна, с использованием множества тестовых коллекций и выбраны оптимальные параметры для русского языка. Стандартные параметры word2vec могут отличатся от оптимальных для русского языка.

2. Нет необходимости в длительном обучении модели. Обучение модели векторных представлений размерности 500 на корпусе текстов из 150 Гб с тремя итерациями по корпусу занимает до нескольких дней на инстансе r3.8xlarge Amazon EC2 с 32 ядрами и 244 Гб оперативной памяти. Вычисление ближайших соседей для миллиона наиболее частотных слов занимает еще несколько дней для векторной модели сопоставимой размерности (вектора размерности 500, лексикон из 7 млн слов).

3. Эффективность в использовании. Многие полезные приложения векторных представлений слов используют только список ближайших слов и

могут обойтись без векторов слов. Например, для лексического расширения запросов и других видов коротких текстов достаточно знать список ближайших слов к целевому слову. Граф подобия слов занимает на порядок меньше памяти (1 Гб оперативной памяти для графа слов по сравнению с 20 Гб для векторов) и не требует ресурсоемких вычислений близости между векторами.

4. Наличие гиперонимов. В состав RDT входят гиперонимы извлеченные из того же корпуса, который был использован для обучения векторных представлений слов.

5. Наличие смыслов слов. В отличие от стандартной модели word2vec, дистрибутивный тезаурус русского языка содержит версии, в которых для каждого слова известно несколько смыслов (например, "ключ" для замка и "ключ" как источник воды). Для каждого из значений представлен список ближайших соседей релевантных данному смыслу.

Проекты:

-NLPub поддерживает открытые проекты по созданию и развитию русских языковых ресурсов. На инфраструктуре NLPub размещены следующие проекты

-Yet Another RussNet — открытый электронный тезаурус русского языка;

-RUSSE — мероприятие по сравнению методов оценки семантической близости русских слов;

-RTLOD — тезаурусы русского языка в виде открытых связанных данных;

-Mechanical Tsar — движок для краудсорсинговой разметки данных;

-Russian Distributional Thesaurus — открытый дистрибутивный тезаурус русского языка;

-LRWC — суждения людей о семантических отношениях.

Кроме того, NLPub является информационным партнёром конференций AINL и ISMW.

Семантическое поле и проблемы синонимии.

В последнее время в лингвистической литературе большое распространение получила полевая модель системы языка, которая имеет разнообразные интерпретации и применения.

Предмет исследования в теории поля в современной лингвистике составляют группировки языковых единиц, объединяемых на основе общности выражаемых ими значений (семантический принцип) или по общности выполняемых ими функций (функциональный принцип), или на основе комбинации двух признаков (функционально-семантический принцип). «В современном языкознании семантическое поле определяется как совокупность языковых единиц, объединенных общностью содержания и отражающих понятийное, предметное или функциональное сходство обозначаемых явлений». Выделяемые на основе этих признаков группировки – поля – представляют собой системные образования с характерными для любой системы связями и отношениями и вместе с тем обладающие собственными специфическими чертами. Семантическое поле характеризуется следующими основными свойствами. Поле имеет особую структуру – ядро и периферия – для которой характерна максимальная концентрация полеобразующих признаков в ядре и неполный набор признаков поля или возможное ослабление их интенсивности на периферии. Специфика поля как способа существования объекта характеризуется явлением аттракции, которое заключается в том, что, «благодаря существованию данной группы элементов с общим признаком в него включаются новые элементы с таким же признаком» [16, с. 101].

Другой важной характеристикой поля служит возможность пересечения отдельных полей, которое приводит к образованию общих сегментов, зон семантического перехода. Переход от ядра к периферии также осуществляется постепенно, вычленяется ряд периферийных зон, в разной степени удаленных от ядра. Семантическая однородность единиц семантического поля вместе с тем не исключает возможности их грамматической разнооформленности. Поле может объединять в своем

составе разнородные языковые средства, принадлежащие к различным грамматическим классам и уровням языка. «Лексико-семантическое поле – это объединение слов разных частей речи, которые на уровне своих лексических значений имеют хотя бы одну общую (интегральную) сему» [14, с. 59].

В теории семантического поля, в зависимости от природы исходной единицы, определяющей семантическую и словообразовательную деривацию его элементов и его частеречный состав, выделяются такие категориальные типы полей, как процессуальные (с доминантой глаголом), предметные (с доминантой именем существительным), признаковые (с доминантой именем прилагательным). Учет типологии семантических полей позволяет выявить для каждого из них детерминанту, как главное свойство, определяющее состав единиц поля и категориальный характер их семантики. В одних доминируют глаголы, а другие члены производны, вторичны, периферийны, во-вторых – существительные, в-третьих, – прилагательные.

Межуровневость является специфической характеристикой функционально-семантических и грамматико-лексических полей [5, с. 17]. Как всякое системно-структурное объединение, поле имеет определенную структуру.

Понятие «структура поля» подразумевает существование определенных группировок элементов внутри данного множества, пересечение отношений в его структуре, наложение связей. Поле может иметь в своем составе несколько микрополей, обладающих относительной самостоятельностью. При изучении отношений между элементами поля до недавнего времени основное внимание уделялось аппозитивным отношениям. «При описании семантических противопоставлений в рамках поля удобно использовать семантические компоненты, описанные выше (речь идет об интегральных и дифференциальных семантических признаках), в частности, бинарные признаки». Вместе с тем признак связности значений между собой по определенному числу однозначных противопоставлений кладется в основу

определения замкнутых групп слов, в лексической семантике называемых лексико-семантической парадигмой. Однако «под понятие парадигмы, накладывающее строгие ограничения на характер связей между словами, подводится лишь небольшая часть группировок слов, объединяемых на основе общности их значений».

Общим понятием, применимым к более широкому кругу парадигматических отношений, является семантическое поле. Частями семантического поля являются группа синонимических корреляций, родовидовые корреляции, корреляции несовместимости, корреляции «часть – целое», антонимические корреляции, конверсивные корреляции, корреляции семантической производности. Изучение семантических корреляций поля позволило установить типы структурных связей, которые формируют все перечисленные группировки – вхождение, схождение и расхождение. Вхождение включает гиперо-гипонимическую связь, пересечение, синонимическую, градуальную и партитивную связь. Схождение включает фазовую связь, тяготение, темпоральную, локальную, инструментальную и реминисцентную связь. Расхождение объединяет антонимическую связь, несовместимость и противодействие [7].

В названии словаря содержится не только термин «синонимы», но и термин «сходные по смыслу выражения», которые Н. Абрамов понимал очень широко, включая в их состав следующие синонимические средства русского языка: 1) аналоги (тематически близкие слова); 2) когипонимы (названия одного из видов того рода, к которому относится исходное слово, например шашки и шахматы); 3) гиперонимы (слова с родовым значением по отношению к исходному); 4) гипонимы (слова с видовым значением по отношению к исходному); 5) названия частей того, что обозначено ключевым словом; 6) конверсивы (У меня есть = Я имею); 7) перифразы ключевого слова (обычно глагола), в которых участвуют его актантные и синтаксические дериваты (помогать, оказывать помощь, подавать помощь); 8) фразеологические синонимы (ошибаться – сбиться с пути, впасть в

ошибку, дать промах, дать маху); 9) синонимы, различающиеся сочетаемостью (Война: внутренняя, газетная, кровопролитная, сухопутная, морская, партизанская, таможенная, наступательная, оборонительная. Вспыхнула война. Война … не знает перемирий.); 10) слова, выражающие тот же или похожий смысл вне зависимости от способа его выражения (очень, весьма, безгранично, бесконечно, крайне, сильно, страшно, ужасно, адски, донельзя, больно, на диво, на чем свет стоит, из рук вон (плохо), тыс.у раз (прав), беда как (умен), ужас как (скуп), давнымдавно, полным-полно, черным-черно); 11) лексико-синтаксические конструкции (приблизительно; лет этак через двадцать) [6].

Понятие антонимии.

За последние годы появилось много исследований по антонимии как в русском, так и в других, в частности английском, языках (Миллер Е. Н., Иванова В. А., Ермаков Н. Ф.). Они отражают эволюцию взглядов лингвистов на само понимание антонимии, дают более глубокое осмысление этого феномена в системе языка.

Антонимы – слова с противоположным значением. В учебниках по лексикологии они определяются и как слова «разного звучания, которые выражают противоположные, но соотносительные друг с другом понятия» [8].

Чтобы и понятия противоположного значения, и противоположность адекватно отражали семантику антонимов, служили средством для их опознавания, необходимо вложить в них, по нашему мнению, строго терминированный смысл, семантически ограничить. Одним из таких ограничений является указание на соотносительные противоположные значения [9].

Рассмотрим развитие взглядов на проблему определения антонимов. Во втором издании Большой советской энциклопедии, вышедшей в начале 50-х годов, говорится, что антонимы бывают только у слов, содержащих в своём значении указание на качество, и являются «словами обязательно разных

корней» [10]. Л. А. Булаховский считал, что «антонимы в основном относятся к выражению качеств, но возможны также, например, при названии действий и состояний отрицательного или отменяющего характера» [11].

Д. Н. Шмелев полагает, что под антонимией понимают не простое противоположение, которое может быть выражено прибавлением отрицания (например, белый – небелый, говорить – не говорить), а противопоставление допускающих это значений, выраженных различными корнями (например, бедный – богатый, сухой – мокрый) [12].

Л. Ю. Максимов в основу определения антонима помещает признак качества, однако в это понятие он включает не только разнокоренные, но и однокоренные слова – антонимы [13]. Вслед за ним мы также полагаем, что признание антонимами только разнокоренных слов не отражает действительного положения вещей, т. к. выражение антонимических отношений в русском языке также осуществляется при помощи приставок (не- и др.), а в английском выражение антонимии при помощи аффиксов (un-, im, – a, – dis, mis) является определяющим.

В противовес точке зрения, что антонимы могут иметь лишь слова, выражающие качество, мы можем привести мнение Л. А. Вараксина, занимающегося проблемой глагольной антонимии в русском языке. В своей работе «Однокорневые префиксальные глаголы – антонимы в современном русском языке» он совершенно справедливо и оправданно приходит к выводу, что «сущность явления антонимии не может быть сведена лишь к качественной противоположности лексических единиц», и далее – «разряд слов, являющихся выражением противоположных понятий, при таком ограничении совершенно неоправданно сужается, что затрудняет лингвистический анализ антонимии как единого, целостного явления» [14]. Однако, по нашему мнению, уделение чрезмерного внимания качественному признаку несколько сузило понимание противоположности, а, следовательно, и ограничило понятие антонимии. В «Словаре лингвистических терминов»

О. С. Ахмановой находим следующее определение антонимов: 1. Это слова, имеющие в своем значении качественный признак и потому способные противопоставляться друг другу как противоположные по значению. Например, хороший – плохой, близкий – далекий, добрый – злой, а также 2. Слова, противопоставленные друг другу как коррелятивные (брат – сестра), как обозначающие противоположно направленные действия (уходить – приходить) и т. п. [15].

Наряду с данными работами были предприняты попытки описательного определения антонимов в русском языке через перечисление их различных свойств, в частности, предпринятые П. А. Введенской в её введении к словарю антонимов русского языка. Как видно из вышеназванных примеров, определения антонимов в русском языке весьма разнообразны. Рассмотрим, какую альтернативу данному вопросу УДК 80/81 предлагает зарубежная лингвистика. Обзор работ по антонимии в зарубежном языкознании даёт Х. Геккелер в своей книге «Структурная семантика и теория семантического поля» [16]. Он отмечает, что по сравнению с синонимией антонимия изучалась значительно меньше. Она стала самостоятельным объектом исследования в основном в последнее время в связи с развитием структурной лингвистики. И, тем не менее, определение этого явления имеет по сравнению с принятым традиционным достаточно дифференцированный характер.

По данным словаря лингвистических терминов Дж. Кноблоха, «антонимом считается слово, которое вступает с другим словом в отношение контрадикторной, контрарной или коррелятивной противоположности [17]. Несмотря на то что важность антонимии как особого противопоставления, особой ассоциации слов подчеркивалась неоднократно, например, в связи с теорией семантических полей Трира, а также в связи с проблемой идентификации и классификации фактов в лингвистике (Ш. Балли), лингвисты и в настоящее время исходят часто из весьма недифференцированного понятия антонимии. Так, А. де Винценз

предпочитает говорить просто об антонимичных парах, вообще не обсуждая понятия антонимии. Вместе с тем следует отметить ряд работ, которые содержат объективные предложения в решении этого вопроса. Это прежде всего, на наш взгляд, те труды, которые расширяют прежнее узкое понимание антонимии.

В своём исследовании о синонимах И. Филлипец намечает несколько структурных типов антонимов, которые выражают не только качество, но и количество, оценку, пространственные и временные отношения, противоположности действий и состояний [18]. Аналогичную картину находим в очерке по семантике французского языка О. Духачека. Он даёт схему классификации французских антонимов и отмечает, что в антонимические отношения могут вступать слова, обозначающие качество, количество, оценку, чувства, действия, состояния, пространственные и временные отношения [19].

Что касается именно англистики, то важную лепту в исследование английских антонимов внёс Дж. Лайонз, который имел тенденцию дифференцировать глобальное понимание антонимии. В своих работах «Структурная семантика» и «Введение в теоретическую лингвистику» Дж. Лайонз представил своего рода классификацию всех антонимов английского языка по типам противоположностей. Он выделил три класса антонимов – комплементарные, собственно антонимы и конверсивы, описал их свойства [6].

Далее в нашей работе мы подробно рассмотрим данные группы антонимов в сравнении с русскими эквивалентами. Говоря об исследовании проблем определения антонимов в английском языке, нельзя не упомянуть и о советском учёном В. Н. Комиссарове. Он полагал, что слово является антонимом только в том случае, если регулярно употребляется в антонимических контекстах, а также если имеет одинаковую сферу лексической сочетаемости [7].

Во множестве сборников по лексикологии английского языка (под ред. Арнольд, Гинзбурга) данная точка зрения принимается, как базовая и другая не предусматривается. Мной она рассматривается как не совсем правильная, т. к., например, критерий одинаковой лексической сочетаемости не всегда применим для всех слов, являющихся антонимами, а употребление их в пяти антонимичных контекстах не всегда требуется для доказательства антонимии слов, т. к. она может следовать из жизненного опыта людей.

Итак, данные существующие определения антонимов, по нашему мнению, можно условно разделить на три группы. Первая группа – определения, относящие к антонимам слова с противоположными значениями. Подобные определения имеют частные вариации, но главное в том, что отмечается наличие противоположных значений. Это определения антонимов Реформатского, Шанского, Степанова. «Антонимы – это слова противоположного значения». «Антонимы являются словами разного звучания, которые выражают противоположные, но соотносительные друг с другом понятия» [14]. «Антонимы – это слова, противоположные по сигнификативному значению» [13]. К данной группе также возможно отнести вышеуказанные определения В. А. Ивановой, Д. Н. Шмелева.

Так как, данные методы помогают провести анализ тональности текстов они представляют ценность для работы. Некоторые из алгоритмов и методов будут использованы в ходе анализа тональности корпуса рецензий на английском языке в практической части работы.

Выводы по главе 2

Каждый отдельный метод рассматривает корпус текстов, слова-маркеры и семантические связи под разным углом. Каждый метод имеет свои плюсы и минусы.

Компьютерная лингводидактика как синтез двух наук должна развиваться. Для этого требуется проведение больших экспериментов. Результат уже становится заметным. Анализ информации из научных

источников на предмет тональности приобретает адекватный вид и существенно помогает в понимание научных текстов, посредством предоставления человеку характеристики текста. Три основных компонента должны существовать и быть исследованы параллельно. Так как все они играют важную существую роль в итоговом результате. Этими тремя компонентами являются корпус текстов, либо документ, слова-маркеры, семантика слов. На этих трех компонентах в большинстве случаев и строятся представленные методы и подходы.

Так же не мало важно использование и развитие компьютерных технологий. По природе человека он может держать в голове порядка 8-15 слов. Этим выражена так называемая оперативная память человека. Конечно, человек в силах предоставит анализ текста, разбить его на тезисы, максимально сжать текст, и выделить слова-маркеры, главные моменты в документе. Но когда речь идет о корпусе документов, которые как минимум состоит из порядка 1-3 миллионов символов, данная задача кажется просто невыполнимой. Что же касается машины она может произвести данные действия в течении короткого количества времени.

Сам процесс обучения ускорится, если человеку необходимо будет понять тональность текста из большого количества, маловажных моментов. Так как на сегодняшний день список кандидатов на соискание ученых степеней резко увеличился, а, следовательно, и количество диссертационных работ и всякого рода публикаций также увеличилось.

Обработка данной информации ставит не простую задачу перед созданными программами. Тем более, что многие из них находятся на данный момент лишь на стадии разработки. Хотя существует и уже готовые к использованию методы, такие как например латентно семантические анализ. Однако и эта система имеет свои недостатки. Такие как например смешение семантики слова и ее замена.

Так же сейчас пишется много работ, на тему как улучшить эту сферу компьютерной лингводидактики. Каждый текст либо же документ, или

корпус текстов, должен быть подвержен анализу, должен быть разбит на части. Не каждый метод учитывает данные условия. Поэтому нередко явление смешение семантических связей, и как следствие выделение не самой главной и важнейшей информации.

Мое мнение заключается в том, что машина не всегда способна определить главную идею, которую человек заложил в какой-либо документ. Особенно проблема возрастает, когда дело касается корпуса текстов, который может быть написан не одним, не двумя, а десятками авторов. Необходимо изобрести такой метод, который будет улавливать правильно семантику слов-маркеров, что позволит максимально приблизится к желаемому результату.

Исследования, которые проводятся на сегодняшний день развиваются в правильном направлении. Главным плюсом будет приобретением из каждого метода абсолютных неповторяющихся компонентов. Объединение структур метод, так как в большинстве своем они несут одну идею и раскрываются по одному алгоритму.

Алгоритм в данных программах так же имеет немаловажную роль. Каждый шаг должен быть заранее и четко продуман автором разработки. Правильная и адекватная структура самой программы позволит более тщательно и скрупулёзно отбирать необходимую информацию.

Так же в новейших подходах необходимо учитывать взаимосвязь компонентов и их взаимоотношения при нахождении в одной среде, в нашем случае программе.

Немаловажно учитывать структуру каждого отдельного языка, его особенности и принципы изложения исходной информации. Культурные нормы, нормы морали, принципы общения и донесения информации между людьми в одной языковой среде. Поток научной информации с развитием технологий возрос в большинстве стран мира. Американские и английские ученые не исключение в этом списке. Нам так же необходимо получать важнейшую информацию от западных авторов и ученых. Обмен знаний

достаточно важен в развитии науки. Программы, методы, подходы позволят на достойном уровне ознакомится с идеями как наших, так и зарубежных авторов.

Компьютерные технологии позволяют развиваться науке в правильном направлении. Не так далек тот день, когда машины если не полностью заменят человека в вопросах выделения информации из корпуса текстов, то по крайне мере выдут с ним на один уровень по способности распознавания тональности.

Не только ученые из сферы компьютерной лингводидактики, но и из различных других сфер науки привносят вклад в решение данного вопроса. Например, люди, занимающиеся информационными технологиями, математикой, физикой и другими отраслями науки.

Преобладающим элементом в ходе выделение информации из научных текстов является структурный анализ этого корпуса. Каждая программа разбивает текст следуя своему собственному алгоритму. Например, в научной статье Квятковской и Седовой предлагается рассматривать корпус текстов в виде матрицы. Где каждый элемент занимает свою конкретную позицию. Это позволяет рассматривать каждый элемент матрицы в виде вектора.

Существует особый векторов в m-мерном пространстве, где m – количество терминов во всех документах корпуса. Представлении корпуса документов в качестве векторов, позволяет использовать аппарат векторной алгебры для векторного анализа корпуса документов.

В ходе проделанной работы были изучены понятия корпуса текстов, слов-маркеров, семантики слов. Эти три понятия являются ключевыми в моей теме. Были изучены разные взгляды на извлечении информации из корпуса текстов. Так же в нашей работе были рассмотрены разные способы, использующиеся для более продуктивного извлечения информации. На сегодняшний день данная проблема является актуальной и активно изучается многими учеными.

В ходе работы основной проблемой оказалось описать, сравнить и понять способ извлечения информации и подходы разных ученых.

Были изучены гибридные методы извлечения слов-маркеров из корпуса текстов. Методы машинного перевода бывают следующих видов: деревья решений, методы опорных векторов, использование нейронных сетей, байесовские методы. Так же интересным моментом было знакомство с латентным семантическим анализом и применением его для автоматического индексирования корпуса текстов и проведения лингвистического анализа. Весь проанализированный материал будет в дальнейшем использован для написания практической части ВКР. В которой будет рассмотрены такие понятия, как совместная встречаемость, будут исследованы пути выявления концептуальных связей и будет осуществлен поиск примеров. Основная проблема, которая будет рассмотрена – отношения между слова, понятия синонимии и антонимии.

ГЛАВА 3 АНАЛИЗ СЛОВ-МАРКЕРОВ ДЛЯ ИССЛЕДОВАНИЯ ТОНАЛЬНОСТИ ТЕКСТА

3.1 Лексические значения слов-маркеров

В практической работе был использован корпус текстов. Корпус текстов состоит из рецензий на научные статьи и книги на разную тематику. Так среди тем научных статей есть: изучение дискурса, изучение метафор, логическое и визуальное представление информации, представлены рецензии на научные журналы и книги о логике, языке информации, так же о мире животных и многих других научных тем научного познания.

В ходе работы с рецензиями, одним из важных аспектов является, какого рода информацию данная рецензия несет. Представляет ли она собой положительные отзывы или отрицательные. Так, по косвенным и прямым признакам были изучены разные уровни языка и извлечены элементы, которые указывают на ту или иную характеристику.

Уровни изучения языка для извлечения информации, для дальнейшей ее оценки были следующие: лексический, морфологический, лексико-морфологический. Так же при взаимодействии с различными уровнями языка был использован метод компонентного анализа и метод семантических полей. Два данных метода, позволяют понять более полную картину представленной в рецензии характеристики и определить является слово, предложение или суффикс, положительным или отрицательным маркером.

Первый и наиболее объемным уровнем языка, на котором проводилось исследование был лексический уровень языка. Отдельные слова и части речи, в особенности определения (прилагательное, наречие и другие слова выступающие в роли определения).

Для более полного разбора слов по частям речи и по их принадлежности к той или иной группе, необходимо построить таблицу слов характеристик.

В большинстве случаев характеристика в рецензии не зависимо от того положительная она или отрицательная представлена с помощью

прилагательного и наречий интенсификаторов. Так, в английском языке представлены группы слабых и сильных прилагательных (weak and strong adjectives). Прилагательные данного вида так же могут называться градуируемыми и не градуируемыми (gradable and non-gradable adjectives).

Первая группа обозначает качества, которыми можно владеть в большей или меньшей степени, такие прилагательные могут образовывать сравнительную и превосходную степени сравнения. Приведем примеры: short — shorter — shortest — ridiculously short, shocked — deeply shocked и т.д. В свою очередь неградуируемые или абсолютные прилагательные (absolute adjectives) выражают признаки, которые не могут иметь большей или меньшей степени: dead, wooden, red.

У прилагательных и наречий существует комплект значения. Слова, составляющие комплект значения того или иного прилагательного или наречия, могут быть определены с помощью семантического анализа слова и поиска семантических полей.

К примеру, возьмем характеристику на статью из рецензии

1) Handily.

В русском языке слово Handily передается как ловко, умело, удобно. Сделать определенный вывод будет ли слово относится к положительной и отрицательной характеристике не представляется возможным. Однако если изучить компоненты значения и синонимы данного слова, можно будет сделать определенные выводы. Так же для определения тональности взятого для примера слова можно посмотреть, как оно используется в контексте предложения. “Koller also handily shows the reader that sports metaphors are linked to aggressive competition and war." Исходя из контекста и при его передаче на русский язык до конца невозможно определить является слово положительной или отрицательной характеристикой.

Рассмотрим синонимы слова Handily. Синонимами выступают такие слова, как dexterously, skillfully, easily. Данный синонимичный ряд более полно раскрывает значение сова handily. Они несут значения “мастерски,

умело". Так же данные слова выступают составляющими семантического поля слова. Все эти признаки, которые были найдены в ходе проведения структурного и семантического анализа слова дают основания полагать что оно несет в себе положительную характеристику [16].

3.2 Функционирование в предложении слов-маркеров

Рассмотрим пример из рецензии на научную статью "Journal of Logic, language, and Information", автор Flow J. van Eijick and A. Visser [17].

Some chapters are genuine representation overviews, but some other chapters seemed to be more a development of the author's personal research than a presentation suitable for general consumption.

В данном примере необходимо рассмотреть, как положительную, так и отрицательную характеристику на статью, представленную в рецензии.

Genuine-sincere, true, real.

Рассмотрим конструкцию со словом equally:

LaCastro offers a clear answer to the first question. In an answer to the second, she is equally explicit.

Раннее в работе уже было рассмотрено слово clear, которое несет в себе положительную характеристику. Необходимо рассмотреть его в комбинации с наречием equally. Для этого нужно рассмотреть какие значения несет в себе слово equally. В русском языке данная конструкция передается, как "в равной степени".

Согласно, Oxford Dictionary "equally" является наречием. Meaning- in the same manner or to the same extent. Данная характеристика помогает понять читателю научной статьи что, в следующем предложении будет содержаться положительная характеристика научного труда, которая представляется в рецензии LaCastro.

Благодаря, синонимичному ряду и проведенную с его помощью компонентному анализу, можно сделать вывод что слово "genuine", несет в себе положительную характеристику. Следовательно, первая часть данного

предложения является положительным маркером. Между частями стоит союз (but), сама по себе вторая часть данного предложения не несет в себе ни отрицательную ни положительную характеристику. Рассмотрим союз but подробнее.

Согласно, Oxford Dictionary: But (conj.) – used to introduced an added statement, usually something that is different from what was said before, contrasting opinion.

Из определения можно сделать вывод, что после союза but, который может передаваться на русский язык, как (но), следует часть предложения, которая обозначает что-либо противоположное или отличное от того что было сказано в первой части предложения. Так как в первой части предложения было определенно, что слово “genuine” является положительным маркером и несет в себе положительную характеристику, то вторая часть предложения, которая если бы перед ней не стоял союз but была бы нейтральной, приобретает отрицательную характеристику и становится отрицательным маркером.

Конструкция as already observe- в рецензии на статью “Applied Linguistics” рассмотренную Richard F.Young [20].

As already observed, the breadth of Young’s review of literature an and around language learning and teaching in this book is formidable.

Already- before or by now or the time in question.

Observe-notice or perceive (something and register is as being significant); take note of or detect (something) in the course of a scientific study. На русский язык передает как “как уже было сказано”.

Так как в рецензии до этого было указана положительная характеристика, с помощью конструкции “as already observed” становится ясно, что последующее предложение будет так же нести положительную характеристику.

Конструкции as well as — в рецензии на статью “Applied Linguistics” рассмотренную Richard F.Young.

Рассмотрим предложение из рецензии: His descriptions of these studies are thorough as well as accurate and he provides reference to other studies that exemplify the approaches.

Thorough – detailed and careful, complete, very great, or very much (Oxford Dictionary).

As well as – in addition; and also (Oxford Dictionary)

Так как конструкция, as well as подтверждает и согласует предыдущую информацию с последующей, то можно сделать вывод что, во второй части предложения будет сохраняться положительная характеристика. Мы можем доказать это рассмотрев слово “accurate” Accurate- correct, exact, and without any mistakes.

Конструкция on the contrary.

On the contrary, I find it a very interesting contribution to our knowledge.

On the contrary – used to show that you think or feel the opposite of what has just been stated.

В русском языке конструкция “on the contrary” передается как “наоборот, напротив, в противоположность этому, даже наоборот, напротив того”, следовательно, она опровергает то что было сказано до этого. Данная конструкция может опровергать как положительную характеристику, так и отрицательную. В предложенном примере, необходимо рассмотреть последующую за конструкцией часть предложения и определить какую характеристику она несет. Даже не рассматривая предыдущего предложения можно будет определить его тональность и важность в рецензии.

Рассмотрим конструкцию “a very interesting contribution”. Необходимо провести компонентный анализ и найти семантические поля для каждого слова-участника данной конструкции чтобы определить тональность всего высказывания.

Very – 1.used to add emphasis to a noun) exact or particular (Ox. Dictionary);

2. used to describe or emphasize the furthest point of something.

Used to show the exact point or furthest point of view. (Усиливает значение последующего за ним прилагательного.)

Interesting – attract, wonder, motivated, exciting, fascinating)-синонимичный ряд

Interesting – holding ones attention (adj.)

Contribution – indemnity, assistance, promote, fee, donation-синонимичный ряд.

Contribution – something that you contribute or do to help produce or achieve something together with other people, or to help make something successful:

Исходя из значения все составляющих компонентов конструкции, можно сделать вывод что “a very interesting contribution”, которая передается на русский язык как “очень интересный вклад” несет в себе положительную тональность и является положительным маркером [19].

Исходя из этого можно сделать вывод что так как конструкция on the contrary, опровергает то что было сказано в предыдущем предложении.

Данный вывод можно проверить компонентным анализом предложения, которое предшествует данному.

The foregoing remarks, pertinent or not, may give the impression I do not like the book.

Do not – конструкция отрицания

Like – want, prefer, wish, attract

Рецензент утверждает, что его замечания, которые были высказаны в ходе рецензии могут создать впечатление, что ему не нравится данная книга. Но они лишь могут создать впечатление, но на самом деле не являются выводами и определением рецензии как положительной или отрицательной. Это автор и опровергает в последующем предложении, с использованной упомянутой конструкции.

Конструкция but может так же содержать двойное отрицание. В рецензии на статью Hammer, but используется в сочетании с частицей no. Рассмотрим пример из рецензии: Thus, Hammers book does not fulfil its remit, but it is no

small achievement. Слово Thus, обозначающее “таким образом”, дает основания полагать, что автор подводит итоги в своей рецензии и в данном предложении будет описывать положительные и отрицательные моменты статьи.

Does not – слово сочетание представляющее собой отрицание.

Fulfil – to do something that is expected, hoped for, or promised, or to cause it to happen.

В первой части предложения автор использует отрицательною характеристику рецензии, что подтверждает словосочетание does not fulfil.

Во второй части предложения, которая отделена союзом but содержится положительна характеристика рецензии. Однако данная характеристика выражена не часто употребляемой и свойственной конструкцией с двойным отрицание. С помощью компонентного анализа, поиска синонимов и изучения значения частей конструкции в словаре докажем это.

But – used to introduce an added statement, usually something that is different from what you have said before. Conjunction. Синонимичный ряд для компонентного анализа- however, nevertheless, and, though.

No – used to give negative answers. Синонимичный ряд: nope, nether, nix.

Small – not very important or not likely to cause problems (Oxford Dictionary).

Синонимичный ряд: little, low moderate, narrow, weak, feeble, faint. (Adjective)

Achievement – something very good and difficult that you have succeeded in doing.

Синонимичный ряд: accomplishment, progress, breakthrough, effort, success, luck, hit, achievement, advance, prosperity.

Таким образом, в ходе компонентного анализа и изучения значения слов в словаре Oxford Dictionary, можно сделать вывод что в данной конструкции во второй части предложения, but. no, используется для приведения положительной характеристики. Следовательно, тональность предложения находится под влиянием синтаксической структуры предложения и его

компонентов. Данный анализ предложения так указывает на то, что положительными маркерами могут являться и слова, которые несут отрицательное значение, если они используются в одной конструкции.

Положительная тональность высказывания может быть определенна как с помощью структуры предложения, и с помощью его составляющих частей и их разбора, компонентного анализа, поиска значения в словаре и более глубоко изучения их внутренней структуры.

Так же необходимо сгруппировать слова извлеченные, из корпуса научных статей описанные в рецензиях по частям речи. Одной из самых многочисленных групп будут прилагательные. Они более полно и четко раскрывают мысли и взгляды автора рецензии на рассматриваемую им статью.

Второй по численности в ходе изучении текстов являлась группа причастий.

Для определения численности слов в данном корпусе научных рецензий был использована программа подбирающая список частотности в алфавитном и обратном алфавитном порядке.

У наречий и прилагательных существует комплект значения. Комплект значения можно определить с помощью компонентного анализа и поиска семантических полей. За счет изучения синонимичного ряда слов можно определить значение слова более полно и определить его структуру.

Наречия можно разделить на две группы: обстоятельственные и определительные. Обстоятельственные в свою очередь подразделяются на наречия: времени, места, цели причины. Определительные же на качественные, способа и образа действия и меры, и степени.

Для более детального разбора характеристики и тональности текста необходимо подразделить извлеченные слова-маркеры по группам.

1. Strength – a good or beneficial quality or attribute of a person or thing.

2. Very detailed account:

1) very – in a higher degree,

2) detailed – having many details or facts showing attention to detail; executed with many minor features.

3. Wide range of insights.

1) wide (adj) – to the full extent;

2) range – the area of variation between upper and lower limits in a particular scale); (the ability to have) a clear, deep, and sometimes sudden understanding of a complicated problem or situation)

4. Makes it clear – 1) clear – easy to perceive, understand and interpret; leaving no doubt, obvious, unambiguous.

5. Moves us well forward in our understanding of this –1) moves to (cause to) progress, change, or happen in a particular way or direction; 2)forward (adjective) future; relating to the future [15].

6. Handily shows (handily, обращает внимание на положительную характеристику, синонимы: dexterously, skillfully, easily)

7. Kollner provides a detailed outline

Следующие характеристики, которую автор вводит в своей рецензии на научную статью. Производился анализ семантических полей слов-маркеров. Изучались компоненты значения слов, выявлялась их положительная или отрицательная конотация.

1. What the author leaves unclear (not clear, vague, uncertain, and fuzzy)

Not obvious or easy to see or know

2. Another strong aspect of the book-strong – effective; of a good quality or level and likely to be successful:

3. Her conclusion, which analyzes the negative effect and suggest more positive alternatives.

1) suggest – to mention an idea, possible plan, or action for other people to consider

2) more – used to form the comparative of many adjectives and adverbs

Positive – certain and without any doubt

Alternatives – something that is different from something else, especially from what is usual, and offering the possibility of choice [16].

1) Reasonable – based on or using good judgment and therefore fair and practical.

2) The first chapter, by the Editors, gives an excellent overview of the work to follow in the rest of the book.

3) Excellent (adjective) – extremely good

Some chapters are genuine overviews, but some others chapters, but some other chapters seemed to be more a development of the authors personal opinion than a presentation suitable for general consumption.

Genuine-sincere, true, real

Suitable (adj.) – acceptable or right for someone or smth, само по себе несет положительную характеристику, но приобретает оттенки отрицания из за предшествующей конструкции (but) [17].

1) Makes a bold attempt – совершить смелую попытку (подлежащее, сказуемое, дополнение) положительный маркер, так же может иметь оттенок отрицательный, согласно Oxford Dictionary: 1. not frightened of danger

2) In spite of its weaknesses, Pinker's Words and Rules is a significant contribution to the field of cognitive science, and is a valuable resource book for every scholar striving to unravel the mysteries of the human mind.)

1) Significant – important or noticeable; contribution- something that you contribute or do to help produce or achieve something together with other people, or to help make something successful [20].

3.3 Морфологическая структура слов-маркеров

Работа с корпусом текстов проводилась на разных уровнях языка. Одним из наиболее важных в раскрытие значения слов является морфологический уровень. Главным приемом в ходе работы со словами, извлеченными из корпуса текстов являлся разбор слов на его составляющие части, такие как суффикс, корень, приставка, окончание. Одним из важнейших составляющих

в слове, при изучении его значения и определения его в положительную или отрицательную характеристику является суффикс. Были изучены суффиксы слов и значения, которые они могут принимать [22].

Суффикс Un.

Данный суффикс в большинстве случаев придает слову отрицательную окрасу и характеристику. Например, в рецензии на научную работу Вероники Келлер, обозреватель использует следующую конструкцию описывая одну из частей работы автора: What the author leaves unclear.

Словом-маркером в определении значения все конструкции выступает прилагательное. В данном случае – unclear. Слово состоит из двух компонентов. Это префикс и корень. В предыдущих частях, уже делался акцент на слово “clear” и его положительную характеристику, опираясь на Oxford Dictionary. Данное слово часто используется в научных статьях, трудах и книгах для освещения материала в позитивном свете. Согласно Oxford Dictionary (clear- certain, having no doubt, or obvious), на русском языке может передаваться как “четкий, ясный, понятный, не вызывающий сомнений”. Далее рассмотрим суффикс “un” (used to add the meaning "not", "lacking", or "the opposite of" before adjectives, adverbs, verbs, and nouns). Согласно Оксфордскому словарю префикс “un” одержит в себе отрицательное значение, следовательно, все слово приобретает негативную окраску, что говорит нам о отрицательные характеристики в рецензии одной из частей научной статьи написанной автором.

В еще одной статье используется конструкция: “There are uncomfortably many misprints”. Слово “uncomfortably” состоит из двух частей, корня и префикса. Корень “comfortably” несет в себе положительную характеристику. Так согласно Оксфордскому словарю “comfortably”: in a comfortable way (рус. удобным способом, путем). Префикс “un” придает слову отрицательную характеристику, следовательно, все слово приобретает отрицательный окрас.

Префикс mis.

Рассмотрим пример из рецензии на научную статью “Journal of Logic, language, and Information”, автор Flow J. van Eijick and A. Visser [21].

There are uncomfortably many misprints.

Mis — ассоциируется с чем-то «неправильным», «ошибочным», т. к. чаще всего она именно в этом значении употребляется, mis- работает только с существительными, глаголами и причастиями:

Misfortune – неудача; misdial – набрать неправильный номер; misleading – вводящий в заблуждение.

Misprint — a mistake, such as a word that is spelled wrong, in a printed text (Oxford Dictionary)

Так как, приставка (mis) несет в себе значение чего-либо ошибочного, следовательно, слово misprint будет отрицательным маркером в данной рецензии.

Суффикс able.

На примере слова formidable, используемого в рецензии на статью “Applied Linguistics” рассмотренную Richard F.Young

Formidable- causing you to have fear or respect for something or someone because that thing or person is large, powerful, or difficult

Суффикс -able, -ible характеризует свойство либо доступность для какого-либо действия:

Edible — съедобный

Portable — переносной, портативный

Admirable — вызывающий восхищение

Данный уровень так же необходим к рассмотрению. Он показывает скрытые характеристики языка. За счет этого будет достигаться новизна данной работы. В ходе изучение стилистического уровня языка, было определенно, что слова, которые чаще всего несут положительную характеристику, под влиянием структуры предложения и его частей могут изменять свое значение. Этого может происходит и наоборот, когда отрицательные слова приобретают положительную характеристику.

Рассмотрим предложение “The sustained focus on the interface between pragmatics and sociolinguistics, provides book with its undoubtness strengths, as I will discuss below”

Undoubtness strengths – неоспоримая сила работы.

Как ранее было рассмотрено в работе суффикс “un” представляет отрицательную характеристику той или иной работы. В данном случае рассматривается пример, когда суффикс приобретает и придает положительную характеристику в контексте всего высказывания. Для того, чтобы доказать данную гипотезу, разберем синонимичный ряд и семантическое поле слова “undoubtness” и сделаем его морфологический разбор.

“Undoubtness” – (not doubted, genuine, undisputed). Примеры использования в других конструкциях “the undoubted truth, an undoubted friend”. Данные примеры доказывают, что данная характеристика статьи является положительной. Соответственно, можно говорить о том, что в некоторых контекстах суффикс “un” может представлять в совокупности всего слова положительную характеристику.

Основным выводом, который можно сделать по работе с лексико-морфологической группой является то, что все слова можно разбить на две группы. Первая группа - это слова, содержащие в себе положительную характеристику, и используемые без префиксов и суффиксов. Вторая группа, слова, используемые с префиксами и суффиксами. Слова из обеих групп выражают положительную характеристику в рецензиях. Но их можно разбить на две группы по силе высказывания. Для примера, возьмем слова unclear и vague. Слово, используемое без суффикса, несет в себе более усиленную характеристику статьи. Следовательно, когда автор использует его он вкладывает более интенсифицируемую характеристику и делает акцент на характеристиках, которые описывает.

Если разбить слова на две группы по силе высказывания, получится создать более совершенный метод извлечения ключевых слов из корпуса

текстов. Так как конструкции с более углубленной и точной характеристикой будут оказывать больше влияния на восприятие текста в целом, а, следовательно, они будут занимать более высокие позиции в перечне ключевых слов корпуса текстов.

Выводы по главе 3

Лексический уровень, на котором проводилась основная часть исследований оказался самым большим по значению, для всей работы. За счет рассмотрения данных уровней была достигнута новизна работы, и решены поставленные цели и задачи. Были собраны комплекты значения слов с помощью разного рода словарей. Найдены слова и контексты, которые можно определить под другим углом, приобретающие в контексте новые значения. Данные исследования помогли определить тональность как отдельных компонентов текста, так и в целом тональность и отношение авторов к рецензиям.

ЗАКЛЮЧЕНИЕ

В работе было проведено исследование на предмет извлечения информации из корпуса научных текстов. Базу, использующуюся в работе, составили корпусы научных рецензий на научные статьи, написанные на английском языке. Ключевым понятием, используемым в практической части, являлась тональность отдельных компонентов корпуса текстов. При работе с корпусом рецензий использовались методы сплошной выборки, а также метод компонентного анализа языка. Слова-маркеры извлекались из корпуса научных статей на основе их основных признаков, положительных или отрицательных.

Основными единицами языка через которые проводилось исследование и распознавалась тональность текста, являлись прилагательные и наречия. С помощью построения списка частотности встречаемости слов в корпусе рецензий в алфавитном и обратном алфавитном порядке была определенна и характеристика, которая приводилось в рецензиях.

Частотность встречаемости прилагательных и наречий повлияла и на все исследование в совокупности. Лексический уровень, на котором проводилась основная часть исследований оказался самым большим по значению, для всей работы.

Также, в работе рассматривались синтаксический уровень, морфологический, и смежные уровни, такие как лексико-морфологический. За счет, рассмотрения данных уровней, была достигнута новизна работы, и решены поставленные цели и задачи.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Антрушина, Г. Б. Лексикология английского языка: учебное пособие для студентов / Г.Б. Антрушина, О.В. Афанасьева, Н.Н. Морозова. – 3-е изд., стереотип. – М. : Дрофа, 2001. – 288 с.

2. Арнольд, И.В. Стилистика. Современный английский язык : учебник для вузов / И.В. Арнольд. – 9-е изд. – М. : Флинта: Наука, 2009. – 384 с.

3. Арутюнова, Н.Д. Язык и мир человека / Н.Д.Арутюнова. – М. : Языки русской культуры, 1999. – 896 с.

4. Булаховский Л. А. Введение в языкознание / Л. А. Булаховский, А.С. Чикобава. — М.: Учпедгиз. – 1952. – С.163–166

5. Вараксин, Л. А. Однокорневые префиксальные глаголы, антонимы в современном русском языке: автореф. дис. … канд. филол. наук : 19.03.2013 / Варакасин Леонид Анатольевич. – Куйбышев, 1970. – 22 с.

6. Иванова, В.А. Антонимия в системе языка [Текст] / В.А. Иванова. Кишиневский Государственный Университет, 1982. – С.163– 166.

7. Кабыш, М.Ю. Фоносемантическая тональность как способ выразительности в поэтических произведениях первой половины XX века / М.Ю. Кабыш // Austrian Journal of Humanities and Social Sciences. – 2010. – 77–81 с.

8. Кормалев, Д.А. Технология извлечения информации из текстов, основанных на знаниях [Текст] / Д.А. Кормалев, Е.П. Куршев, И.В. Трофимов. – М. : Программные продукты и системы. – 2009. – №5. – С. 62–66.

9. Максимов, Л. Ю. Антонимия как один из показателей качественности прилагательных [Текст]: (На материале рус. яз.): автореферат дис. …к. фил.н. / Моск. гос. пед. ин-т им. В. И. Ленина. – Москва, 1958. – 19 с.

10. Москвитина, Т.Н. Ключевые слова и их функции в научном тексте / Т.Н. Москвитина // Вестник Челябинского государственного педагогического университета. – 2009. – №11. – С. 277–283.

11. Самойлик, Е. Е. Оценочные определения как средство экспликации оппозиции «свой» / «чужой» в англоязычном политическом дискурсе/ Е. Е. Самойлик // Вес. Мин. гос. лингв. ун-та. Сер. 1, Филология. – 2008. – №3. – С. 134 – 141.

12. Седова, Я.А. Системный анализ корпуса текстов научного знания / Я.А. Седова, И.Ю. Квятковская // Вестник саратовского государственного технического университета. – 2010. – Т. 4, №2(50). – С. 196–203.

13. Солодовникова, А. Н. Тональность текста социальной рекламы / А. Н. Солодовникова // Вестник Нижегородского университета им. Н.И. Лобачевского. – 2011. – Ч. 2. – Т. 2, №6. – С. 652–655.

14. Шанский, Н. М. Лексикология современного русского языка / Н.М Шанский // Учебное пособие. — 4-е изд., доп. — М.: Изд- во ЛИБРОКОМ. – 2009. – 312 с.

15. Шкуропацкая, М.Г. Семантическое поле и проблемы синонимии [Текст] / М.Г. Шкуропацкая, Н.В. Цепелева // Вестник Кемеровского государственного университета. – 2012. – С. 233–240.

16. Шмелёв, Д. Н. Современный русский язык. Лексика / Д.Н. Шмелев. – М.: Просвещение, 1977. – 335 с.

17. Blench, R. The Origins and Development of African Livestock: Archaeology, Genetics, Linguistics and Ethnography. – L.: University College London Press, 2000. – 546 p.

18. Fabiszak, M. Language and Meaning: Cognitive and Functional Perspectives. – Frankfurt am Main: Peter Lang, 2007. – 344 p.

19. Filipec, J. Ceska Synonyma z hlediska stylistiky a lexikologie. – Praha: CSAV, 1961. – 383 s.

20. Geckeler, H. Structurelle semantik und Wortfeldtheorie. – Munchen: Fink, 1971. – 178 s.

21. Hyland, K. Disciplinary Identities: Individuality and Community in Academic Discourse. – Cambridge: Cambridge University Press, 2012. – 238 p.

22. Koller, V. Metaphor and gender in business media discourse: A critical cognitive study. – Hampshire: Palgrave Macmillan, 2004. – 260 p.

23. LaCastro, V. Pragmatics for language educators: a sociolinguistic perspective. – NY.: Routledge, 2012. – 231 p.

24. Lyons, J. Structural Semantics. – Oxford: Blackwell, 1963. – 245 p.

25. Pinker, S. Book review words and rules: the ingredients of language Linguistics and Language Development Department. – Jose: Jose State University, 1999. – 380 p.

26. van Eijck, J. Journal of Logic, Language, and Information. – Cambridge: Foundation of Computing Series, – 1997. – 343 p.

27. Young, F. Discursive practice in language learning and teaching. – Oxford: Oxford University Press, 2012. – 626 p.

Словари и энциклопедии:

28. Большая Советская Энциклопедия: В 30-ти т./ гл. ред. А. М. Прохоров.– Изд. 3-е. – М.: Сов. Энциклопедия, 1970 – 1978. – 630с.

29. Комиссаров В.Н. Словарь антонимов современного английского языка. - М.: "Международные отношения", 1964. – 538 с.

30. Ахманова, О.С. Словарь лингвистических терминов. –М.: Советская энциклопедия, 1969. — 608 с.