Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)
САЕ «Институт человека цифровой эпохи»
Автономная магистерская программа
«Компьютерная и когнитивная лингвистика»

ДОПУСТИТЬ К ЗАЩИТЕ В ГЭК
Руководитель ООП
д-р филол. наук, профессор
_____ З. И. Резанова
подпись
«10» июня 2022 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТРА
(МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ)

РАЗГРАНИЧЕНИЕ ПОЛИСЕМИИ И ОМОНИМИИ ДИСКУРСИВНОГО СЛОВА ***ВОТ*** С ПОМОЩЬЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ (НА МАТЕРИАЛЕ БИМОДАЛЬНОГО КОРПУСА УСТНОЙ РЕЧИ ТЮРКСКО-РУССКИХ БИЛИНГВОВ RuTuBic)

по направлению подготовки 45.04.03 – Фундаментальная и прикладная лингвистика

Жильцова Наталья Владимировна

Руководитель ВКР
канд. филол. наук, доцент
_____Ю.В. Филь
подпись
«10» июня 2022 г.

Консультант
науч. сотр. лаборатории
лингвистической антропологии
_____А. А. Степаненко
подпись
«10» июня 2022 г.

Автор работы
студент группы № 132083
_____ Н.В. Жильцова
подпись
«10» июня 2022 г.

Томск – 2022

Министерство науки и высшего образования Российской Федерации
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (ТГУ)
САЕ «Институт человека цифровой эпохи» Автономная магистерская программа «Компьютерная лингвистика»

УТВЕРЖДАЮ
Руководитель АМП
доктор филол. наук,
профессор,зав. кафедрой
_____З.И. Резанова
«01» февраля 2022 г.

ЗАДАНИЕ
по подготовке магистерской диссертации студенту
Жильцовой Наталье Владимировне, группы 132083

1. Тема магистерской диссертации: «Разграничение полисемии и омонимии дискурсивного слова вот с помощью методов машинного обучения (на материале бимодального корпуса устной речи тюркско-русских билингвов RuTuBiC)».
2. Срок сдачи студентом выполненной магистерской диссертации:
   a) на кафедре: 17.06.2022 г.
   b) в ГЭК: 13.06.2022 г.
3. Исходные данные к работе:
   Объект исследования: функционально-семантические особенности дискурсивого слова *вот*.
   Предмет исследования: контекстные употребления дискурсивного слова *вот* в материалах корпуса устной речи тюркско-русских билингвов.
   Цель: выявление эффективного метода машинного обучения для решения задач автоматического разграничения полисемии и омонимии слова *вот*.
   Задачи 1-го этапа:
   1) проанализировать семантическую структуру дискурсивного слова *вот*, в том числе определить его ядерные и периферийные значения / оттенки значений;
   2) рассмотреть и систематизировать контексты, в которых употребляется данное слово в корпусе устной речи тюрско-русских билингвов;
   3) предложить параметры, позволяющие разграничить значения слова *вот* и функции, которые оно выполняет.

   Задачи 2-го этапа:
   1) сформировать корпус текстов, содержащих слово *вот*;
   2) создать алгоритм, автоматически определяющий, какое из значений *вот* используется в том или ином случае;
   3) обучить модель для разрешения омонимии и полисемии слова *вот*;
   4) представить результаты работы классификатора на новых, незнакомых модели контекстах.

   Использованные группы методов: общенаучный - описательный (сбор, анализ и синтез материала); лингвистические: структурно-семантический, реализованный через дефиниционный, дистрибутивный и контекстный анализ; метод машинного обучения (библиотека fastText, использующая нейронную сеть для встраивания слов)

4. Краткое содержание работы:

В магистерской диссертации представлены теоретические основания работы, включающие описание дискурсивных маркеров и их классификаций по разным параметрам, описание возможных подходов к разрешению омонимии и полисемии при автоматической обработке текста (Глава 1). Во второй главе произведён анализ семантических особенностей слова. Также данная глава посвящена анализу контекстов, полученных сплошной выборкой из корпуса тюркско-русских билингвов, проанализировано, в каких значениях чаще используется слово «вот», какие выражения с данным словом присутствуют в речи билингвов, каковы контекстуальные особенности данной единицы в корпусе. Предложены параметры, по которым можно разграничить омонимию и полисемию данного слова. Глава 3 посвящена работе с библиотекой FastText, созданию кода, позволяющего разграничить полисемию и омонимию слова. В заключении подводятся итоги работы и определяются перспективы дальнейшей разработки. В заключении подводятся итоги работы и определяются перспективы дальнейшей разработки.

5. Предприятие/организация, по заданию которого выполняется работа: Томскийгосударственный университет.
6. Дата выдачи задания: «01» февраля 2022 г.

Руководитель магистерской диссертации,
канд. филол. наук, доцент / Филь Ю.В.

Задание принял к исполнению / Жильцова Н.В.

АННОТАЦИЯ

Данное исследование посвящено анализу значений и функций дискурсивного слова *вот* и направлено на создание алгоритма, с помощью которого можно автоматически разграничить полисемию и омонимию данного слова.

Работа состоит из введения, трех глав, заключения и списка использованной литературы.

Первая глава представляет теоретическую базу исследования, основанную на описании существующих классификаций и наименований дискурсивных маркеров, явления хезитации и классификаций вербальных сбоев в устной речи. Вторая часть главы отводится на характеристику существующих способов разграничения полисемии и омонимии слов во время автоматической обработки текстов.

Во второй главе рассматриваются причины, по которым отбиралось исследуемое слово из ряда других дискурсивных слов и содержит описание семантических особенностей *вот*, основанное на данных лексикографических источников и записях устной речи тюркско-русских билингвов. В главе также описаны функции дискурсивного слова *вот* и правила, по которым можно разграничить значения слова и его функции.

В третьей главе представлен ход работы по разработке алгоритма машинного обучения, способного автоматически разграничить типовые значения и функции дискурсивного маркера *вот*.

В заключении обобщаются полученные результаты, подводятся итоги исследования и рассматриваются перспективы его развития.

Оглавление

ВВЕДЕНИЕ 5
1. ДИСКУРСИВНЫЕ СЛОВА В АСПЕКТЕ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА 10
1.1 Дискурсивные слова как объект лингвистического исследования 10
1.2 Автоматическое разрешение омонимии и полисемии 16
Выводы к главе 1 21
2. ФУНКЦИОНАЛЬНО-СЕМАНТИЧЕСКИЕ ОСОБЕННОСТИ ДИСКУРСИВНОГО СЛОВА *ВОТ* 23
2.1 Характеристика частотности слова *вот* 23
2.2 Анализ семантических особенностей дискурсивного слова *вот* 24
2.3 Функции дискурсивного слова *вот* 35
2.4 Анализ контекстов употребления *вот* в бимодальном корпусе устной речи тюркско-русских билингвов 38
Выводы к главе 2 41
3. СНЯТИЕ ОМОНИМИИ И ПОЛИСЕМИИ С ПОМОЩЬЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ 43
Выводы к главе 3 43
ЗАКЛЮЧЕНИЕ 49
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ И ЛИТЕРАТУРЫ 51

ВВЕДЕНИЕ

Дискурсивные маркеры считаются достаточно частым объектом лингвистических исследований, интерес к которым объясняется в первую очередь когнитивно-дискурсивной направленностью современного языкознания и широкими функциональными возможностями указанных единиц. Как правило, основными функциями дискурсивных маркеров являются обеспечение смысловой целостности коммуникации, выражение эмоционального отношения говорящего к ситуации, а также регуляция процесса понимания сказанного. Эти функции позволяют говорить о «направленности» дискурсивных маркеров либо на самого говорящего (выражение эмоционально-оценочного отношения), либо на адресата (маркеры регулируют процесс понимания), либо на содержательную сторону высказывания или его языковое оформление. В этих случаях дискурсивные маркеры обеспечивают когерентность дискурса либо используются как метакоммуникативные средства. Яркими особенностями данных единиц являются их многозначность (в том числе использование как неполнозначных единиц – хезитативов) и зависимость от контекста (в некоторых случаях при этом отмечается «мерцательность» передаваемых ими смыслов), что ведет к отсутствию единых критериев их классификации и осложняет их типологизацию. В силу сказанного дискурсивные маркеры, с одной стороны, являются важным коммуникативным индикатором, а с другой стороны, представляют собой проблему для автоматической обработки естественного языка, что значимо для корпусной лингвистики.

В рамках данного исследования предполагается создание системы, автоматически отличающей хезитативное употребление дискурсивного слова от нехезитативного (разграничение омонимии), а также отличающей типичные значения слова при его нехезитативном употреблении (разграничение полисемии). Исследование проводится на материале корпуса устной речи тюркско-русских билингвов.

Бимодальный корпус устной речи тюркско-русских билингвов Южной Сибири (RuTuBiC)» создается в рамках проекта «Языковое и этнокультурное разнообразие Южной Сибири в синхронии и диахронии: взаимодействие языков и культур». Цель корпуса – фиксация типов проявлений русско-тюркской интерференции в системе отклонений от речевого стандарта в речи билингвов, соотнесение их с другими источниками отклонений, социолингвистическое портретирование билингва, фиксация его языкового опыта.

Текстовый материал корпуса – записи устных интервью информантов, носителей русско-тюркского билингвизма (татарско-русских, шорско-русских, хакасско-русских билингвов), жителей Южной Сибири. Что касается аннотирования корпуса, то в нем представлена морфологическая разметка на основе Mystem, ручная разметка отклонений от речевого стандарта, метаразметка (возраст, пол, образование), которая включает многоаспектную характеристику билингвального опыта информанта, а также тематическая разметка. Однако, несмотря на то что в корпусе представлена автоматическая морфологическая разметка, некоторые случаи частеречной и падежной омонимии, а также полисемии подлежат ручной обработке, поэтому необходимо предложить другие варианты разрешения неоднозначности сложных случаев в разметке.

Таким образом, **актуальность** исследования обусловлена особым вниманием современных учёных к проблеме описания дискурсивных слов и востребованностью исследований и проектов по решению задачи снятия многозначности дискурсивных маркеров в системах машинного перевода, информационного поиска, автоматической обработки текстов.

Новизна исследования определяется обращением к поискам автоматических способов разграничения полисемии одного из частотных дискурсивных слов – *вот* и способов разграничения омонимии данного слова на новом корпусном материале через описание и систематизацию семантических и контекстных особенностей указанной единицы, а также

создание алгоритма, позволяющего решить указанную проблему на материале корпуса устной речи тюркско-русских билингвов.

Объектом исследования являются функционально-семантические особенности дискурсивого слова *вот*.

Предмет исследования – контекстные употребления дискурсивного слова *вот* в материалах корпуса устной речи тюркско-русских билингвов.

В качестве объекта изучения в настоящей работе выступает явление омографии в русском языке. Предметом исследования являются контекстные употребления омографов в русском языке. Цель данной работы состоит в выявлении эффективного метода машинного обучения для решения задач автоматического снятия омографии.

Цель – выявить и предложить эффективный метод машинного обучения для решения задач автоматического разграничения полисемии и омонимии дискурсивного слова *вот*.

Поставленная цель предполагает решение следующих **задач**:

– проанализировать семантическую структуру дискурсивного слова *вот*, в том числе определить его ядерные и периферийные значения / оттенки значений;

– рассмотреть и систематизировать контексты, в которых употребляется данное слово в корпусе устной речи тюркско-русских билингвов;

– предложить параметры, позволяющие разграничить значения слова *вот* и функции, которые оно выполняет;

– сформировать корпус текстов, содержащих слово *вот*;

– создать алгоритм, автоматически определяющий, какое из значений *вот* используется в том или ином случае;

– обучить модель для разрешения омонимии и полисемии слова *вот*;

– представить результаты работы классификатора на новых, незнакомых моделиконтекстах.

Исследование носит междисциплинарный характер, поэтому среди

методов, используемых в магистерской диссертации, представлены следующие.

- Общенаучные: описательный;
- Лингвистические: структурно-семантический, реализованный через дефиниционный, дистрибутивный и контекстный анализ;
- метод машинного обучения (библиотека fastText, использующая нейронную сеть для встраивания слов).

Многозначность рассматриваемого дискурсивного слова потребовала обращения к разнообразным лексикографическим источникам для определения всего спектра значений и оттенков значения *вот*, примерам его употребления. **Материал исследования** составили дефиниции русского слова *вот* и иллюстративный словарный материал, представленные в толковых словарях русского языка, Русской грамматике, а также тексты, извлеченные из корпуса устной речи тюркско-русских билингвов.

Источники материала. Материал для исследования собирался путем сплошной выборки из следующих источников: Большой академический словарь русского языка, Малый академический словарь русского языка, Толковый словарь русского языка Д.Н. Ушакова, Толковый словарь русского языка Т.Ф. Ефремовой, Словарь русского языка С.И. Ожегова, Большой толковый словарь русского языка С.А. Кузнецова; а также «Бимодальный корпус устной речи тюркско-русских билингвов Южной Сибири (RuTuBiC)»

Теоретической основой в работе послужили научные труды в области анализа дискурсивных маркеров, направленные на выявление особенностей дискурсивных слов, их функций, создание типологии дискурсивных маркеров (Н.В., Богданова-Бегларян, А.А. Кибрик, И.М. Кобозева, В.А. Плунгян, Е.В., Рахилина Е.В. и др.); исследования и научные проекты в области автоматической обработки естественного языка, корпусной лингвистики, машинного обучения (Найденова К.А., Невзорова О.А., Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С.)

Значимость исследования заключается во вкладе проводимого исследования в дальнейшее изучение проблемы автоматической обработки естественного языка, связанной с разрешением омонимии и полисемии дискурсивных слов на материале корпуса устной речи тюркско-русских билингвов.

1. Дискурсивные слова в аспекте автоматической обработки естественного языка

1.1 Дискурсивные слова как объект лингвистического исследования

Дискурсивные маркеры являются предметом многих лингвистических исследований, как отечественных, так и зарубежных. Термин «дискурсивные маркеры» применяется к грамматически неоднородному классу языковых единиц, объединенных общими функциональными особенностями. Данный термин не является общепринятым, наряду с ним используются термины «дискурсивные слова» и «дискурсивные частицы» [Баранов, Плунгян, Рахилина, 1993; Малов, Горбова, 2007], «модальные частицы» и «модальные слова» [Schourup, 1999; Дараган, 2000], «логические частицы», «прагматические маркеры» [Schiffrin, 1987; Frazer, 1990], «дискурсивные коннекторы» [Blakemore, 2002], «ключевые маркеры», «прагматические выражения» [Erman, 1987] и др. Разнообразие в терминах является следствием разнообразия подходов к изучению и описанию дискурсивных маркеров и влечёт за собой варьирование объема понятия дискурсивных маркеров от исследования к исследованию. Как следствие, под одним и тем же названием часто фигурируют близкие, не идентичные явления, либо для одного и того же явления используются синонимичные термины.

Первоначально интерес к дискурсивным маркерам возник в рамках изучения дискурса. Дискурс в данном случае понимается как коммуникативное событие, «сложное единство языковой формы, значения и действия», которое включает в себя «социальный контекст, дающий представление как об участниках коммуникации (и их характеристиках), так и о процессах производства и восприятия сообщения» [Дейк, 2000, с. 113]. Одно из первых упоминаний о дискурсивных маркерах как языковых единицах можно обнаружить в работе У. Лабова и Д. Фэншела [Labov & Fanshel, 1977]. Интенсивное изучение этой проблемы началось в 1980-х годах

и продолжается по сей день. Оно представлено самыми разнообразными подходами: когнитивным, функциональным, прагматическим, семантическим. Исследования проводятся на материале разных языков, в рамках описания конкретных дискурсов, на материале корпусов письменных текстов и устной речи.

На данный момент не существует единой, общепризнанной классификации дискурсивных маркеров. Б. Фрейзер рассматривает прагматические маркеры как широкий класс подобных единиц, выделяя дискурсивные маркеры только как один из подклассов, утверждая, что «прагматические маркеры, взятые отдельно от пропозиционального содержания предложения, являются лингвистически закодированными подсказками, которые сигнализируют о потенциальных коммуникативных намерениях говорящего» [Frazer, 1990, с. 386 – 387] В своей книге «Прагматические маркеры» (Pragmatic Markers) он выделяет: 1) базовые маркеры, указывающие на характер базового сообщения (тип речевого акта, перформативы, прагматические идиомы), например, I believe, I hope; 2) маркеры-комментарии, комментирующие базовое сообщение, например, fortunately, sadly, clearly, definitely; 3) параллельные маркеры, которые дополняют базовое сообщение, например, now, so, well; 4) дискурсивные маркеры, указывающие, как базовое сообщение связано с контекстом, например, thus, moreover, however. Дискурсивные маркеры, в свою очередь, делятся на контрастирующие маркеры (but, however, although), детализирующие маркеры (in other words, particularly, also) и маркеры вывода (that's why, so, accordingly) [Frazer, 1996].

М. Халлидей и Р. Хасан, исследуя аналогичные единицы, выделили четыре значения «соединительных единиц» (слов, помогающих создавать текст, указывая на семантические связи в основной понятийной структуре): добавочное (также, кроме того), противительное (однако, с другой стороны), причинно-следственное (в связи с этим, поэтому) и временное (далее, затем) [Halliday, Hasan, 1976]. Данные значения подразделялись на более

конкретные. Так, например, причинно-следственное значение включало в себя общее причинно-следственное значение (простые и эмфатические единицы) и определённое причинно-следственное значение (единицы, обозначающие причину, результат и намерение).

Дебора Шиффрин считает основной функцией дискурсивных маркеров обеспечение текстовой когезии на разных уровнях речи. В основу ее классификации положен тип логического соотношения соединяемых клауз. Она делит дискурсивные маркеры на следующие типы: 1) соединительные (and, but, or); 2) восклицательные (oh); 3) наречные (now, then); 4) лексические фразы (y'know, I mean) [Schiffrin, 1987].

В целом следует заметить, что большинство классификаций дискурсивных маркеров основываются на функциональных, реже – на семантических признаках.

Помимо дискурсивного употребления эти единицы могут употребляться как полнозначная лексика. Выделение значений подобных дискурсивных маркеров в определенной степени субъективно, при этом в словарях может быть представлено разное количество значений. О полисемии дискурсивных маркеров говорят многие современные исследователи. Так, И.М. Кобозева считает, что подобные единицы играют значимую роль при понимании смысла, принципиально важны при переводе. На примере слова *вот* исследователь рассматривает возможность разрешения синтаксической и семантической неоднозначности дискурсивных слов в контексте предложения [Кобозева, 2007], используя анализ клауз и разметки по типу морфологической разметки Национального корпуса русского языка.

Дискурсивные маркеры в качестве основной имеют задачу обеспечения связи между элементами дискурса и выражение отношения говорящего к положению дел или к высказыванию собеседника. Однако существует тип слов, относящихся к дискурсивным маркерам, который, в отличие от простых дискурсивных маркеров, непосредственно связан с самим процессом

говорения и ничего нового в сообщаемое не вносит. Данный тип слов связан с «проблемами» устного дискурса. В устной спонтанной речи возможно возникновение своеобразных «сбоев» вследствие несоответствия некоего фрагмента высказывания изначальному плану. Такой «внутренний перебив» в процессе речепорождения принято назвать хезитацией. Наличие хезитации объясняется тем, что порождение речи рекурсивно, и когда говорящий «планирует» высказывание, он не представляет его себе как окончательный продукт своей речевой деятельности. Устный неподготовленный дискурс характеризуется быстротой процесса отражения явлений, и в ходе реализации коммуникативного намерения говорящего могут меняться ресурсы, связанные с темой, объемом речи, стилем подачи и т.д.

Для корректировки сбоев в устной спонтанной речи применяются слова, которые имеют много наименований в научной литературе: «незнаменательная лексика», «лишние слова», «пустые частицы», «вставные элементы», «слова-заменители», «слова-паразиты», «вербальные хезитативы» и т.д. Понятие спонтанной речи, вслед за Н.А. Хан, определяется следующим образом: это «та речь, главным условием реализации которой является ее неподготовленность, выражающаяся на психолингвистическом (спонтанность речи относится к этапам речевой деятельности, начиная с замысла до ее реализации), дискурсивном (речевое поведение говорящего) и языковом уровнях (проявление спонтанности в фонетике, лексике, синтаксисе)» [Хан, 2013, с. 8].

Сам термин «хезитация» имеет узкую трактовку (непреднамеренные паузы в речи, во время которых говорящий осуществляет выбор следующей речевой единицы) и широкую (всякого рода перестройки «на ходу» структуры высказывания – добавления, исправления, возвраты, повторы и т. д.») [Сапунова, 2009]

Паузы хезитации принято делить на *незаполненные* (отсутствие акустического сигнала, ноль звука, перерыв в звучании) и *заполненные*. Согласно О. А. Александровой и В. В. Иваницкому [Александрова,

Иваницкий, 2009], заполненные паузы хезитации могут быть представлены следующими явлениями.

1. Вокализация (*э-э, м-м, гм*).

2. Затяжки звуков – удлинение гласных и согласных звуков (*По телеви-изору нра-авятся э-эти познавательные передачи*).

3. Паузы вербального поиска, а именно:

– семантизированные модальные и вводные слова и выражения; указательные, притяжательные, определительные местоимения и указательные местоименные наречия, которые часто называют словами-паразитами или пустыми словами (*значит, ну, это, вот, этот, самый, как бы, какие-то, как-то, там…*);

– метатекстовые комментарии (*как вы знаете, видите ли, понимаете*...);

– непреднамеренные повторы, среди которых выделяется полный или частичный повтор слова, а также повтор словосочетания (*я имею в виду ду ду ду*);

– плеоназмы (*очень прекрасный, более интереснее*);

– стереотипы, штампы (*ради Бога, реклама – двигатель торговли*).

4. Невербальные паузы хезитации – покашливания, вздохи, смех, прочищение горла, цоканье языком.

5. Самопрерванные конструкции: незаконченные высказывания, рестарты (повторные начала), фальстарты (неудачные начала).

Явление хезитации активно изучается на материалах устных корпусов. Так, Н. В. Богданова-Бегларян описала «поисковую функцию» вербальных хезитативов на материале двух блоков Звукового корпуса русского [Богданова-Бегларян, 2013]. В данном случае они маркируют поиск говорящим разных компонентов текста:

– поиск глагола-сказуемого (или его части);

– поиск предиката (не действия);

– поиск наименования (чаще – имени существительного);

– поиск определения;

– поиск обстоятельства;

– поиск предикативной единицы.

В.И. Подлесская, А.А. Кибрик рассматривают основные типы речевых сбоев, описывают принципы классификации одного из наиболее распространенных типов речевых сбоев и предлагают основанные на этой классификации правила разметки различных речевых сбоев при транскрибировании образцов устного дискурса для нужд корпусных исследований [Кибрик, Подлесская, 2007].

Ю.В. Дараган классифицировала так называемые «слова-паразиты» [Дараган, 2000]. Исходя из модели порождения высказывания А.А. Леонтьева было выдвинуто предположение, что употребление слов-паразитов затрагивает этап замысла (программы), этап реализации программы и этап сопоставления реализации замысла с самим замыслом. Исходя из этого слова-паразиты были разделены на 3 большие группы – *установочные*, *программирующие* и *гипертекстовые* маркеры.

Установочные маркеры отражают коммуникативные и психологические установки говорящего. Программирующие маркеры можно разделить на 5 подгрупп: маркеры планирования, которые означают формирование сценария реализации замысла (*так, ну*); маркеры концептуализации, которые означают формирование в сознании говорящего понятия, соответствующего плану содержания одной или нескольких языковых единиц, которое затем будет вербализовано (*как бы, такой, вот*); маркеры контроля, которые означают правильность реализации сценария (*значит, вот*); маркеры поиска, которые означают припоминание нужного слова в процессе реализации (*это (самое), этот (самый)*) и маркеры коррекции, которые означают коррекцию реализации самого сценария (*ну, это*). К гипертекстовым маркерам относятся слова, предоставляющие слушающему дополнительную информацию об отношении говорящего к

произносимому им тексту. Они делятся на: *оценочные* маркеры, выражающие оценку говорящим собственных высказываний с точки зрения удачности реализации (*как бы, так сказать, скажем*), важности сообщаемой информации, а также степени ее актуализованности (вот, там) и др.; маркеры-апеллятивы, передающие те или иные требования/просьбы говорящего к адресату.

Итак, исследуемые единицы обладают целым рядом семантических и функциональных особенностей, мотивированы и порождены определенным типом дискурса, как правило, характеризуются многозначностью и слабой смысловой дифференцируемостью, напрямую зависят от контекста, могут выступать в роли неполнозначной единицы, вследствие чего дискурсивные маркеры трудно квалифицируются в речи, их статус может варьироваться даже в рамках одного высказывания. Таким образом, данные единицы составляют тот тип единиц, которые вызывают особую сложность для их выделения и квалификации при автоматической обработке естественного языка, корпусном аннотировании.

1.2 Автоматическое разрешение омонимии и полисемии

Проблема устранения лексической неоднозначности является неразрешенной проблемой обработки естественного языка. Используя понятие неоднозначности, мы имеем ввиду то, что одному означающему соответствуют разные означаемые. Основные виды такого соответствия – это полисемия (или многозначность) и омонимия. Учитывая способность языка непрерывно развиваться, данные языковые явления превращаются в одну из самых насущных проблем не только общего языкознания, но и компьютерной лингвистики.

Иногда для снятия лексической неоднозначности необходим морфологический анализ. В корпусе устной речи тюркско-русских билингвов Южной Сибири используется морфологический анализ, выполненный

морфологическим анализатором русского языка Mystem. Морфологическим анализом в компьютерной лингвистике называется получение леммы или основы (псевдоосновы) заданного токена, а также морфологических параметров [Большакова Е.И. и др., 2017, с. 33]. Mystem приводит слова к начальной форме, прописывая их грамматические характеристики (часть речи, время, падеж, число, наклонение, форма прилагательного, лицо, род, вид, залог, одушевленность). Также он может указывать на различные особенности слова: например, редко встречающееся слово, сокращение, устаревшая форма, обесцененная лексика и т.д. Что касается принципов работы, то программа использует словарь часто встречаемых русских слов, а для остальных предлагает гипотетическую начальную форму.

Прежде чем провести морфологический анализ, необходимо выделить из текста отдельные слова. В связи с этим вместе с системой морфологического анализа иногда поставляется подсистема графематического анализа. Входной поток символов разбивается на токены нескольких классов: буквенные последовательности, числа, цифробуквенные комплексы, пунктуация, разделители, иероглифы. При этом каждый класс токенов имеет собственный набор тегов, в частности, для слов это может быть язык (кириллица или латиница) и регистр.

Одному и тому же токену может соответствовать несколько разных словоформ, что и называется лексической неоднозначностью.

Снятие (или разрешение) омонимии – одно из явлений лексической неоднозначности — этап анализа текста, на котором проводится выбор единственного варианта морфологического анализа для каждого токена.

Существует 4 класса омонимии:

– Неоднозначные по параметрам – словоформы с различными множествами грамматических параметров, но совпадающей леммой и частью речи.

– Неоднозначные по части речи — словоформы, совпадающие по лемме, но отличающиеся по части речи. Так как части речи не совпадают, то

наборы параметров у словоформ также будут отличаться. В связи с этим сравнение параметров проводиться не может.

– Неоднозначные по лемме — словоформы, отличающиеся по лемме, но имеющие одинаковую часть речи. Здесь параметры могут как совпадать, так и отличаться.

– Неоднозначные по части речи и лемме — словоформы, отличающиеся как по лемме, так и по части речи. Сравнение параметров здесь также проводиться не может.

В нашем случае слова имеют одну и ту же лемму, используясь в разных функциях и в разных значениях.

Для разрешения проблемы омонимии и полисемии существует три основных подхода:

– основанный на правилах;

– основанный на статистике;

– основанный на машинном обучении.

Первый метод основан на том, что применяется определенный набор правил, и на их основе определяются неподходящие варианты морфологического разбора слов и выбираются правильные. Однако данный метод предполагает ручное составление правил, привлечение знаний экспертов, долгую и трудоемкую работу. Для каждого правила необходимо создать самостоятельный программный модуль, и пополнение системы правил становится всё труднее с каждым новым правилом. Ввиду этого подобные методы не получили широкого распространения.

Статистические методы и методы, основанные на машинном обучении имеют большую популярность, что связано с наличием открытых, размеченных корпусов, объём которых позволяет построить довольно точные модели. Простейший способ снятия морфологической омонимии – подсчёт статистики разных вариантов разбора по корпусу. Вместе с тем вычисляются апостериорные вероятности каждого из разборов по размеченному корпусу со снятой омонимией. В настоящее время существует несколько способов

подсчёта таких вероятностей, однако все они не оказывают значительного влияния на точность снятия омонимии. Здесь более важен сам корпус: его представительность, объём, точность разметки.

Хорошие результаты дает также учёт контекста слова. Чаще всего для учета контекстов используют триграммную модель — анализ слова и его контекста из ещё двух слов. Триграммная модель показывает значительно лучшие результаты, чем уни- или биграммные, а четырехграммная модель занимает больше места, не давая значительного прироста в точности.

Таким образом, если сравнивать эти два подхода, то статистический имеет два преимущества: он не требует определенных знаний, необходимых для составления правил, и почти не зависит от предметной области анализируемых текстов.

Кроме статистических методов снятия омонимии и полисемии в настоящее время применяются и другие методы классификации. Допустим, что каждая словоформа – результат морфологического анализа выбранного токена – принадлежит одному из двух классов: корректное предсказание и некорректное предсказание. Соответственно, проводится бинарная классификация словоформ, у которой в качестве параметров могут использоваться грамматические параметры данного или соседних слов в некотором окне, их леммы, признаки наличия знаков препинания и проч. Для обучения используются размеченные корпуса, а в качестве метода классификации могут использоваться различные методы обучения, например, нейронные сети, скрытые марковские поля и т.д.

В нашем случае мы использовали NLP-библиотеку FastText от Facebook Research, которая стала большим шагом в развитии векторных семантических моделей и машинного обучения в обработке текста.

FastText представляет собой бесплатную библиотеку с открытым исходным кодом от Facebook AI Research (FAIR) для изучения встраивания слов и классификации слов. Благодаря библиотеке можно создавать алгоритм обучения без учителя или контролируемого обучения для получения

векторных представлений слов.

FastText может применяться в следующих случаях:

- для поиска семантических сходств;
- для классификации текста (например, для фильтрации спама);
- для обучения большого количества данных за короткий промежуток времени.

Для получения векторного представления слов одновременно используются модели skipgram и CBOW (Continuous Bag-of-Words). Модель skipgram учится предсказывать целевое слово благодаря соседнему слову. С другой стороны, модель CBOW предсказывает целевое слово в соответствии с его контекстом. Контекст представлен в виде набора слов, содержащихся в окне фиксированного размера вокруг целевого слова. CBOW часто используется в контекстных словах для предсказания средних слов.

Также для модели векторных представлений слов используется skip-gram с негативным сэмплированием. Негативное сэмплирование — это способ создать для обучения векторной модели отрицательные примеры, то есть показать ей пары слов, которые не являются соседями по контексту. Для каждого положительного примера (когда слова в тексте стоят рядом, например, «плюшевый медведь») мы подбираем несколько отрицательных («плюшевый телевизор», «плюшевый автомобиль», «плюшевое бегство»). Всего подбирается от 3 до 20 случайных слов, что не требует много компьютерного времени и позволяет ускорить работу FastText.

Skip-gram игнорирует структуру слова, но в некоторых языках есть составные слова, как, например, в немецком. Поэтому к основной модели была добавлена subword-модель. Subword-модель — это представление слова через цепочки символов (n-граммы) с n от 3 до 6 символов от начала до конца слова плюс само слово целиком. Например, слово замок с n = 3 будет представлено n-граммами <за, зам, амо, мок, ок> и последовательностью <замок>. Таким образом, для модели есть разница между последовательностью <зам> в слове зам — и n-граммой зам из слова

замок. Такой подход позволяет работать и с теми словами, которые модель ранее не встречала.

Признаки, полученные при помощи разбиения на n-граммы, имеют огромную размерность (т.е. для текста получается очень большая и тяжелая таблица). Это может замедлить работу обучаемой на этих признаках модели. Для фиксирования размерности признаков применяется хэширование признаков (специальная процедура, которая позволяет кодировать объекты разных размеров с помощью символьных цепочек одинаковой длины). Признаки получают хэш-индексы, что помогает считывать их быстрее.

Выводы к главе 1

Исследования дискурсивных слов на современном этапе развития функциональной лингвистики достаточно актуальны, затрагивают широкий спектр аспектов описания и типологизации дискурсивных маркеров, основываются на методах смежных для лингвистики наук. Предпринятое исследование имеет целью предложить эффективный метод машинного обучения для решения задач автоматического разграничения полисемии и омонимии одного из дискурсивных слов – *вот*.

Данная глава закладывает теоретическую базу работы, в ней представлены определения базовых понятий диссертации – «дискурсивные маркеры», «хезитация»; рассмотрены существующие классификации дискурсивных маркеров и хезитативов; описана полемика в отношении статуса данных единиц, принципиальных различий. Отмечается, что дискурсивные слова – порождение определенного типа дискурса и его маркеры; они обладают сложной смысловой структурой, крайне многозначны, зависимы от контекста и сложно квалифицируются в речи.

В главе рассматривается проблема снятие полисемии и омонимии в рамках создания разметки языковых корпусов, описываются типы возникающей лексической неоднозначности и существующие способы

разграничения полисемии и омонимии при автоматической обработке естественного языка, в том числе в приложении к таком разряду единиц, как дискурсивные слова. В качестве базы использованы контексты, извлеченные из корпуса устной речи тюркско-русских билингвов. Несмотря на наличие трех основных подходов в решении указанной проблемы (основанного на правилах, статистике, машинном обучении), их эффективность для поставленной цели исследования не одинакова с точки зрения трудоемкости процесса, точности получаемых результатов и т.д. В рамках магистерской диссертации для анализа было принято решение создать собственную модель, базирующуюся на методах машинного обучения и позволяющую разграничить значения исследуемого дискурсивного слова *вот*.

2. Функционально-семантические особенности дискурсивного слова *вот*

Далее в работе описание предпринятого разграничения полисемии и омонимии дискурсивных слов методами машинного обучения. В качестве материала были выбраны контексты с дискурсивным словом *вот*, представленные в Бимодальном корпусе устной речи тюркско-русских билингвов (RuTuBic).

2.1 Характеристика частотности слова *вот*

Выбор данного дискурсивного слова для анализа и дальнейшей работы обоснован рядом факторов, одним из которых является его частотность среди других слов, подтвержденная Частотным словарем русской лексики (57 место) [Ляшевская, Шаров, 2009]. Его ipm – общая частота, характеризующая число употреблений на миллион слов корпуса – составляет 1785.1.

Данные Национального корпуса русского языка также подтверждают частотность выбранной единицы: поиск на запрос «вот» в основном корпусе показал 532 280 вхождения (при объеме корпуса 337 025 184 слова). При этом самыми частыми сферами функционирования для *вот* являются

художественная (60.21%),

публицистическая (26.85%)

бытовая (7.28%) сферы, а самая популярная тема для текстов, в которых употребляется изучаемое слово, – *частная жизнь* (11.97%).

Характеризуя функционирование данной единицы в корпусе устной речи тюркско-русских билингвов, также следует отметить его частотность (см. подробнее Главу 2).

Кроме того, для данного слова характерна многозначность, а также частое использование в качестве хезитатива, что также повлияло на выбор данной единицы в качестве объекта описания.

2.2 Анализ семантических особенностей дискурсивного слова *вот*

В качестве материала для определения семантической структуры слова *вот* мы обратились к следующим лексикографическим источникам XX–XXI веков: Словарь современного русского литературного языка (БАС), Малый академический словарь русского языка, Толковый словарь русского языка Д.Н. Ушакова, Толковый словарь русского языка Т.Ф. Ефремовой, Словарь русского языка С.И. Ожегова, Большой толковый словарь русского языка С.А. Кузнецова, Грамматика русского языка (Академграмматика 80). Нами были проанализированы дефиниции рассматриваемой единицы и иллюстративные контексты, предлагаемые источниками.

Остановимся на семантических особенностях *вот*, представленных в Большом академическом словаре русского языка [Словарь современного русского литературного языка, 1950-1965]. У данного слова выделяется два основных значения

Первое значение: Употребляется для указания на наличие кого-, чего-либо поблизости в данный момент времени.

Уточняется, что данная частица иногда используется с повторением. *Вот лещик, потроха, вот стерляди кусочек.*

Также отмечены некоторые оттенки значений.

Например, *вот* употребляется при передаче, вручении, указании каких-либо предметов. *Председатель прислал книжку и бумагу какую то. Вот...*

Для более картинного живого изображения употребляется в повествованиях и описаниях при глаголах, с значением: *смотри, гляди* (иногда с повторением). *Вот стороной дороги бегут две потные, косматые лошади в хомутах с захлестнутыми за шлеи постромками.*

При существительных с отвлеченным значением подчеркивает качественное значение существительного, соответствуя обороту: *вот какой, каков* (нередко существительному предшествует местоимение *он, она, оно*). *Вот история, я опасно болен, а между тем мне не дают никаких лекарств.*

При подведении итога сказанному или в значении прекращения разговора. *Садись чай пить, вот и разговор весь.*

В сочетании с местоимениями и местоименными наречиями усиливает и уточняет их: *вот кто, вот что, вот где, вот как* и т. п. (с ударением на слове «вот»). Да кто выпустил, – вот кто выпустил: эти молодцы!

Употребляется при 2-м лице повелительного или изъявительного наклонения глагола при указании на какой-либо предмет. *Вот посмотри на эту картину.*

Отмечается, что в просторечии может сопровождать указательный жест. *Да, помню, как однажды при мне покойник-батюшка из сада принес яблоко – вот! Отец складывал вместе оба кулака, чтоб дать понятие о величине яблока.*

Второе значение, отмеченное в словаре: Служит для выделения, усиления местоимений (при логическом на них ударении, соответствуя иногда по значению слову «именно»).

Вот то-то мне и духу придает, что я, совсем без драки, могу попасть в большие забияки.

Далее представлены оттенки значений:

В сочетании с наречиями времени, места, причины и т. п. уточняет и усиливает их.

Вот где я нашел деньги.

В том же указательно-усилительном значении употребляется с существительными типа: *вздор, ерунда, невидаль* и т. п.

Вот невидаль: мышей! Мы лавливали и ершей.

При восклицаниях усиливает положительный или иронический оттенок.

Вот прелесть!

В Малом академическом словаре помимо значений и оттенков значений, указанных в Большом академическом словаре, выделяется значение, в котором *вот* употребляется при указании на последовательность

действий, на смену явлений, происходящих как бы перед глазами (может повторяться перед каждым следующим членом) [Словарь русского языка, 2016].

Вот раздалося «ау!» вдалеке; Вот над колосьями в синем венке черная быстро мелькнула головка.

Также выделяется значение, в котором *вот* употребляется для привлечения внимания собеседника.

Вот взгляните: старинное лото, в которое еще играла с нами покойная мать.

Анализ дефиниций вот в Толковом словаре Ушакова [Толковый словарь русского языка, 2008] показывает наличие значений, которые можно добавить к выше отмеченным значениям.

1. Служит для указания на то, по поводу чего высказываются какие-нибудь мнения, что является поводом к речи.

Вот теперь вздумали квартиру менять, а разве на старой тесно?

2. Ставится рядом с тем словом, на которое падает логическое ударение, для большего его выделения (разг.).

Денег вот дать не обещаю.

3. В сочетании с последующими или предшествующими указательными местоимениями *тот, этот, такой, таков* и наречиями *там, туда, оттуда, тогда, так* усиливает их указательное значение.

Прочтите-ка рассказ вот этот не спеша.

Словарь Ефремовой [Новый словарь русского языка. Толково-словообразовательный, 2000] и словарь Ожегова [Толковый словарь русского языка, 1994] показывают, что *вот* употребляется в значении связки при именном сказуемом.

Мир на земле – вот наша цель.

Также оба словаря отмечают то, что в некоторых случаях *вот* выражает истинность, отсылает к истинности того, что утверждалось ранее. *Все-таки была игра. – В. А я что говорил! Озяб? В. говорил: надень плащ.*

Большой толковый словарь Кузнецова [Большой толковый словарь русского языка, 1998] дополняет данные остальных словарей, выделяя значение, при котором *вот* выражает угрозу рассчитаться с кем-л. или злорадство по поводу совершившейся расплаты, возмездия.

Академграмматика трактует данную единицу как совмещающую значения указательного наречия и подчеркивающей, усиливающей частицы. В предложении *Вот мельница!* это слово выступает как наречие; в предложениях *Вот (еще) новости!, Вот (она, современная) молодежь!, Вот и добился!* – как частица. Однако отсутствие резкой грани между наречием и частицей обнаруживается здесь в том, что в первом предложении с наречием *вот* связано значение непосредственного указания, обязательно совпадающего с моментом речи; именно поэтому в это предложение не могут быть введены показатели прошедшего или будущего времени [Грамматика русского языка, 1980].

Таким образом, семантический анализ показал наличие у слова *вот* двух основных (ядерных) значений.

Первое – указание на объект, находящийся в непосредственной близости, как бы перед глазами. *Вот идёт поезд; вот наш дом; вот здесь пойдем*. Отметим, что в словаре Ожегова в данном значении *вот* квалифицируется как местоименное наречие [Словарь русского языка, 2015], в отличие от других словарей, где *вот* в данном значении оценивается как указательная частица.

Во втором значении *вот* ставится рядом с тем словом, на которое падает логическое ударение, для большего его выделения. Оно уточняет и усиливает указательные и относительные местоимения, а также наречия места, времени, причины и т.п. Таким образом, *вот* в сочетании со словом, стоящим рядом, выполняет экспрессивную, выделительную и организующие функции: выделяя смысловой центр синтагмы, логическое ударение вместе с тем организует вокруг выделенного слова всю синтагму. *Вот что я тебе скажу; денег вот дать не обещаю; вот вас-то мне и надо*. В этом значении оно является частицей и иногда соответствует по значению слову «именно».

В ходе работы по данным словарей была составлена таблица, содержащая все выделяемые значения с указанием части речи, стилистических помет, сочетаний, с которыми употребляется слово, и других примечаний.

Словарь	Часть речи	Помета	С чем употребл	Значение	Пример	Примечание
Ожегов	мест. нареч.			указывает на происходящее или находящееся в непос	В. идёт поезд. В. наш дом: В. здесь пойдет	
	мест. нареч.		в сочетании с в	придаёт им смысл подчёркнутого указания на что-н. в	В. что я тебе с	[всегда ударное]
	частица.		в сочетании с м	выражает высокую степень оценки, удивления	В. сколько орехов! В. в какую глушь заехал	
	частица.			открывая предложение, употр. для выделения того, на	В. послушайте, что говорят. В. вас-то мне и	
	частица.	разг.		выражает истинность, отсылает к истинности того, что	Все-таки была игра. -В. А я что говорил! Оз	
	частица.	разг.		то же, что хорошо	Послушай, что было. Звонит он вечером. В	
	в знач, связки		при именном ск	казуемом	Благополучие всех - в. наша цель.	
МАС				употребляется при указании на кого-, что-л. находяще	Вот стороной дороги бегут две потные косм	
				употребляется при указании на последовательность де	Вот раздалося	может повторяться перед ка
			с указательным	для уточнения и усиления их значения	Мы скоро подружились вот по какому случа	
				употребляется для привлечения внимания собеседник	Вот взгляните: старинное лото, в которое ещ	
				употребляется в восклицательных предложениях для у	Жениться ведь вздумал? Вот вздор: совсем	
	частица указат		в сочетании с ч	образует сложные частицы, подчеркивающие эмоцион	Вот то-то, все вы гордецы! Спросили бы, ка	
Словарь Ушако				служит для указания на что-н., находящееся или прои	Вот наш дом.	
				в восклицаниях служит для такого указания на что-н.,	Вот странное уничиженье! Вот дурень! Вот	
				служит для указания на то, по поводу чего высказыва	Вот теперь вздумали квартиру менять, а ра	
		разг.		ставится рядом с тем словом, на к-рое падает логичес	Денег вот дать не обещаю. Вот вас-то мне	
	зательная част		в сочетании с г	усиливает указательное значение местоимений	Прочтите-ка ра	ударение падает на местоим
Словарь Ефрем				употребляется при указании на кого-либо или на что-либ	о, находящихся или происходящих в непо	
				употребляется при подчеркивании или усилении качественно-количес	обычно в сочетании со сл.: к	
				употребляется при выделении главного, существенного,	того, на чём необходимо сосредоточить в	
				употребляется при выражении истинности, соответствуя	по значению сл.: вот именно	
	част.	разг.		употребляется в значении связки при именном сказуем	ом	
БАС	указательная частица			либо поблизости в данный момент времени	[Алеша] напра	иногда с повторением
	оттенок значения			при передаче, вручении, указании каких-либо предмет	Председатель	прислал книжку и бумагу ка
	оттенок значения			употребляется в повествованиях и описаниях при	пристальнее,	иногда с повторением
	оттенок значения		при существит	существительного, соответствуя обороту: вот какой,	опасно	нередко существительному
	оттенок значения		местоимениям	наречия : вот кто, вот что, вот где, вот как и т. п.	выпустил, —	с ударением на слове «вот»
	оттенок значения		при 2-м лице по	при указании на какой-либо предмет	Вот видишь приближающегося к нам челов	
	оттенок значен	в просторечии		сопровождает указательный жест.	как однажды	
	указательная частица			ний, соответствуя иногда по значению слову	и духу	при логическом ударении на
			наречиями	уточняет и усиливает наречия	нашел	
			в соединении с	предшествующим междометием «ну»	] Послушай,	
			существитель	указательно-усилительное значение	невидаль:	
Большой толко	частица			происходящее перед глазами, в непосредственной	поезд.	
			с местоим. и на	последующего слова или высказывания в целом	тебе скажу.	обычно ударная
			с частицами и	употребляется для усиления восхищения, удивления,	Вот это новость!	
			с частицами и	указывает на неожиданный результат, нежелательный	Вот какое дел	обычно с оттенком горечи, ра
			с частицами	употребляется для подтверждения предыдущих слов	Вот то-то и оно!	
			с личными мес	выражает угрозу рассчитаться с кем-л. или злорадство	Вот тебе, получи сполна!	
	связка при именном сказуемом				Мир на земле - вот наша цель.	
				в составе сложных союзов: (и) вот почему. (и) вот оттого. (и) вот поэтому		

Рисунок 1. – Слово *вот* в словарях XX – XI вв.

Помимо ядерных значений единицы, зафиксированных во всех словарях, выделяются периферийные значения, представленные либо отдельными значениями, либо оттенками значений.

1. *Вот* выражает высокую степень оценки, удивления. употребляясь в

сочетании с местоименными словами “какой”, “как” и некоторыми другими в восклицаниях. Служит для такого указания на что-нибудь, что сопровождается иронией, насмешкой, досадой и т.п., а также усиливает эмоциональную окрашенность восклицательных предложений.

Вот сколько орехов! Вот в какую глушь заехали! Вот как обрадовался! Вот странное уничиженье! Вот дурень! Вот уж не думал, что так случится!

2. *Вот* выражает истинность, отсылает к истинности того, что утверждалось ранее, вот именно.

Озяб? Вот говорил: надень плащ.

Отмечается, что в данном значении слово употребляется как разговорное.

3. *Вот* означает то же, что *хорошо*.

Послушай, что было. Звонит он вечером. Вот (ну вот). Оказывается, у него важное дело.

Является разговорным в данном значении.

4. *Вот* употребляется при указании на последовательность действий, на смену явлений, происходящих как бы перед глазами.

Вот одни человек прошел, вот другой.

При этом *вот* может повторяться перед каждым следующим членом.

5. *Вот* в сочетании с частицами *ведь, то-то, уж* и т.п. образует сложные частицы, подчеркивающие эмоциональную сторону речи, усиливающие выразительность.

Вот то-то, все вы гордецы! Спросили бы, как делали отцы?

6. *Вот*, употребляясь с частицами и местоимениями, указывает на неожиданный результат, нежелательный итог, обычно с оттенком горечи, разочарования, недоумения.

Вот какое дело получается.

7. *Вот* употребляется с частицами для подтверждения предыдущих слов собеседника, как *именно так*.

Вот то-то и оно!

8. *Вот* выражает угрозу рассчитаться с кем-либо или злорадство по поводу совершившейся расплаты, возмездия, употребляясь с личными местоимениями.

Вот тебе, получи сполна!

Помимо вышеперечисленных значений, в которых *вот* выступает в статусе указательных частиц, оно употребляется в значении связки при именном сказуемом.

Благополучие всех – вот наша цель.

Следует отметить целый ряд отдельных выражений с *вот*, имеющих в каждом контексте тонкие смыслы пожелания, иронии, удивления и др.: *(Ну) вот ещё!; вот (оно) как! или вот (оно) что!; вот (оно) что! – каково!; вот бы; вот вам Бог; вот вам!; вот ведь; вот где сидит* (кто-, что-л.); *вот ещё; вот и; вот и весь сказ; вот и всё; вот и вся недолга; вот и…; вот именно; вот как; вот оно как; вот оно что; вот поди ж ты!; вот почему; вот так; вот так вот; вот так да; вот так так; вот так штука!; вот так; вот так-то; вот те (и) на́! вот те и; вот тебе (вам)!; вот тебе (и) на!; вот тебе (те) раз!; вот только б (бы); вот то-то (же); вот то-то и оно; вот уж; вот уже; вот ужо!; вот что; вот это; вот это да; вот я тебя (его, их и т.д.)!; вот, вот; вот... так; вот-вот; и вот; и вот почему (оттого, поэтому); так вот* и др.

Например, выражение *вот именно* употребляется в подтверждение чего-либо, отрицаемого собеседником.

Это ко всем не относится. Равные люди-то! – Вот именно ко всем... Все одинаковы мы...

Вот-вот (в слитном произношении) – сейчас, тотчас, еще немного и.... *Раза два мы отдыхали, потом опять шли, в надежде, что вот-вот покажется море.*

Вот, вот (в разделении) – подтверждает высказываемые собеседником

мысли и соответствует по значению словам: «так, так», «именно так».

Чувствую, что сегодняшнее мое напряжение разрешится чем-нибудь... Или я сломаю что-нибудь, или ... — Вот, вот, это именно и нужно.

То-то вот оно и есть употребляется при подведении итога, заключения сказанному.

Ведь вы по будням-то из одной чашки едите, в восьмом часу спать ложитесь. Только в именины да по праздникам и раскошеливаетесь. А мы, видишь, как живем. Смотри! То-то вот оно и есть.

Вот еще! – при отрицании чего-либо, при несогласии с чем-либо или при отрицательном отношении к чему-либо.

Я тебя к себе заберу. – Не поеду. Вот еще!

Вот то-то (же) – при усилении подтверждений в смысле: именно так, в том-то и дело.

Служить бы рад, прислуживаться тошно. – Вот то-то, все вы гордецы!

Вот как! Вот (оно) что! – при выражении удивления или признания неясной прежде мысли.

Ах! Вы не знаете? Горюнова Мишу произвели – вот мы сегодня и отличаемся, – в восторге добавил Волков. – Вот как! сказал Обломов.

Вот тебе! вот вам! – получай(те) наказание по своим заслугам.

Что взял! а! что взял! Вот тебе, собака! Ну, теперь подавись моими деньгами.

Вот тебе (и) на! Во т тебе (или те) раз! Вот так так! Вот так штука! – выражает удивление по поводу чего-либо, поражающего своей неожиданностью.

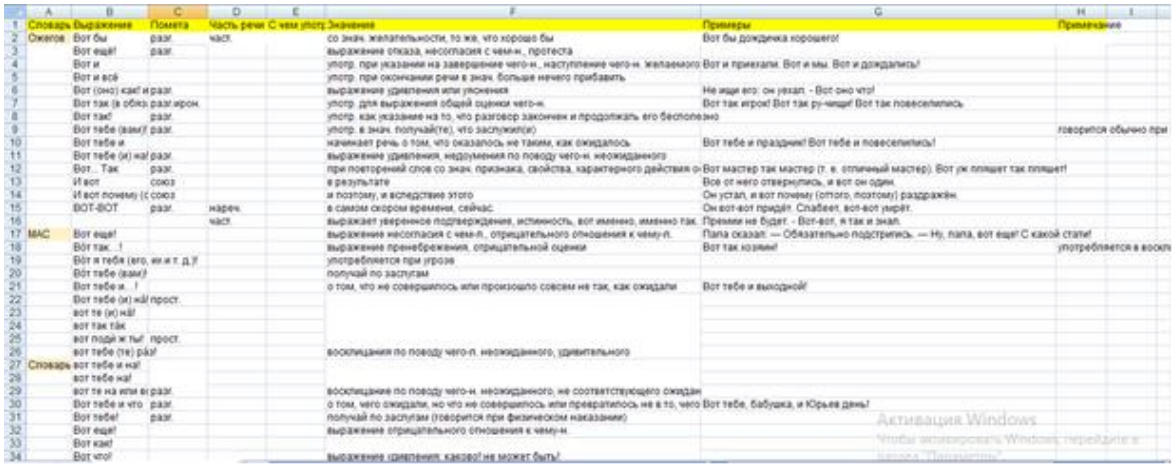
Вот тебе и... – о чем-либо ожидаемом, которое совсем не происходит или изменяется в неожиданную сторону.

Вот тебе и подарок!

Вот так... – употребляется в иронических восклицаниях при выражении пренебрежительной оценки.

Вот так хозяин!

В основном данные выражения используются в разговорной речи и отмечены в лексикографических источниках такими пометами, как «разговорное», «просторечие», «фамильярное», «ироничное». По своему морфологическому статусу они могут являться разными частями речи – союзами, частицами, междометиями и наречиями.



	A	B	C	D	E	F	G	H	I
1	Словарь	Выражение	Помета	Часть речи	С чем употр	Значения	Примеры	Примечание	
2	Ожегов	Вот бы	разг.	част.		со знач. желательности, то же, что хорошо бы	Вот бы дождика хорошего!		
3		Вот ещё!	разг.			выражение отказа, несогласия с чем-н., протеста			
4		Вот и				употр. при указании на завершение чего-н., наступление чего-н. желаемого	Вот и приехали. Вот и мы. Вот и дождались!		
5		Вот и всё				употр. при окончании речи в знач. больше нечего прибавить			
6		Вот (оно) как! и	разг.			выражение удивления или упрека	Не ищи его: он уехал. - Вот оно что!		
7		Вот так (в обяз	разг.-ирон.			употр. для выражения общей оценки чего-н.	Вот так игрок! Вот так ру-чищи! Вот так повеселились		
8		Вот так!	разг.			употр. как указание на то, что разговор закончен и продолжать его бесполезно			
9		Вот тебе (вам)!	разг.			употр. в знач. получай(те), что заслужил(и)		говорится обычно при	
10		Вот тебе и				начинает речь о том, что оказалось не таким, как ожидалось	Вот тебе и праздник! Вот тебе и повеселились!		
11		Вот тебе (и) на!	разг.			выражение удивления, недоумения по поводу чего-н. неожиданного			
12		Вот.. Так	разг.			при повторении слов со знач. признака, свойства, характерного действия о	Вот мастер так мастер (т. е. отличный мастер). Вот уж плящет так плящет!		
13		И вот	союз			в результате	Все от него отвернулись, и вот он один.		
14		И вот почему (с	союз			и поэтому, и вследствие этого	Он устал, и вот почему (оттого, поэтому) раздражён.		
15		ВОТ-ВОТ	разг.	нареч.		в самом скором времени, сейчас	Он вот-вот придёт. Слабеет, вот-вот умрёт.		
16				част.		выражает уверенное подтверждение, истинность, вот именно, именно так	Премии не будет. - Вот-вот, я так и знал.		
17	МАС	Вот ещё!				выражение несогласия с чем-л., отрицательного отношения к чему-л.	Папа сказал: — Обязательно подстригись. — Ну, папа, вот ещё! С какой стати!		
18		Вот так...!				выражение пренебрежения, отрицательной оценки	Вот так хозяин!	употребляется в воскл	
19		Вот я тебя (его, ее и т. д.)!				употребляется при угрозе			
20		Вот тебе (вам)!				получай по заслугам			
21		Вот тебе и...!				о том, что не совершилось или произошло совсем не так, как ожидали	Вот тебе и выходной!		
22		Вот тебе (и) на!	прост.						
23		вот те (и) на!							
24		вот так так							
25		вот поди ж ты!	прост.						
26		вот тебе (те) раз!				восклицания по поводу чего-л. неожиданного, удивительного			
27	Словарь	вот тебе и на!							
28		вот тебе на!							
29		вот те на или вот	разг.			восклицание по поводу чего-н. неожиданного, не соответствующего ожидан			
30		Вот тебе и что	разг.			о том, чего ожидали, но что не совершилось или превратилось не в то, чего	Вот тебе, бабушка, и Юрьев день!		
31		Вот тебе!	разг.			получай по заслугам (говорится при физическом наказании)			
32		Вот ещё!				выражение отрицательного отношения к чему-н.			
33		Вот как!							
34		Вот что!				выражение удивления, какого! не может быть!			

Рисунок 2. – Многообразие выражений со словом *вот* в словарях XX – XI вв. (фрагмент)

Проведенный анализ слова *вот* в устной речи тюркско-русских билингвов подтвердил то, что данная единица действительно часто употребляется как полнозначное слово. Среди отмечаемых в научной литературе значений мы обнаружили употребление *вот* в следующих значениях (при этом частотность употребления в разных значениях варьируется).

1. Данная частица служит для указания на что-нибудь, находящееся или происходящее перед глазами или как бы перед глазами в данную минуту, на наличие чего-нибудь в момент речи.

Мм из любимых концертных записей я могу назвать ээ «Пятнадцать лет Агате Кристи», «Филини», «БИ-2», «Сплин» две тыщи третьего года,

*ну ещё, Александр Красовицкий, **вот** это «Никогда».*

*Так я спела, он грит: «О, **вот** у тебя и козырь!».*

2. *Вот* ставится рядом с тем словом, на которое падает логическое ударение, для большего его выделения (разг.).

*Я **вот** занимаюсь футболом и хочу начать смотреть футбол.*

*Потом, если вернуться, то можно увидеть, что кого-то нет, потому что там мм волки могли погрызть или что-то такое. И **вот** поэтому папа ходит в тайгу на лошади именно, чтобы посмотреть за ними.*

*Нет, у нас город Нерюнгри, там, как я **вот** помню до сих пор почему-то, как-то мне это… запомнилось. Классные часы у нас всегда проводили, ну, посвященные дружбе народов.*

3. *Вот* служит для указания на то, по поводу чего высказываются какие-н. мнения, что является поводом к речи (как указание на следующую ситуацию, как разделитель ситуации).

***Вот** у Стругацких, например, один учитель, например, воспитывает четырёх людей с самого младшего возраста и до… до выпуска.*

*Ну **вот** в Киргизию приехали, на лицо же я не русская, они и не разговаривают по-русски.*

*Ну да. Он есть, ввели щас, по-моему. Ввели, но опять же **вот** эти двести тысяч или сто тыщ, я не знаю, объясняла нам педагог. На пять лет это тоже накладно.*

4. В сочетании с последующими или предшествующими указательными местоимениями и наречиями *вот* усиливает их указательное значение (причем ударение падает на местоимение или наречие, а *вот* произносится без ударения.

*Мм последнее я прочитала это «Туманность Андромеды» советского писателя, поэтому там тоже коммунизм процветает по всей планете, но только не так, как у Стругацких, там, уже коммунизм был по всей планете **вот** в это время, в две тысячи восемнадцатом году, а у него где-то через шестьсот лет и более, наверное.*

Вот *в этих книгах научно-фантастических много времени и объёма уделяется теории и описанию, поэтому мне и нравится читать, а сюжеты у них такие типичные.*

Также частица *вот* в сочетаниях с другими простыми частицами может вносить в предложение значения подчеркивания, акцентирования, положительной оценки, отрицания, иронии, следствия, подготовленности и т.д.; подчеркивает исчерпывающий характер сказанного.

У нас, шорцев, я видела только две тётки и мамин брат. Дядя, получается. ***Вот и всё.***

«Вот так вот» означает «именно так, а не как-нибудь иначе; именно таким образом».

int *Ну тоже шорец, считается, что кумандинский диалект шорского языка, да?*

resp *Да. Либо кумандинский, потому что у его бабушки, это кумандинцы воспитывали его у бабушки. Ну, как бы, в гости постоянно у бабушки. Как-то* ***вот*** *так* ***вот****, смешение немног… небольшое.*

Таким образом, анализ материалов корпуса демонстрирует крайнее разнообразие семантических оттенков единицы.

Что касается хезитации, то анализ материала показал, что *вот* употребляется и для корректировки сбоев в устной спонтанной речи, т.е. выступает в качестве хезитатива. При этом данное слово в основном выполняет поисковую функцию. Анализ материала показывает, что чаще всего происходит поиск предиката или наименования чего-либо.

Предиката

Даже я её устраивала на работу, и она ***вот*** *мои* ***вот*** *эти,* ***вот*** *то, что я вот конкурс, я когда предложила, грю: «Давайте конкурс делать красавиц, красавиц, шорочек»;*

А мои дети тоже были там, ***вот*** *эта старшая была и сын был,* ***вот*** *этот, который вот сидит, сноха у нас русская, вот;*

*Красный атлас такой, рубахи ребятам **вот** это самое, поём про Ленина, одну песню, только вспомнили одну песню, собрали по крупицам, ни одной песни не было, ничего;*

Наименования

*Уже **вот** это ушло, прошлое, слава тебе господи.*

*Вот нынче отправила **вот** этого, Артёма, который в Кемерово учится в институте, на третьем курсе.*

2.3 Функции дискурсивного слова *вот*

Учитывая весь спектр значений и оттенков значений дискурсивного слова *вот*, охарактеризуем функции, выполняемые им. В научной литературе указывается на 3 основные функции *вот*. Согласно точке зрения А.А. Рачёвой, это следующие функции [Рачёва, 2016].

Первая функция – **указательная**.

И.М. Кобозева описывает употребления *вот* с указанием на объект следующим образом: «Говорящий привлекает внимание адресата к объекту X (или его изображению), находящемуся в поле зрения Говорящего и/или Слушающего в момент речи» [Кобозева, 2007, с. 252].

Другой функцией, отмечаемой некоторыми словарями для *вот*, является **уточняющая**.

Данный тип употреблений слова схож с указательными, однако в данном случае само указание осуществляется с помощью наречий *там* и *здесь* или указательного местоимения *этот*, а *вот/вон* лишь усиливает, уточняет идею указания. В отличие от указательных употреблений, где слово *вот* акцентно выделено, в данном случае оно оказывается фразово-безударным.

Также существуют фрагменты в речи, где субъект речи с помощью рассматриваемого дискурсивного слова старается обратить внимание адресата на некоторый аспект ситуации (время/участники/состояние/признак

объекта). Подобные употребления отличаются от указательных, во-первых, просодически (маркер является фразово-безударным), во-вторых, функционально: нет идеи прямого указания на определенный объект, идея указательности сохраняется только в форме акцентирования внимания на некотором аспекте ситуации. Такую функцию можно назвать **акцентирующей**.

Помимо указанных, *вот* может быть маркером экземплификации. Подобный тип употреблений описывает И.М. Кобозева: «*Вот* обращает внимание Слушающего (Читателя) на X как на пример того, о чем говорилось ранее» [Кобозева, 2007, с. 252]. В целом механизм такого использования можно охарактеризовать следующим образом: субъект речи эксплицирует представления о некоторой ситуации, а затем с помощью *вот* указывает на определенный объект/ситуацию, являющийся иллюстрацией для описанного ранее положения вещей.

Кроме данных употреблений, выделяются следующие типы употреблений единицы, которые можно охарактеризовать как функции *вот*.

1. *Финальное вот*, использующееся говорящим для указания «на то, что некоторый отрезок дискурса завершен и говорящий переходит к следующему» [Кибрик, Подлесская, 2009, с. 148].

2. *Маркер открытия новой темы*. Чаще всего данный тип употреблений является фразово-безударным или имеет второстепенное ударение [Кобозева, 2007]. Говорящий при этом вносит новую информацию, логически связанную с предыдущим отрезком дискурса.

3. *Анафорическое вот.* Значение анафорического *вот* И. М. Кобозева описывает следующим образом: «Говорящий утверждает, что конкретным воплощением категории У в данной ситуации является только что упомянутый или ситуативно актуализованный X, а не что-либо иное» [Кобозева, 2007, с. 254]. По И. М. Кобозевой, *вот* в данном случае несет контрастное ударение.

4. *Маркер возвращения к теме*, подчеркивающий кореферентность разнесенных по тексту объектов, а нередко и событий.

5. *Маркер поиска.* Используется в ситуациях, когда говорящий испытывает трудности в подходящей вербализации, происходит поиск правильной номинации, нужных слов.

6. *Маркер чужой речи*. К данной группе относятся употребления, где *вот* используется на границе собственной речи говорящего и фрагмента, в котором он передает чужую речь [Рачева, 2016].

Анализ материала корпуса устной речи тюркско-русских билингвов показывает, что все описанные в литературе функции *вот* нашли подтверждение в конкретных текстах. Однако чаще всего в текстах встречается *вот,* выполняющее следующие функции.

1. **Указательная**

А вот это чё за буква? А здесь вот в Кабырзе мы.

2. **Уточняющая**

Вот этот мальчик сейчас в Академии учится, маленький такой, серьёзный. *Нет, этот вот Виталька раньше, а потом после смерти матери.*

3. **Выделительная**

Вот почему мне понравился Волгоград, мы в доме Павлова были. Какая вот была студенческая жизнь?

4. **Поисковая**

И вот эта, Любовь Георгиевна тоже была, по русскому языку и литературе, вообще молодец. Я, вот, я с ней ни ни разу не, это, не, не, не ругалась ничего.

Остальные случаи, например, финальное *вот (То есть, у нас, как бы, наша смена считается самой лучшей в больнице, вот.)* тоже встречаются в корпусе, однако гораздо реже.

2.4 Анализ контекстов употребления *вот* в бимодальном корпусе устной речи тюркско-русских билингвов

На следующем этапе исследования был проведен анализ речи русско-тюркских билингвов с целью выявления типовых контекстов, в которых используется маркер *вот*. В результате были выявлены следующие типовые ситуации.

1. В качестве указательной частицы *вот* используется перед именительным падежом существительного или местоимения – *вот* + И.п. = указательная частица.

Вот икра, попробуйте.

А это вот тэйкмек.

У нас это самое, знаешь, башкирский, это самый, язык, какая-то, так, а вот казанский язык – они прямо отличаются.

А вот варенье наварила, оно вообще никуда.

Они считаются городскими, папа считается городской, потому что он здесь вот ... городской парень.

2. *Вот* усиливает указательное значение в сочетании с последующими или предшествующими указательными местоимениями и наречиями – *вот + тот, это такой, таков, там, туда, оттуда, тогда, так / тот, это такой, таков, там, туда, оттуда, тогда, так + вот.*

А шорский язык вот, поступила когда... так вот ...учила.

Вот так мы и работали здесь.

Вот такие дела у нас.

Ну, вот такие вещи тоже были.

То есть вот так попасть в конфликтную ситуацию, это называется в струю попасть?

3. *Вот* в сочетаниях с другими словами составляет различные семантические единства. Например, *вот так вот* (нареч. разг.) = именно так, а не как-нибудь иначе; именно таким образом.

Ну потом был пер.. какой-то период, в это время, возник, когда все упрощалось: лепнину ту содрали, сделали современный потолок, там че-то такое короче, ну вот так вот, вот в здании, в этом здании был зал, хороший большой концертный, типа концертного зала.

То есть внуки у меня языка практически не знают, хотя вот эти двое внуков пою, поют на татарском языке, а говорить – не говорят. И не совс, и мало что понимают. Вот так вот.

По росту достаточно высокий, есть на фотографии, где он был када в санатории, вот так вот он, он стоит в центре и он выше всех.

Отметим, что одной из главных задач исследования является разграничение хезитативных употреблений от нехезитативных. Анализ контекстов показал, что *вот* является хезитацией, если оно представлено следующим образом.

1. Составляет отдельное предложение.

Не сказать, что настолько, ну ладно я-то, ладно, пойму, а вот старые люди уже не понимают. Вот.

Ну естественно, там какой-то ремонт был и так далее. Вот.

Вот, семьями уезжают даже, потому что работать негде. Вот…;

2. Стоит в начале, в середине или в конце предложения, обособляясь в корпусе запятыми.

Вот, в деревнях вот, он очень богатый этот хакасский язык.

Вот, по- поэтому свёкр тама, ну.

Вот, э, здесь вот мы пое-ну, проезжали Хакассию, поехали в Горный Алтайск.

До какого года я точно тоже тоже не помню уже, вот...

Это шестнадцатый год, значит в девяносто шестом году, вот.

А сейчас родители достаточно рано ушли из жизни, поэтому спросить не у кого, вот.

Там есть такой зал большой, ну, видимо, был для приемов, или как вот,

знати, где балы проводились, вот, и подобное.

Ну, мама у него тоже не высокая, по-моему, вот, мы не застали в живых его деда по материнской линии.

Особенно в детстве, сейчас-то она так выглядит, как все обычные дамы, вот, а в детстве она больше похожа была на русскую девочку именно.

3. Используется в выражениях *так вот* или *вот так вот* (в конце предложения или отдельным предложением).

А тада же, т- , тканей хороших, красивых не было, он откуда -то выписывал, женщинам шили костюмы, вот- эти , платья шили, вот так вот.

В принципе, так вот.

Ну, вот так вот.

Поэтому она вот собралась, ей надо – собралась, это, ну и дядя Миша, он, видно, вот как-то так вот.

Как интересно. Всё. Вот так вот.

Таким образом, мы видим, что хезитативное употребление *вот* отличается от других употреблений тем, что в данном случае *вот зависит от знаков препинания и положения в предложении.* Это значительно упрощает автоматическое разграничение хезитативного употребления *вот* от нехезитативного.

Однако остается проблема разграничения полисемии вот, т.е. дифференциации основных значений полнозначного *вот* в рассматриваемом корпусе. Для этого на основе отмеченных значений единицы в лексикографических источниках с учетом семантических особенностей *вот* в корпусе были выделены 2 основных (как наиболее типичных и частотных для употребления вот в материалах корпуса).

Первое значение – указательное и уточняющее. К нему можно отнести употребления *вот* в случаях, когда оно указывает на что-то, находящееся по близости или как бы перед глазами, или в случаях, где оно усиливает,

уточняет идею указания.

Вот эти манты, они ненатуральные или *Вот в этих книгах научно-фантастических много времени и объёма уделяется теории и описанию, поэтому мне и нравится читать, а сюжеты у них такие типичные.*

Второе значение связано с логическим ударением, когда какое-нибудь рядом стоящее слово или слова выделяются силой голоса. Вот у него, как бы ну я бы сказала гипертрофированное чувство э-э-э реализации самого себя.

Математика, физика, знание иностранного языка, в принципе, тоже, немецкий спрашивали, я вот со словарем это сдал.

В результате проведенного анализа было принято решение собрать контексты для трех случаев использование *вот*: указательное / уточняющее значение, акцентирующее значение и хезитативное употребление.

Выводы к главе 2

Данная глава представляет собой лингвистический анализ дискурсивного слова, разграничению омонимии и полисемии которого посвящена работа.

Во главе были описаны значения *вот*, представленные в современных лексикографических и научных источниках, проведен их сопоставительный анализ в аспекте отражения базовых значений рассматриваемой единицы и ее отдельных оттенков, позволяющий, с одной стороны, говорить о наличии у вот значительного количества значений, в том числе исключительно контекстуальных, а с другой стороны, представить отмеченные значения как ядерные и периферийные.

Дальнейший анализ контекстов реализации *вот* в иллюстративном материале современных словарей и в корпусе устной речи тюркско-русских билингвов, его функций позволил отобрать наиболее типичные для исследуемого дискурсивного маркера значения и функции, которые лягут в

основу классификации с помощью одного из алгоритмов машинного обучения, выбранного для проводимого исследования.

3. Снятие омонимии и полисемии с помощью методов машинного обучения

Основная задача настоящей диссертационной работы – снятие омонимии и полисемии – рассматривается как задача классификации контекстов дискурсивного слова *вот* с тем или иным значением. Для решения данной задачи были использованы методы машинного обучения, эффективность которых во многом зависит от качества и количества датасетов.

В результате анализа был собран датасет, содержащий 426 употреблений ***вот*** в качестве хезитации (отмечается в тексте как **вот_min**), 178 употреблений ***вот*** в качестве указательно-уточняющей функции (**вот_equal**), 186 употреблений в качестве логического ударения и выделительной по отношению к соседним элементам контекста функции (**вот_plus**).

Структуру разработанного кода можно разделить на несколько этапов, которые будут описаны далее.

Предварительно весь датасет делится на обучающую (Рисунок 3) и тестовую (Рисунок 4) выборку в соотношении 80:20.

	Unnamed: 0	tmp	all
0	1	Мм из любимых концертных записей я могу назва...	вот_plus
1	2	вот_plus в этих книгах научно-фантастических м...	вот_plus
2	3	вот_plus она хитрая такая была, вот_plus этот ...	вот_plus
3	4	А меня заставили так вот_plus.	вот_plus
4	5	А в городе бывало, что придешь :, мне, ну, как...	вот_plus

Рисунок 3. Обучающая выборка с контекстами слова *вот*

	Unnamed: 0	tmp	all
0	12	Да, к-, как , до шестнадцати, так они были –...	вот_plus
1	15	У меня как раз мама попала ну я об этом тол...	вот_plus
2	19	Лодка бы так стояла б.	вот_plus
3	22	Они туда проехали, смотрим оттуда , типа, как...	вот_plus
4	32	А как, в принципе, он, идёт однотонно, идёт, и...	вот_plus

Рисунок 4. Тестовая выборка с контекстами слова *вот*

Сначала импортируются нужные библиотеки:

```
from gensim.utils import tokenize
from gensim import utils
from gensim.test.utils import datapath
from gensim.models import FastText
import pandas as pd
```

Далее создается итератор, который должен каждый раз выдавать список строк, где каждая строка должна быть отдельным словом:

```
class MyIter:
    def __iter__(self):
        for line in df['tmp']:
            yield list(tokenize(line))
```

Производится векторизация текста с помощью библиотеки FastText:

```
model4 = FastText(vector_size=4, window=3, min_count=1)
```

Строится словарь, состоящий из слов, находящихся в созданном корпусе текстов:

model4.build_vocab(corpus_iterable=MyIter())

Метод build_vocab() устанавливает атрибуты модели corpus_count.

total_examples = model4.corpus_count
model4.train(corpus_iterable=MyIter(), total_examples=total_examples, epochs=5)

Производится токенизация текста – процесс разделения предложений на слова-компоненты – и делается предсказание:

```
df2=pd.read_csv(r"C:\Users\Admin\Downloads\vottest.csv", sep=',')
df2['predicted']='d'
nltk.download('punkt')
for i, row  in df2.iterrows():
  z = model4.wv.most_similar(positive=word_tokenize(df2.iat[i,1]), topn=100000)
  matches0 = [match for match in z if "вот_plus" in match]
  matches1 = [match for match in z if "вот_min" in match]
  matches2 = [match for match in z if "вот_equal" in match]
  full = matches0,matches1,matches2
  b = sorted(full, key=lambda x: x[0][1])
  z = b[2]
  df2['predicted'][i] = z
  print(i)
  i=i+1
  z = ''
```

Для оценки точности классификатора используются такие метрики, как accuracy, precision, recall, F-measure и confusion matrix.

Значения точности и полноты классификатора удобнее рассчитывать с использованием confusion matrix – это матрица размера N на N, где N – это количество классов. Столбцы этой матрицы – экспертные решения, а строки – решения классификатора (Рисунок 6).

Confusion Matrix and Statistics

	Reference		
Prediction	вот_equal	вот_min	вот_plus
вот_equal	2	18	10
вот_min	0	58	7
вот_plus	3	17	6

Рисунок 6. Матрица неточностей для классификатора

Accuracy рассчитывается по следующей формуле: Accuracy = P / N, где, P – количество контекстов по которым классификатор принял правильное решение, а N – размер обучающей выборки.

Precision – это количество контекстов, принадлежащих данному классу относительно всех контекстов, которые система отнесла к этому классу. Recall – это доля найденных классификатором контекстов, принадлежащих классу относительно всех контекстов этого класса в тестовой выборке. Точность и полнота определяются следующим образом:

Precision = TP / TP + FP; Recall = TP / TP + FN, где TP – истино-положительное решение; TN – истино-отрицательное решение; FP – ложно-положительное решение; FN – ложно-отрицательное решение.

Чем выше точность и полнота, тем лучше работает классификатор. Следовательно, необходима F-мера – метрика, объединяющая в себе

информацию о точности и полноте алгоритма. Она стремится к нулю, если точность или полнота стремится к нулю. Рассчитывается по формуле: F = 2(Precision × Recall / Precision + Recall). Вышеперечисленные метрики для данного классификатора представлены в Таблице 1.

Таблица 1 – Полученные значения метрик классификатора

Class	вот_equal	вот_min	вот_plus
Precison	0.06667	0.8923	0.23077
Recall	0.40000	0.6237	0.26087
F1	0.11429	0.7342	0.24490
Accuracy	0.57931	0.6868	0.52839

Таким образом, мы видим, что данный классификатор довольно эффективен в распознавании *вот* в качестве хезитации. Это может быть связано с тем, что хезитативное *вот* в корпусе обособляется знаками препинания, а также часто используется в сочетаниях *так вот* и *вот так вот*.

Причиной того, что *вот* в указательно-уточняющей и выделительной функциях распознается хуже, может являться недостаточное количество отобранных контекстов (а также их разнородность и вариативность в самом корпусе). Также мы считаем, что стоит попробовать использовать другие популярные NLP-модели, например, RNNLM – модель, которая принимает на вход векторные представления предыдущих слов и может «понимать» семантику предложения. Обучение данной модели базируется на алгоритме непрерывного мешка слов. В ней контекстные слова подаются на вход нейронной сети, которая предсказывает центральное слово. Существует модель, цель которой прямо противоположная RNNLM – предсказать окружающие слова на основе центрального – Word2vec. Она основывается на предположении, что слова, которые часто находятся в одинаковых контекстах, имеют схожие значения. С Word2vec тесно связан алгоритм

GloVe, минимизирующий разницу между произведением векторов слов и логарифмом вероятности их совместного появления с помощью стохастического градиентного спуска.

Выводы к главе 3

Данная глава отражает результаты практической части работы, в ней описано создание классификатора, позволяющего разграничить различные значения слова *вот*, в том числе дифференцировать его как единицу, имеющую значение, и хезитатив. Для этого использовались методы машинного обучения.

Создание и апробация алгоритма содержала следующие этапы: был собран датасет, основанный на результатах анализа *вот*, представленного во второй главе и разделенный на обучающую и тестовую выборку; разработан код, импортированы необходимые для этого библиотеки; создан итератор, позволяющий выдавать список слов; проведена векторизация текста с помощью библиотеки FastText; построен словарь, включающий слова созданного корпуса текстов, проведена токенизация текста и сделано предсказание. Для оценки точности классификатора использовались accuracy, precision, recall, F-measure и confusion matrix, это показало, что данный классификатор более эффективен в распознавании *вот* в качестве хезитации (т.е в разграничении омонимии) и менее эффективен для дифференциации указательно-усилительной и выделительной функции вот (т.е. в разграничении полисемии). Данная ситуация, как представляется, разрешима при добавлении контекстов употребления *вот* в указанных значениях и использовании других популярных NLP-моделей.

ЗАКЛЮЧЕНИЕ

Диссертация посвящена решению проблемы разграничения омонимии и полисемии дискурсивного слова *вот* и выполнена в рамках разработки разметки корпуса «Бимодальный корпус устной речи тюркско-русских билингвов Южной Сибири (RuTuBiC)». Проводимое исследование отражает современную направленность лингвистики – выход за рамки традиционно лингвистических объектов исследования и методов их анализа. С одной стороны, в работе проведен структурно-семантический анализ конкретной языковой единицы с учетом зафиксированных в лексикографических источниках ее значений, а также контекстов ее использования, с другой стороны, предложен классификатор, направленный на разрешение лексической неоднозначности данной единицы в языковом корпусе. Таким образом, используются лингвистические методы и методы машинного обучения, сочетание которых представляется результативным.

Изучение лексической неоднозначности языковых единиц в данной работе сфокусировано на разграничении омонимии и полисемии дискурсивного слова *вот*, использующегося как в своих основных значениях, так и в качестве хезитатива, что представляет проблему для морфологической разметки и разметки ошибок в бимодальном корпусе устной речи тюркско-русских билингвов.

Анализ лексикографических источников продемонстрировал многообразие значений и оттенков слова, которое может употребляться в качестве частицы, местоименного наречия и связки, еще раз демонстрируя «диффузность» семантики подобных единиц. Существует также большое количество сочетаний с другими словами, употребление которых подтверждается данными корпуса, что демонстрирует контекстную зависимость выбранного для анализа слова.

Кроме того, проведенный анализ показал наличие у *вот* 9

выполняемых им функций: указательная, уточняющая, акцентирующая, финальное *вот*, маркер открытия новой темы, анафорическое *вот*, маркер возвращения к теме, маркер поиска и маркер чужой речи.

Для создания алгоритма были взяты две функции, встречающиеся в корпусе чаще, чем остальные, – указательно-уточняющая и выделительная. Анализ контекстов позволил выявить, что в указательно-уточняющей функции *вот* используется перед именительным падежом существительного или местоимения, а также в сочетании с последующими или предшествующими указательными местоимениями и наречиями. Помимо употреблений *вот* в этих двух функциях, были взяты употребления данного слова в качестве хезитатива. Анализ контекстов показал, что *вот* является хезитацией, если оно составляет отдельное предложение, стоит в начале, в середине или в конце предложения, обособляясь запятыми, или используется в выражениях *так вот* или *вот так вот* (в конце предложения или в качестве отдельного предложения).

На основе полученных данных по специфике семантики и особенностям реализации в исследуемом корпусе далее был создан алгоритм, предсказывающий употребление *вот* в том или ином значении. Формальная метрика доказывает жизнеспособность используемого алгоритма, однако, для увеличения точности следует увеличить объем разметки. Также следует попробовать применение других алгоритмов машинного обучения и векторизации, например, RNNLM, Word2vec и GloVe, результаты обращения к которым позволят повысить эффективность классификатора.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ И ЛИТЕРАТУРЫ

1. Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models // Proceedings of the International Conference on Uncertainty in Artificial Intelligence. – 2009. – Pp. 27–34.

2. Balikas G., Amini M., Clausel M. On a topic model for sentences // Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. – SIGIR ’16. – New York, NY, USA: ACM, 2016. – Pp. 921-924.

3. Blakemore Diane Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers, Cambridge University Press, 2001 – URL: https://www.academia.edu/273837/Diane_Blakemore_2002_Relevance_and_Linguistic_Meaning_the_Semantics_and_Pragmatics_of_Discourse_Markers_Cambridge_Studies_in_Linguistics_99_Cambridge_Cambridge_University_Press (дата обращения: 15.05.2021)

4. Chemudugunta C., Smyth P., Steyvers M. Modeling general and specific aspects of documents with a probabilistic topic model // Advances in Neural Information Processing Systems. – Vol. 19. – MIT Press, 2007. – Pp. 241-248.

5. Dietz L., Bickel S., Scheffer T. Unsupervised prediction of citation influences Proceedings of the 24th international conference on Machine learning. – ICML ’07. – New York, NY, USA: ACM, 2007. – Pp. 233–240.

6. Erman B Pragmatic Expressions in English: A Study of “You Know”, “You See” and “I Mean” in Face-to-Face Conversation. Almqvist & Wiksell International: Stockholm, Sweden. 1987

7. Fraser B. Pragmatics Markers // Pragmatics, 1996. – Vol.6, №2. –Pp.167-190.

8. Frazer B. An approach to discourse markers // Journal of Pragmatics. – 1990. – № 14. Pp. 386–387.

9. Halliday M. A. K. and Hasan R. Cohesion in English. London: Longman, 1976

10. Labov W. & Fanshel D. Therapeutic discourse: Psychoterapy as conversation. New York: Academic Press, 1977. – URL: https://www.researchgate.net/publication/278497337_Therapeutic_Discourse (дата обращения: 11.05.2021)

11. Schiffrin D. Discourse Markers. Cambridge: Cambridge University Press, 1987.

12. Schourup L. Discourse markers // Lingua. – 1999. – № 107(3). – 227-265.

13. U. Malik. Python for NLP: Working with Facebook FastText Library [Интернет-источник] URL: https://stackabuse.com/python-for-nlp-working-with-facebook-fasttext-library/

14. Аброскин А. А. Поиск по корпусу: проблемы и методы их решения // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы : сб. ст. – СПб. : Нестор-История, 2009. – С. 277–282.

15. Александрова О.А., Иваницкий В. В. Пауза колебания комплексный феномен современной коммуникации // Вестник НовГУ. –2003. – № 25. URL: https://cyberleninka.ru/article/n/pauza-kolebaniya-kompleksnyy-fenomen-sovremennoy-kommunikatsii (дата обращения: 20.01.2021).

16. Антонова А.Ю., Соловьев А.Н. Использование метода условных случайных полей для обработки текстов на русском языке // Информационные технологии и системы, Калининград, 2013. – С. 321-325.

17. Апресян Ю.Д., Богуславский И.М., Иомдин Б.Л. [и др.] Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003–2005. – М. :Индрик, 2005. – С. 193–214.

18. Баранов А.Н., Плунгян В.А., Рахилина Е.В. Путеводитель по дискурсивным словам русского языка. Москва: Помовский и партнеры, 1993. – 207 с.

19. Большой толковый словарь русского языка / Гл. ред. С.А. Кузнецов. – СПб.: Норинт, 1998. – 1536 с.

20. Богданова-Бегларян Н. В. Кто ищет – всегда ли найдет? (о поисковой функции вербальных хезитативов русской спонтанной речи) // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Междунар. конф. «Диалог» (Бекасово, 29 мая – 2 июня 2013 г.) / гл. ред. В. П. Селегей. М.: РГГУ, 2013. – Вып. 12(19). – С. 125–136.

21. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных : учеб. Пособие. – М.: МИЭМ, 2011. – 272 с.

22. Виноградов В.В. Русский язык : (Граммат. учение о слове) : [Учеб. пособие для высш. учеб. заведений]. – М. ; Л. : Учпедгиз, 1947. – 784 с.

23. Гришина Е. А., Ляшевская О. Н. Грамматический словарь новых слов русского языка» // Словари на основе Национального корпуса русского языка / Российская академия наук, Институт русского языка им. В.В. Виноградова РАН. – [М.], 2008. – URL: http://dict.ruslang.ru/gram.php (дата обращения: 07.04.2022).

24. Гришина Е. А. Корпус устных текстов в НКРЯ: состав и структура / Е. А. Гришина, С. О. Савчук // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы : сб. ст. – СПб.: Нестор-История, 2009. – С. 129–149.

25. Гришина Е. А. О маркерах разговорной речи (предварительное исследование подкорпуса кино в Национальном корпусе русского языка) / Е. А. Гришина // Компьютерная лингвистика и интеллектуальные технологии : труды международной конференции «Диалог'2007». Бекасово, 30 мая – 03 июня 2007 г. – М., 2007. – С. 147–156.

26. Гришина Е. А. Перспективы развития Национального корпуса русского языка / Е. А. Гришина, В. А. Плунгян // Национальный корпус русского языка: 2003–2005. Результаты и перспективы : сб. ст. – М.: Индрик,

2005. – С. 330–339.

27. Дараган Ю. В. Функции слов-«паразитов» в спонтанной речи // Тр. междунар. семинара «Диалог’2000». Т. 1: Теоретические проблемы. Протвино, 2000. – С. 67–73. URL: http://www.dialog-21.ru/Archive/2000/Dialogue%202000-1/67.htm (дата обращения 26.02.2021).

28. Дейк ван Т. Язык. Познание. Коммуникация. – Благовещенск : БГК им. И. А. Бодуэна де Куртенэ, 2000. – 308 с.

29. Ефремова Т.Ф. Новый словарь русского языка. Толково-словообразовательный. – Москва: Русский язык, 2000. URL: https://www.efremova.info/ (дата обращения: 24.01.2021).

30. Рахилина Е. В. , Кустова Г. И., Ляшевская О. Н., Резникова Т. И., Шеманаева О. Ю. Задачи и принципы семантической разметки лексики в НКРЯ // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы : сб. ст. – СПб.: Нестор-История, 2009. – С. 215–239.

31. Кибрик А. А., Подлесская В. И. Самоисправления говорящего и другие типы речевых сбоев как объект аннотирования в корпусах устной речи // Научно-техническая информация. Сер. 2. – 2007. – № 2. – С. 2-23.

32. Кобозева И.М. Полисемия дискурсивных слов и попытка ее разрешения в контексте предложения (на примере слова вот) / И.М. Кобозева // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2007». – Бекасово, 30 мая – 3 июня 2007 г. – С. 250-255.

33. Когут С. В. Дискурсивные маркеры в письменном научном дискурсе // Сибирский филологический журнал. – 2016. – № 2. – С. 157–163.

34. Беликов В. И., Копылов Н. Ю., Пиперски А. Ч., Селегей В. П., Шаров С. А. Корпус как язык: от масштабируемости к дифференциальной полноте // Компьютерная лингвистика и интеллектуальные технологии : сборник материалов ежегодной международной конференции «Диалог-2013». Бекасово, 29 мая – 02 июня 2013 г. – М., 2013. – Вып. 12 (19). – С. 84–95.

35. Коченкова Юлия Евгеньевна Цена паузы // МИРС. – 2010. – №2. URL: https://cyberleninka.ru/article/n/tsena-pauzy (дата обращения: 29.04.2022).

36. Кретов А. А. Анализ семантических помет в НКРЯ // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы : сб. ст. – СПб.: Нестор-История, 2009. – С. 240–257.

37. Кутруева Н. Г. О восприятии некоторых русских вербальных хезитативов носителями других языков // Вестник Пермского университета. Российская и зарубежная филология. – 2015. – №2 (30). URL: https://cyberleninka.ru/article/n/o-vospriyatii-nekotoryh-russkih-verbalnyh-hezitativov-nositelyami-drugih-yazykov (дата обращения: 11.03.2022).

38. Лазурский А. В., Бердичевский А.С., Крейдлин Л. Г., Митюшин Л.Г., Сизов В.Г. Интерактивное разрешение лексической и синтаксической неоднозначности в системах автоматической обработки естественного языка // Интернет-математика 2005. Автоматическая обработка веб-данных. – М., 2005. – С. 58-79.

39. Ляшевская О. Н., Плунгян В. А., Сичинава Д. В. О морфологическом стандарте Национального корпуса русского языка // Национальный корпус русского языка: 2003–2005. Результаты и перспективы : сб. ст. – М. : Индрик, 2005. – С. 111–135.

40. Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка) // Словари на основе Национального корпуса русского языка / Российская академия наук, Институт русского языка им. В.В. Виноградова РАН. – М. : Азбуковник, 2009. – URL: http://dict.ruslang.ru/freq.php (дата обращения: 07.06.2022).

41. Малов Е.М., Горбова Е.В. Дискурсивные слова в русской разговорной речи (на материале анализа спонтанной разговорной речи) // Труды Первого междисциплинарного семинара «Анализ разговорной русской речи» (АРЗ – 2007). СПб., 2007. С. 31–36.

42. Мушаев В.Н. Полисемия и вопросы многозначности (на материале калмыцкого языка) // Вестник КалмГУ. – 2018. – №4 (40). URL: https://cyberleninka.ru/article/n/polisemiya-i-voprosy-mnogoznachnosti-na-materiale-kalmytskogo-yazyka (дата обращения: 13.04.2022).

43. Найденова К.А., Невзорова О.А. Машинное обучение в задачах обработки естественного языка: обзор современного состояния исследований // Учен. зап. Казан. ун-та. Сер. Физ.-матем. науки. – 2008. – №4. URL: https://cyberleninka.ru/article/n/mashinnoe-obuchenie-v-zadachah-obrabotki-estestvennogo-yazyka-obzor-sovremennogo-sostoyaniya-issledovaniy (дата обращения: 17.06.2022).

44. Национальный корпус русского языка (НКРЯ) [Электронный ресурс] / Институт русского языка им. В. В. Виноградова РАН. – [М.], 2001–2021. – URL: http://www.ruscorpora.ru (дата обращения: 14.11.2021).

45. Новый словарь русского языка. Толково-словообразовательный / Ефремова Т. Ф. – М.: Русский язык, 2000. – 1222 с.

46. Обухова И.А. Хезитационные явления при порождении текста на родном и иностранном языках // Социо- и психолингвистические исследования. – 2021. – №9. URL: https://cyberleninka.ru/article/n/hezitatsionnye-yavleniya-pri-porozhdenii-teksta-na-rodnom-i-inostrannom-yazykah (дата обращения: 10.05.2022).

47. Порохнин А. А. Анализ статистических методов снятия омонимии в текстах на русском языке // Вестник АГТУ. Серия: Управление, вычислительная техника и информатика. – 2013. – № 2. – С. 168-174.

48. Порохнин Алексей Александрович Анализ статистических методов снятия омонимии в текстах на русском языке // Вестник АГТУ. Серия: Управление, вычислительная техника и информатика. – 2013. – №2. URL: https://cyberleninka.ru/article/n/analiz-statisticheskih-metodov-snyatiya-omonimii-v-tekstah-na-russkom-yazyke (дата обращения: 09.06.2022).

49. Рачёва А.А. Употребление маркеров вот и вон как дискурсивный выбор говорящего // Сибирский филологический журнал. – 2016. – №2. URL: https://cyberleninka.ru/article/n/upotreblenie-markerov-vot-i-von-kak-diskursivnyy-vybor-govoryaschego (дата обращения: 10.05.2022).

50. Русская грамматика : в 2-х т. / под ред. Н. Ю. Шведовой. – М. : Наука, 1980. – Т. 1 : Фонетика. Фонология. Ударение. Интонация. Словообразование. Морфология. – 783 с.

51. Сапунова Е.М. Разновидности пауз хезитации в спонтанном монологе (на материале чтения) // Материалы XXXVIII Междунар. филол. конф.СПб.: Изд-во С.-Петербург. ун-та, 2009. Вып. 22: Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков /отв. ред. А.С. Асиновский, науч. ред. Н.В. Богданова-Бегларян. – С. 138–146.

52. Словарь русского языка : в 4 х т. / под ред. А. П. Евгеньевой. – 3 е изд., стереотип. – М. : Русский язык, 1985–1988. – Т. 1 : А–Й. – 696 с.; Т. 2 : К–О. – 736 с.; Т. 3 : П–Р. – 752 с.; Т. 4 : С–Я. – 800 с.

53. Стилистический энциклопедический словарь русского языка / [Л. М. Алексеева, В. И. Аннушкин, Е. А. Баженова и др.] ; под ред. М. Н. Кожиной. – М. : Наука : Флинта, 2003. – 694 с.

54. Толковый словарь русского языка / Гл. ред. Б.М. Волин и проф. Д.Н. Ушаков. – М.: ГИС, 1935–1940. – Т. 1–4.

55. Толковый словарь русского языка [Текст] : 72500 слов и 7500 фразеологических выражений / С. И. Ожегов, Н. Ю. Шведова ; Российская АН, Ин-т рус. яз., Российский фонд культуры. – 2-е изд., испр. и доп. – Москва : Азъ, 1994.

56. Хакимов Б.Э., Гильмуллин Р.А., Гатауллин Р.Р. Разрешение грамматической многозначности в корпусе татарского языка // Учен. зап. Казан. ун-та. Сер. Гуманит. науки. – 2014. – №5. URL: https://cyberleninka.ru/article/n/razreshenie-grammaticheskoy-mnogoznachnosti-v-korpuse-tatarskogo-yazyka (дата обращения: 03.03.2022).

57. Хан Н. А. Устные спонтанные монологи разного типа в коммуникативно-дискурсивном аспекте (на материале Звукового корпуса русского языка): дис. … канд. филол. наук. – СПб., 2013. – 277 с.

58. Хан Н.А. Спонтанные монологи разного типа в коммуникативно-дискурсивном аспекте (на материале Звукового корпуса русского языка): автореф. дис. … канд. филол. наук. – СПб., 2013. – 24 с.

59. Яковлева Э.Б. Речевые хезитации: формальный и функциональный аспекты // Речевые хезитации: формальный и функциональный аспекты. – 2016. – № 2016. URL: https://cyberleninka.ru/article/n/rechevye-hezitatsii-formalnyy-i-funktsionalnyy-aspekty (дата обращения: 16.02.2022).

60. Яцко В.А. Предметная область компьютерной лингвистики // Вестник Иркутского государственного лингвистического университета. – 2014. – №2. – С. 24-35.

61. Оценка классификатора (точность, полнота F-мера): блог Дениса Боженова [Электронный ресурс]. — URL: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html (дата обращения: 07.06.2022).

ПРИЛОЖЕНИЕ А

Предметный указатель компетенций

Таблица 1 – Предметный указатель компетенций выпускной квалификационной работы магистра на тему «Разграничение полисемии и омонимии дискурсивного слова вот с помощью методов машинного обучения (на материале бимодального корпуса устной речи тюркско-русских билингвов RuTuBic)»

Компетенция	Структурный элемент работы (номера глав, параграфов работы, в которых раскрывается компетенция)
ОК-1 способностью к абстрактному мышлению, анализу, синтезу	Глава 1, глава 2, глава 3
ОК-2 готовностью действовать в нестандартных ситуациях, нести социальную и этическую ответственность за принятые решения	Введение
ОК-3 готовностью к саморазвитию, самореализации, использованию творческого потенциала	Глава 1, глава 2, глава 3
ОПК-1 способностью свободно вести коммуникацию в устной и письменной формах на государственном языке Российской Федерации и иностранном языке для решения задач профессиональной деятельности	Глава 1, глава 2, глава 3
ОПК-2 готовностью руководить коллективом в сфере своей профессиональной деятельности, толерантно воспринимая социальные, этнические, конфессиональные и культурные различия	Заключение
ОПК-3 владением знаниями об истории лингвистических учений и современного состояния лингвистической науки	Параграф 1.1., параграф 2.2
ОПК-4 способностью к осознанию современного состояния в области компьютерной лингвистики и информационных технологий	Параграф 1.2, глава 3
ОПК-5 способностью анализировать, сопоставлять и критически оценивать различные лингвистические направления, теории и гипотезы	Параграф 1.1.
ОПК-6 способностью адаптироваться к новым теориям и результатам мировой науки и расширять сферу научной деятельности, участвовать в междисциплинарных исследованиях на стыке наук	Параграф 1.2., параграф 2.1., глава 3
ОПК-7 способностью выбирать оптимальные теоретические подходы и методы решения конкретных научных задач в области лингвистики и новых информационных технологий	Глава 3
СПК-1	

способность разрабатывать проекты прикладной направленности в области прикладной и компьютерной лингвистики	Глава 3
СПК-2 способность результативно применять современные технические средства и информационные технологии в целях оптимизации лингвистической деятельности	Глава 3
ПК-1 способностью проводить самостоятельные исследования и получать новые научные результаты в области теории языка, лингвистики конкретных языков, прикладной и компьютерной лингвистики	Глава 2, глава 3
ПК-2 способностью изучать и осваивать современные технические средства и информационные технологии, служащие для обеспечения лингвистической деятельности	Глава 3
ПК-3 способностью самостоятельно планировать и проводить научные эксперименты (в том числе, при наличии подобного оборудования, с использованием высокоточных методов регистрации мозговой активности (электроэнцефалография (вызванные потенциалы) и функциональная магнитно-резонансная томография) и движений глаз)	
ПК-7 способностью разрабатывать лингвистические компоненты электронных языковых ресурсов (лингвистические корпуса, словари, онтологии, базы данных)	Глава 3
ПК-8 способностью разрабатывать системы автоматической обработки звучащей речи и письменного текста на естественном языке, лингвистические компоненты интеллектуальных и информационных электронных систем	Глава 3
ПК-9 способностью разрабатывать и внедрять в практику компьютерные системы обучения	Глава 3
ПК-10 способностью разрабатывать и совершенствовать системы автоматизации и информационной поддержки лингвистических исследований	Глава 3
ПК-11 владением навыками перевода научной литературы по лингвистике и смежным дисциплинам с иностранных языков на государственный язык Российской Федерации и с государственного языка Российской Федерации на иностранный язык со снабжением ее необходимым редакторским и издательским комментарием и научным аппаратом	Глава 1

Обработан файл:
Жильцова Н.В. Магистерская диссертация.doc.

Год публикации: 2022.

Оценка оригинальности документа - 98.77%

Процент условно корректных заимствований - 0.0%

Процент некорректных заимствований - 1.23%

Просмотр заимствований в документе

Время выполнения: 31 с.

Документы из базы

Источники

Источники заимствования	В списке литературы	Заимствования
1. **Функции слов-“паразитов” в русской спонтанной речи** Авторы: Ю.В.Дараган. Год публикации: 2000. Тип публикации: доклад на конференции. http://www.dialog-21.ru/digest/2000/articles/daragan/	–	1.23%