МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «ВГУ»)

УТВЕРЖДАЮ

Заведующий кафедрой
теоретической и прикладной лингвистики
Шилихина К.М.

10.06.2024 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
Б1.В.02 Введение в компьютерную лингвистику

1. Код и наименование направления подготовки/специальности:

10.05.04 Информационно-аналитические системы безопасности

2. Профиль подготовки/специализация: Автоматизация информационно-аналитической деятельности

3. Квалификация выпускника: специалист по защите информации

4. Форма обучения: очная

5. Кафедра, отвечающая за реализацию дисциплины: кафедра теоретической и прикладной лингвистики

6. Составители программы: Донина Ольга Валерьевна, кандидат филол. наук, доцент кафедры теоретической и прикладной лингвистики

7. Рекомендована: НМС факультета РГФ, протокол № 8 от 1 апреля 2024 г.

8. Учебный год: 2024/2025 **Семестр(ы):** 2

9.Цели и задачи учебной дисциплины

Целью освоения учебной дисциплины является ознакомление студентов с основными направлениями, а также с основными прикладными практическими задачами современной компьютерной лингвистики, с формальными, программно-реализуемыми подходами к изучению структур и закономерностей естественных языков, с возможностями применения знаний о языке в новых информационных технологиях. В рамках курса предусматривается ознакомление учащихся с современными программными средствами для решения базовых лингвистических прикладных задач.

Задачи учебной дисциплины:

- показать связь между теоретическим и прикладным лингвистическим знанием,
- сформировать у студентов терминологическую базу и навыки использования лингвистически ориентированных программных продуктов для решения практических задач, связанных с обработкой естественного языка,
- познакомить обучающихся с интеллектуальными системами в лингвистической сфере с целью обучения применения таких систем для решения прикладных задач.

10. Место учебной дисциплины в структуре ООП: Блок Б1. Дисциплины (модули). Часть, формируемая участниками образовательных отношений (вариативная).

Для успешного освоения дисциплины требуются базовые навыки работы с компьютером.

11. Планируемые результаты обучения по дисциплине/модулю (знания, умения, навыки), соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями) и индикаторами их достижения:

Код	Название компетенции	Код(ы)	Индикатор(ы)	Планируемые результаты обучения
ПК-3	Способен решать типовые задачи обработки и анализа информации в информационно-аналитических системах государственных органов, обеспечивающих национальную безопасность	ПК-3.1	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	знать: • информационно-лингвистические технологии, технологии автоматической обработки естественного языка и искусственного интеллекта • принципы работы лингвистически ориентированных программных продуктов • основные типы задач обработки и анализа естественно-языковых текстов, • основные виды автоматизированных систем обработки и анализа естественно-языковых текстов уметь: • использовать лингвистически-ориентированные программные системы • подбирать информационно-коммуникационные технологии для наиболее эффективного решения профессиональных

				задач; применять информационно-лингвистические технологии, технологии автоматической обработки естественного языка и искусственного интеллекта в соответствии с решаемой профессиональной задачей • проводить оценку качества и осуществлять выбор автоматизированной технологии обработки текстов в конкретных условиях решения прикладных информационно-аналитических задач • применять автоматизированные технологии обработки текстов при решении прикладных информационно-аналитических задач, владеть: • навыками работы с программными системами, реализующими автоматизированные технологии автоматической обработки естественного языка

12. Объем дисциплины в зачетных единицах/час. (*в соответствии с учебным планом)* — 4 з.е. / 144 ч.

Форма промежуточной аттестации*:* зачет с оценкой.

13. Трудоемкость по видам учебной работы

Вид учебной работы		Трудоемкость			
		Всего	По семестрам		
			семестр	№ семестра	…
Контактная работа					
в том числе:	лекции				
	практические				
	лабораторные				
	курсовая работа				
	др. виды(при наличии)				
Самостоятельная работа					
Промежуточная аттестация *(для экзамена)*					
Итого:					

13.1. Содержание дисциплины

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК *
		1. Лекции	
1.1	Прикладная лингвистика как отрасль научного знания	Соотношение теоретической и прикладной лингвистики. Объект и методы исследования в прикладной лингвистике. История развития прикладного лингвистического знания. Современная трактовка термина «прикладная лингвистика». Прикладная лингвистика как одно из направлений Digital Humanities и искусственного интеллекта. Основные направления прикладной лингвистики. Методы исследования в прикладной лингвистике. Основные понятия теории знаний (виды знаний, структуры представления знаний; модель мира). Разработка моделей коммуникации. Гипертекстовые технологии представления текста	
1.2	Компьютерная лингвистика как раздел прикладной лингвистики	Компьютерная лингвистика как междисциплинарное направление исследований. Задачи компьютерной лингвистики. Сложности моделирования естественного языка. Общие этапы и модули обработки текстов. Базовый терминологический аппарат компьютерной лингвистики. Основные направления компьютерной лингвистики. Подходы к построению модулей и систем компьютерной лингвистики. Признаковая модель текста (в том числе BOW – bag of words). Статистическая языковая модель. Методы выделения устойчивых словосочетаний.	
1.3	Системы автоматической обработки звучащей речи. Диалоговые системы и чат-боты	Общие проблемы автоматической обработки естественного языка. Распознавание и синтез речи. Автоматический анализ и синтез речи. Методы синтеза. Устройство TTS-синтезатора речи. Автоматическое распознавание речи. Лингвистический и статистический подходы к распознаванию речи. Автоматическая обработка устной речи. Особенности диалога на естественном языке. Тест Тьюринга. Архитектура диалоговых систем. Чат-боты. Язык AIML. Вопросно-ответные системы.	

1.4	Системы автоматической обработки письменного текста.	Автоматический анализ естественного языка. NLP (Natural Language Processing). Автоматическая обработка письменного текста. Автоматическое распознавание текста. Автоматическая генерация текстов на ЕЯ. Автоматический анализ и синтез текстов на естественном языке. Морфологический, синтаксический и семантический анализ в системах автоматической обработки текста. Синтаксический парсинг.	
1.5	Аспекты интеллектуального анализа текстов. Text Mining. Применение методов машинного обучения в лингвистике. Методы автоматической классификации	Текст как объект интеллектуального анализа. Классификация и кластеризация текстов. Алгоритмы классификации с учителем / без учителя. Векторизация. Конвейеры векторизации и преобразования. Машинное обучение в лингвистике. Формализация задач машинного обучения. Методы машинного обучения. Самообучающиеся системы: нейронные сети. Глубокое обучение. Способы автоматизации сбора информации из интернет-источников. Обучение моделей на размеченных текстах.	
1.6	Извлечение информации из текстов. Автоматическое аннотирование и реферирование текста.	Автоматизация подготовки и редактирования текстов на ЕЯ. Автоматы, формальные грамматики и языки. Контекстно-зависимый анализ текста. Извлечение информации из текстов. Распознавание сущностей. Автоматическое выявление метафор. Методы разрешения семантической неоднозначности. Разрешение анафоры и кореферентности. Установление референта. Использование экстратекстуальных сигналов. Извлечение отношений и событий. Зависимость категории от контекста.	
1.7	Теория и практика информационно-поисковых систем	Информационный поиск. Основные понятия информационного поиска. Индекс. Разработка информационно-поисковых систем. Классификация запросов. Релевантность, полнота, точность. Фильтрация и ранжирование. Факторы ранжирования. TF-IDF. Лингвистические аспекты информационного поиска. Лингвистическое обеспечение информационно-поисковых систем. Современные информационно-поисковые системы: принципы организации. Типология ИПС.	

1.8	Анализ тональности текста	Выделение мнений. Подходы к автоматическому определению тональности текста. Подход с использованием правил и словарей. Подход с использованием машинного обучения. Алгоритмы автоматического анализа тональности текста. Оценка качества работы алгоритмов. Сферы применения данных о тональности текста. Программы для определения тональности текста.	
1.9	Квантитативная лингвистика	Квантитативная лингвистика: основные понятия и методы описания языковых явлений. Структурно-вероятностная модель языка. Применение статистических данных в лингвистических исследованиях. Статистическая стилистика. Квантитативная морфология. Закон Ципфа; коэффициент Жуйана. Частотные словари (в т.ч. общая частота/ipm). Формальная семантика. Дистрибутивная семантика.	
1.10	Корпусная лингвистика. Лингвистические и филологические ресурсы и программы. Big Data	Корпусная лингвистика: история возникновения и развития. Принципы организации корпуса. Основные свойства корпуса. Классификации корпусов. Средства разметки. Типы разметки. Корпусные менеджеры. Современные корпуса текстов. Интернет как корпус. Мультимодальная лингвистика. Полевая лингвистика. Лингвистические ресурсы. Лингвистические базы данных. Построение и применение лингвистических ресурсов. Тезаурусы и онтологии. Обзор лингвистических баз данных. Примеры создания лингвистических ресурсов. Полуавтоматическое создание размеченного корпуса. Открытые данные и проблема их обработки. Верификация данных. Большие данные в филологии и языкознании.	
1.11	Компьютерная лексикография.	Компьютерная лексикография. Прикладные аспекты лексикографии. Электронные словари: основные типы словарей и принципы их организации. Структура словарной статьи электронного словаря. Электронные словари, доступные в сети. Частотные словари. Иноязычные словари. Лексическая информация в системах ИИ. Терминоведение и терминография: основные направления деятельности. Прикладные аспекты терминоведения. Теория письма; транскрипция; транслитерация.	

1.12	Автоматический перевод. Машинный перевод	Перевод как прикладная лингвистическая дисциплина. Виды перевода. Прикладные аспекты перевода. Лингвистические и нелингвистические проблемы перевода. Автоматический перевод. История создания систем автоматического перевода. Современное состояние отрасли. Машинный перевод. Машинный vs автоматизированный перевод. Определение машинного перевода. Основные подходы к машинному переводу. Основные способы перевода при помощи правил. Главная формула перевода. Модель перевода. Строение машинного перевода. Нейронный машинный перевод. Трансферный подход. Гибридный перевод. Методы оценки качества перевода. Современные системы машинного перевода. Актуальные проблемы машинного перевода. Перспективы развития машинного перевода. Методы оценки качества перевода. Сравнение систем машинного перевода.	
1.13	Лингвистическая экспертиза	Лингвистическая экспертиза: виды и области применения. Экспертные системы. Экспертиза авторства текста. Стилеметрия. Дешифровка. Глоттохронология. Лингвокриминалистика. Психолингвистика. Когнитивная лингвистика. Нейролингвистика. Автоматизированные системы психолингвистического анализа текста. Фоносемантика. Компьютерная текстология. Лингвистическая прагматика. Классификации речевых актов. Вероятностные тематические модели. Социолингвистика. Политическая лингвистика. Речевое воздействие. Языковое манипулирование. Прикладные аспекты речевого воздействия. Семиотика. Основные понятия текстологии. Автоматическое сравнение рукописей. Компьютерная классификация рукописей.	
2. Лабораторные работы			
2.1	Системы автоматической обработки письменного текста.	Форматы данных и кодировки. Регулярные выражения. Макросы (Sublime, MS Excel). Распознавание и графематический анализ текста (OCR: ABBYY Fine Reader Online, New OCR, Online OCR). Сбор информации из интернет-источников. Data Mining. Автоматизация решения DH-задач.	
2.2	Автоматическое аннотирование и	Визуализация данных (диаграммы, гистограммы, графики; облака слов, voyant	

	реферирование текста	tools, Flourish studio, Tableau). Графовые методы анализа текста, сетевой анализ (Gephi). Современные автоматизированные средства визуализации данных.	
2.3	Квантитативная лингвистика	Применение статистических методов в лингвистических исследованиях. Частотность. Закон Ципфа. Стоп-слова. Ipm.	
2.4	Корпусная лингвистика	Характеристика наиболее известных корпусов. Национальный корпус русского языка и другие русскоязычные корпусы. Корпусы иных языков. Обработка корпусных данных. Работа с параллельными корпусами (OPUS, Linguee, НКРЯ). Структура и назначение параллельных корпусов в решении задач ИИ. Google books Ngram Viewer. Выравнивание (Skuuper Cleaner, YouAlign, Champollion, Hunalign, LEOBILINGUA). Корпусные приложения (N-gram Viewer, SketchEngine), корпусные менеджеры (AntConc). Разметка. TEI. Именованные сущности. Markdown. BaseX. Язык запросов XPath	
2.5	Лингвистические ресурсы	Базы данных. Создание каталога лингвистических ресурсов	
2.6	Машинный перевод	Сравнение систем машинного перевода (Google Translate, Bing Translator, Yandex.Translate, Promt)	
2.7	Компьютерная лингвистика как раздел прикладной лингвистики	Стемминг. Лемматизация. Topic Modelling. Текстовая близость. Классификация и кластеризация. Python	
2.8	Мультимодальная лингвистика. Полевая лингвистика	Работа с видео и аудио (Shotcut, ELAN, Praat)	
2.9	Лингвистическая экспертиза	Стилеметрия (в R)	
2.10	Системы автоматической обработки звучащей речи.	Обработка звучащей речи: программы анализа речи, программы синтеза речи. Чат-боты.	
2.11	Введение в NLP (Natural language processing)	Введение в GATE и извлечение информации. JAPE. GATE Cloud services. Машинное обучение в GATE. Python GateNLP. Opinion Mining. Анализ социальных сетей. Анализ оскорблений и дезинформации в онлайн-среде.	
	Защита проекта		

** заполняется, если отдельные разделы дисциплины изучаются с помощью онлайн-курса. В колонке Примечание необходимо указать название онлайн-курса или ЭУМК. В других случаях в ячейки ставятся прочерки.*

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела) дисциплины	Виды занятий (количество часов)				
		Лекции	Практические	Лабораторные	Самостоятельная работа	Всего
1.1	Прикладная лингвистика как отрасль научного знания. Компьютерная лингвистика как раздел прикладной лингвистики	6		4	8	18
1.2	Системы автоматической обработки звучащей речи. Диалоговые системы и чат-боты	4		2	6	12
1.3	Системы автоматической обработки письменного текста.	4		2	6	12
1.4	Аспекты интеллектуального анализа текстов. Text Mining. Применение методов машинного обучения в лингвистике. Методы автоматической классификации. Введение в NLP	4		6	6	16
1.5	Извлечение информации из текстов. Автоматическое аннотирование и реферирование текста.	2		2	6	10
1.6	Теория и практика информационно-поисковых систем	2		2	6	10

1.7	Анализ тональности текста	2		2	6	10
1.8	Квантитативная лингвистика	2		2	6	10
1.9	Корпусная лингвистика. Лингвистические и филологические ресурсы и программы. Big Data	2		4	6	12
1.10	Компьютерная лексикография.	2		2	6	10
1.11	Автоматический перевод. Машинный перевод	2		2	6	10
1.12	Лингвистическая экспертиза. Мультимодальная лингвистика. Полевая лингвистика	2		4	8	14
	Итого:	34		34	76	144

14. Методические указания для обучающихся по освоению дисциплины

Необходимо регулярное посещение лекционных и лабораторных занятий, работа с литературой по дисциплине, выполнение индивидуальных лабораторных работ. Самостоятельная работа обучающихся предусматривает подготовку к аудиторным и лабораторным занятиям; выполнение домашних заданий; подготовку презентаций.

15. Перечень основной и дополнительной литературы, ресурсов интернет, необходимых для освоения дисциплины

а) основная литература:

№ п/п	Источник
1	Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.
2	Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. — М.: Изд-во НИУ ВШЭ, 2017.
3	Леонтьева Н. Н. Автоматическое понимание текстов: системы, модели, ресурсы / Н. Н. Леонтьева. - М. : Академия, 2006. – 302 с.

б) дополнительная литература:

№ п/п	Источник
4	Баранов А. Н. Лингвистическая экспертиза текста. Теоретические основания и практика : учебное пособие / А. Н. - 2-е изд. - М. : Флинта : Наука, 2009. - 591 с.

5	Баранов А. Н. Лингвистическая экспертиза текста. Теоретические основания и практика : учебное пособие / А. Н. Баранов. - 5-е изд. - Москва : Флинта : Наука, 2013. - 591 с.
6	Гуслякова А. В. Информационные технологии и лингвистика XXI века : учебное пособие / А. В. Гуслякова. - Москва : МПГУ, 2016. - 96 с. - URL: http://biblioclub.ru/index.php?page=book&id=469675 (05.08.2019).
7	Баранов А. Н. Введение в прикладную лингвистику / А.Н. Баранов; Моск. гос. ун-т им. М.В. Ломоносова. Фил. фак. — М.: Эдиториал УРСС, 2001 .— 358 с.
8	Шилихина, К.М. Основы прикладной лингвистики : учебное пособие по специальности 021800 (031301) - Теоретическая и прикладная лингвистика / К.М. Шилихина ; Воронеж. гос. ун-т .— Воронеж : ЛОП ВГУ, 2006 .— 51 с.
9	Воевудская О. М. Информационные технологии и ресурсы Интернета в профессиональной деятельности переводчика : учебное пособие / О. М. Воевудская. - Воронеж : Издательский дом ВГУ, 2018 - URL:http://www.lib.vsu.ru/elib/texts/method/vsu/m18-92.pdf.
10	Всеволодова А.В. Компьютерная обработка лингвистических баз данных: учебное пособие / А.В. Всеволодова .— 2-е изд., испр. — М. : Флинта : Наука, 2007 .— 90 с.
11	Герд А. С. Прикладная лингвистика / А.С. Герд ; С.-Петерб. гос. ун-т .— СПб. : Изд-во С.-Петерб. ун-та, 2005 .— 266 с.
12	Жданов А. А. Автономный искусственный интеллект / А. А. Жданов. - 2-е изд. - М. : БИНОМ. Лабораторий знаний, 2009. - 359 с.
13	Захаров, В.П. Корпусная лингвистика. Учебник для студентов гуманитарных вузов / В.П. Захаров ; Богданова С. Ю. – Иркутск : Иркутский государственный лингвистический университет, 2011. – 161 с. // URL: http://biblioclub.ru/index.php?page=book&id=89753
14	Захарова Т. В. Практические основы компьютерных технологий в переводе : учебное пособие / Т. В. Захарова, Е. В. Турлова. - Оренбург : Оренбургский государственный университет, 2017. - 109 с. - URL: http://biblioclub.ru/index.php?page=book&id=481823 .
15	Зиятдинова Ю. Н. Теория перевода : курс лекций : учебное пособие / Ю. Н. Зиятдинова, Э. Э. Валеева. - Казань : Издательство КНИТУ, 2009. - 118 с. - <URL:http://biblioclub.ru/index.php?page=book&id=259076>.
16	Зубов А. В. Информационные технологии в лингвистике / А.В. Зубов, И.И. Зубова .— М.: Academia, 2004 .— 205 с.
17	Калмыков А.А., Коханова Л.А. Интернет-журналистика. – М.: ЮНИТИ-ДАНА, 2005. – 383 с.
18	Копотев М. Введение в корпусную лингвистику / М. Копотев. - Прага : Animedia Company, 2014. - 195 с. - URL: http://biblioclub.ru/index.php?page=book&id=375463 (05.08.2019).
19	Ляшевская О. Н. Корпусные инструменты в грамматических исследованиях русского языка. / О.Н. Ляшевская. - Москва : Издательский Дом ЯСК : Рукописные памятники Древней Руси, 2016. - 520 с.
20	Маннинг Кристофер Д. Введение в информационный поиск = Introduction to Information retrieval / Кристофер Д. Маннинг, ПрабхакарРагхаван, ХайнрихШютце ; [пер. с англ. Д.А. Клюшина] .— М. ; СПб. ; Киев : Вильямс, 2011. - 520 с.
21	Марчук Ю. Н. Компьютерная лингвистика : учебное пособие / Ю.Н. Марчук. - М. : Восток-Запад, 2007. - 317 с.
22	Моисеева И. Ю. Квантитативная лингвистика и новые информационные технологии : учебное пособие / И. Ю. Моисеева. - Оренбург : Оренбургский государственный университет, 2017. - 103 с. - URL: http://biblioclub.ru/index.php?page=book&id=481797 (07.08.2019).

23	Новое в зарубежной лингвистике. Вып. 24. Компьютерная лингвистика. – Москва: Прогресс, 1989. – 432 с. // URL: http://biblioclub.ru/index.php?page=book&id=38638
24	Новожилова, А.А. Информационные технологии в переводе. - Учебно-метод. пособие /, А.А. Новожилова, Е.В. Степанова, Е.А. Шовгенина - Волгоград: Изд-во ВолГУ, 2012. – 162 с.
25	Онтологии и тезаурусы: модели, инструменты, приложения : учебное пособие / Б. В. Добров, В. В. Иванов, Н. В. Лукашевич, В. Д. Соловьев. - Москва : Интернет-Университет Информационных Технологий, 2009. - 173 с. - URL: http://biblioclub.ru/index.php?page=book&id=233056 (07.08.2019).
26	Потапова Р. К. Новые информационные технологии и лингвистика: учебное пособие для студ. вузов / Р.К. Потапова; Моск. гос. лингв. ун-т .— Изд. 2-е .— М.: Едиториал УРСС, 2004 .— 317 с.
27	Потапова Р. К. Новые информационные технологии и лингвистика: учебное пособие для студ. вузов / Р.К. Потапова; Моск. гос. лингв. ун-т.– Изд. 2-е. – М.: Едиториал УРСС, 2004. – 317 с. Новожилова, А.А. Информационные технологии в переводе. - Учебно-метод. пособие /, А.А. Новожилова, Е.В. Степанова, Е.А. Шовгенина - Волгоград: Изд-во ВолГУ, 2012. – 162 с.
28	Потапова Р. К. Речь: коммуникация, информация, кибернетика: Учебное пособие для студентов вузов, обуч. по специальностям "Автоматизированные системы обработки информации и управления", "Лингвистика" / Р.К.Потапова .— М.: УРСС, 2001 .— 564 с.
29	Прикладная и компьютерная лингвистика / Под ред. И.С. Николаева, О.В. Митрениной, Т.М. Ландо. – Москва: URSS, 2017. – 320 с.
30	Теория и практика машинного перевода : учебное пособие / авт.-сост. Э.В. Пиванова - Ставрополь : СКФУ, 2014. - 115 с. - URL: http://biblioclub.ru/index.php?page=book&id=457763
31	Титов В. Т. Романская квантитативная лексикология : (материалы к спецкурсу) : [учебно-методическое пособие для студентов 4 курса дневного и вечернего отделений] / В. Т. Титов ; Воронеж. гос. ун-т.- Воронеж : ЛОП ВГУ, 2006. - 43 с.
32	Berzins, K., Hudson, A. The Use of E-resources: A snapshot of e-resource use among Linking London LLN partner institutions. – London: University of East London, 2011. – URL: http://www.bbk.ac.uk/linkinglondon/resources/esystems-down-loads/report_January2011_The_Use_of_Eresources_among_Linking_London_partners_Continuum.pdf
33	Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.
34	Blei D. (2012) Introduction to Probabilistic Topic Models // Communications of the ACM. — С. 77–84.
35	Burrows, J. (2002) ‘Delta’: a measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17:267-87
36	Eder, M. Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. In: "Digital Humanities 2013: Conference Abstracts". University of Nebraska-Lincoln, Lincoln, NE, pp. 487-89.
37	Eder M. (2011) Style-markers in authorship attribution: A cross-language study of the authorial fingerprint. Studies in Polish Linguistics, 6:99–114.
38	Hoover, D. (2007) The End of the Irrelevant Text: Electronic Texts, Linguistics, and Literary Theory. Digital Humanities Quarterly 1.2.

39	Juola, P. (2006). Authorship Attribution. Foundations and Trends in Information Retrieval
40	Juola, P., Baayen, H. (2005) A Controlled-corpus Experiment in Authorship Identification by CrossEntropy, Literary and Linguistic Computing 20 (Suppl 1), pp. 59-67
41	Kestemont, M. (2014) Function words in authorship attribution. from black magic to theory? In Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL), pages 59–66, Gothenburg, Sweden.
42	Lieberman E., Michel, J.B., Jackson J., Tang T., Nowak M. (2007) Quantifying the Evolutionary Dynamics of Language. Nature. p. 449.
43	Michel, J.B., (2011). Quantitative analysis of culture using millions of digitized books . Science, 331(6014): pp. 176–82.
44	Mosteller F., Wallace D., (1963) Inference in an Authorship Problem. Journal of the American Statistical Association, Volume 58, Issue 302, pp. 275 - 309.
45	Newman M. (2010) Networks: An Introduction. Oxford: Oxford University Press.
46	Воронцов К.В. Вероятностное тематическое моделирование // http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf, 2013.
47	Мухин М. Ю. Лексическая статистика и идиостиль автора: корпусное идеографическое исследование : на материале произведений М. Булгакова, В. Набокова, А. Платонова и М. Шолохова: диссертация на соискание ученой степени доктора филологических наук. [Место защиты: ГОУВПО "Уральский государственный университет"]. - Екатеринбург, 2011. - 383 с.
48	Hamilton, William L.,Jure Leskovec, and Dan Jurafsky. "Diachronic word embeddings revealstatistical laws of semantic change." arXiv preprint arXiv:1605.09096 (2016).
49	Frermann, Lea, and Mirella Lapata. "A Bayesian Model of Diachronic Meaning Change." TACL4 (2016): 31-45.
50	Stewart, Ian, Dustin Arendt, Eric Bell, and Svitlana Volkova. "Measuring, Predicting and Visualizing ShortTerm Change in Word Representation and Usage in VKontakte Social Network." arXiv:1703.07012 (2017)
51	Hall, David, Daniel Jurafsky, and Christopher D. Manning. "Studying the history of ideas using topic models." In Proceedings of the conference on empirical methods in natural language processing, pp. 363-371. Association for Computational Linguistics, 2008.
52	Warner, Julian. Human information retrieval. Cambridge, MA: MIT Press, 2010.
53	Bryant, Fred B., and Paul R. Yarnold. "Principal-components analysis and exploratory and confirmatory factor analysis." (1995).
54	Wasserman, Stanley, and Katherine Faust. Social network analysis: Methods and applications. Vol. 8. Cambridge university press, 1994.
55	Lazer, David, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis et al. "Life in the network: the coming age of computational social science." Science (New York, NY) 323, no. 5915 (2009): 721.
56	Андреев В.С. Моделирование индивидуального стиля (на основе лингвистических характеристик): монография. М.: Флинта – Наука, 2012.
57	Введение в электронные лингвистические ресурсы [Электронный ресурс] / сост. В. Е. Гольдин, О. Ю. Крючкова. Саратов: 2011.
58	Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. Онтологии и тезаурусы: модели, инструменты, приложения. Изд-во ИНТУИТ, 2009.
59	Зубов А.В., Зубова И.И. Основы искусственного интеллекта для лингвистов: учебное пособие для вузов. Москва: Логос, 2007.

60	Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы. М., 2006
61	Мартыненко Г.Я. Основы стилеметрии. Л.: Изд-во Ленинградского ун-та, 1988.
62	Марусенко М.А. Атрибуция анонимных и псевдонимных литературных произведений методами теории распознавания образов. Л.: Изд-во Ленинградского ун-та, 1990.
63	Игнатов Г., Михальча Р. Текст Майнинг. Интеллектуальный анализ текста. Дизайн исследований, сбор данных и методы анализа / Гуманитарный центр, 2021.
64	Осипов Г. С. Методы искусственного интеллекта. М.: Физматлит, 2011. Паттерсон, Д. Глубокое обучение с точки зрения практика / Д. Паттерсон, А. Гибсон. — Москва: ДМК Пресс, 2018.
65	Прикладная и компьютерная лингвистика / Под ред. И.С. Николаева, О.В. Митрениной, Т.М. Ландо. – Москва: URSS, 2017. – 320 с.
66	Survey of Text Mining I: Clustering, Classification, and Retrieval / Ed. by M. W. Berry. — 2004. — Springer, 2003. — 261 p.
67	Aggarwal C. C., Zhai C. Mining Text Data. — Springer, 2012. — 527 p.
68	Пескова О. В. Алгоритмы классификации полнотекстовых документов // Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. — М.: МИЭМ (Московский государственный институт электроники и математики), 2011. — С. 170—212.
69	Kotu V., Deshpande B. Text Mining // Data Science: Concepts and Practice (Second Edition), 2019. – P. 281-305.
70	Kushwaha A.K., Kar A.K., Ilavarasan P.V. Predicting retweet class using deep learning // Trends in Deep Learning Methodologies: Algorithms, Applications, and Systems. Hybrid Computational Intelligence for Pattern Analysis, 2021 – P. 89-112.
71	Nettleton D. Text Analysis // Commercial Data Mining. Processing, Analysis and Modeling for Predictive Analytics Projects, 2014, - P. 171-179
72	Silge J., Robinson D. Text Mining with R // O'Reilly Media, 2017. – 194 p.
73	Федюшкин Н.А., Федосин С.А. Краткий обзор методов и моделей интеллектуального анализа текста // Проблемы и достижения в науке и технике, 2017.
74	Chang, Angel X., Manolis Savva, and Christopher D. Manning. Semantic parsing for text to 3d scene generation. // ACL 2014: 17.
75	Усталов Д., Кудрявцев А. Применение онтологии при синтезе изображения по тексту. // Доклады всероссийской научно–практической конференции Анализ Изображений, Сетей и Текстов. М.: Национальный Открытый Университет ИНТУИТ. 2012
76	Wang H., Can D., Kazemzadeh A., Bar F., Narayanan S. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. // Proceedings of the ACL 2012 System Demonstrations (pp. 115-120). Association for Computational Linguistics.
77	Coppersmith G., Kelly E. Dynamic Wordclouds and Vennclouds for Exploratory Data Analysis // Sponsor: Idibon, 22, 2014.
78	Liu S., Wang X., Chen J., Zhu J., Guo B. TopicPanorama: A full picture of relevant topics. // Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on (pp. 183-192). IEEE.
79	Alexander E., Kohlmann J., Valenza R., Witmore M., Gleicher M. Serendip: Topic model-driven visual exploration of text corpora. // Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on (pp. 173-182). IEEE.

80	Smith A., Chuang J., Hu Y., Boyd-Graber J., Findlater L. Concurrent Visualization of Relationships between Words and Topics in Topic Models // Sponsor: Idibon, 2014, 79.
81	Заславская О.Ю., Пучкова Е.С. Подходы к визуализации объемных текстовых документов // Вестник Российского университета дружбы народов. Серия: Информатизация образования. 2016. №3.

в) информационные электронно-образовательные ресурсы (официальные ресурсы интернет)*:

№ п/п	Ресурс
82	Расширение Web Scraper для Chrome: http://webscraper.io/
83	https://voyant-tools.org/
84	Программное обеспечение в области автоматической обработки текста. www.aot.ru
85	Русский ворднет http://wordnet.ru
86	Технологии Академии «Яндекс» http://company.yandex.ru/technologies/search
87	Системы автоматического аннотирования текстов www.copernic.com
88	Тезаурус английского языка WordNet http://wordnet.princeton.edu
90	База данных языков мира SIL International. URL: http://www.sil.org/
91	Национальный корпус русского языка. URL: www.ruscorpora.ru
92	The Corpus of Contemporary American English. URL: http://corpus.byu.edu
93	Лингвистика Интернета: формирование дисциплинарной парадигмы (http://www.textology.ru/article.aspx?aId=76)
94	Новые возможности лингвистических исследований по исторической семантике с применением электронных ресурсов (http://textualheritage.org/content/view/74/68/lang,ru..)
95	Языковые ресурсы: традиции и инновации (http://elib.grsu.by/katalog/161659-346552.pdf)
96	Контроль использования интернет-ресурсов (http://alexott.net/ru/writings/cf/JI200502.pdf)
97	Компьютерная лингвистика (http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvi..)
98	ICT (http://iite.unesco.org/pics/publications/en/files/321..)
99	Подбор слов с помощью тезауруса. Режим доступа: https://support.office.com/ru-ru
100	Инструкция по использования программ конкорданс. Режим доступа: https://eflnotes.wordpress.com/2013/03/06/building-your-own-corpus-textstat-antconc/
101	British National Corpus. http://www.natcorp.ox.ac.uk/
102	Атлас языков мира. http://wals.info/
103	База данных «Языки мира» www.dblang.ru
104	Интернет-портал «Historia linguisticae». http://histling.nw.ru/
105	Коллекция словарей Института русского языка им. В.В. Виноградова РАН http://slovari.ru
106	Корпуса английского языка https://www.english-corpora.org/
107	Лингвистические данные Lingistic data consortium: https://www.ldc.upenn.edu/language-resources
108	Лингвистический процессор «ЭТАП-3» http://proling.iitp.ru/ru/etap
109	Машинный фонд русского языка http://cfrl.ru
110	Новый частотный словарь русской лексики. http://dict.ruslang.ru/freq.php
111	Программное обеспечение в области автоматической обработки текста. www.aot.ru
112	Проект «Вавилонская башня» http://starling.rinet.ru

113	Словари, созданные на основе Национального корпуса русского языка. http://dict.ruslang.ru/
114	Упсальский корпус русского языка. http://www.slaviska.uu.se/korpus.htm
115	Хельсинский аннотированный корпус русских текстов: slav.helsinki.fi/hanco
116	Чешский национальный корпус. http://ucnk.ff.cuni.cz/english/index.php
117	Каталог лингвистических ресурсов CLARIN https://www.clarin.eu/
118	Каталог лингвистических ресурсов ELRA http://www.elra.info/en/.
119	Вестник Digital Humanities http://vdigital.me/
120	Программы лингвистического анализа и обработки текста http://asknet.ru/analytics/programms.htm
121	Программы анализа и лингвистической обработки текстов https://rvb.ru/soft/catalogue/c01.html
122	ACE - Automatic Content Extraction https://web.archive.org/web/20060308054306/http://www.itl.nist.gov/iad/894.01/tests/ace/
123	Интеллектуальный анализ текстов PROMT https://www.promt.ru/technology/text-analysis/
124	ТОП-5 инструментов для Text Mining http://datareview.info/article/top-5-instrumentov-dlya-text-mining/
125	Способы визуализации текстовой информации https://cs.hse.ru/vitext/visualize
126	Системы машинного перевода текстов и словари https://compress.ru/article.aspx?id=11757

* Вначале указываются ЭБС, с которыми имеются договора у ВГУ, затем открытые электронно-образовательные ресурсы, онлайн-курсы, ЭУМК

16. Перечень учебно-методического обеспечения для самостоятельной работы *(учебно-методические рекомендации, пособия, задачники, методические указания по выполнению практических (контрольных), курсовых работ и др.)*

№ п/п	Источник
127	*Донина О.В. Автоматизация лингвистических исследований : учебное пособие для вузов / О.В. Донина. – Воронеж, 2022. – 125 с.*
128	*Донина О.В. Введение в анализ текстовых данных: учебное пособие для вузов / О.В. Донина, К.А. Сидоров, Н.С. Горбунов. – Воронеж, 2022. – 105 с.*
129	*Воевудская О. М. Информационные технологии в лингвистике [Электронный ресурс] : учебное пособие для вузов / О.М. Воевудская, И.А. Терентьева. - Воронеж : ИПЦ ВГУ, 2012 - <URL:http://www.lib.vsu.ru/elib/texts/method/vsu/m12-10.pdf>.*

17. Образовательные технологии, используемые при реализации учебной дисциплины, включая дистанционные образовательные технологии (ДОТ), электронное обучение (ЭО), смешанное обучение):
При реализации дисциплины применяются технологии смешанного обучения (с использованием образовательного портала «Электронный университет ВГУ» (edu.vsu.ru)). Программа курса реализуется с применением дистанционных технологий.

Для выполнения практических заданий требуется следующее программное обеспечение: Microsoft Office, Praat, GitHub, Notepad++, Sublime, Hunalign, AntConc, Gephi, ELAN, BaseX, Java, R, R Studio, VoyantTools, GATE, KNIME, Orange, RapidMiner, wordsmith, nltk, Stanford coreNLP toolkit, word2vec.

18. Материально-техническое обеспечение дисциплины:

Для проведения лекционных и практических занятий используются аудитории, оснащенные специализированной мебелью.

Для самостоятельной работы используется класс с компьютерной техникой, оснащенный необходимым программным обеспечением, электронными учебными пособиями и законодательно – правовой и нормативной поисковой системой, имеющий выход в глобальную сеть.

19. Оценочные средства для проведения текущей и промежуточной аттестаций

Порядок оценки освоения обучающимися учебного материала определяется содержанием следующих разделов дисциплины:

№ п/п	Наименование раздела дисциплины (модуля)	Компетенция(и)	Индикатор(ы) достижения компетенции	Оценочные средства
1.	Прикладная лингвистика как отрасль научного знания. Компьютерная лингвистика как раздел прикладной лингвистики	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
2.	Системы автоматической обработки звучащей речи. Диалоговые системы и чат-боты	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
3	Системы автоматической обработки письменного текста.	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
4	Аспекты интеллектуального анализа текстов. Text Mining. Применение методов машинного обучения в лингвистике. Методы автоматической классификации. Введение в NLP	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
5	Извлечение информации из текстов.	ПК–3	Владеет способами решения типовых задач обработки и анализа	Устный опрос, выполнение индивидуальных

№ п/п	Наименование раздела дисциплины (модуля)	Компетенция(и)	Индикатор(ы) достижения компетенции	Оценочные средства
	Автоматическое аннотирование и реферирование текста.		информации в информационно-аналитических системах	лабораторных работ
6	Теория и практика информационно-поисковых систем	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
7	Анализ тональности текста	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
8	Квантитативная лингвистика	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
9	Корпусная лингвистика. Лингвистические и филологические ресурсы и программы. Big Data	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
10	Компьютерная лексикография.	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
11	Автоматический перевод. Машинный перевод	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-аналитических системах	Устный опрос, выполнение индивидуальных лабораторных работ
12	Лингвистическая экспертиза. Мультимодальная лингвистика. Полевая лингвистика	ПК–3	Владеет способами решения типовых задач обработки и анализа информации в информационно-	Устный опрос, выполнение индивидуальных лабораторных работ

№ п/п	Наименование раздела дисциплины (модуля)	Компетенция(и)	Индикатор(ы) достижения компетенции	Оценочные средства
			аналитических системах	
Промежуточная аттестация форма контроля – зачет с оценкой				*КИМ*

20 Типовые оценочные средства и методические материалы, определяющие процедуры оценивания

20.1 Текущий контроль успеваемости

Контроль успеваемости по дисциплине осуществляется с помощью следующих оценочных средств *Практикоориентированные задания/домашние задания, Сообщение/доклад/презентация*

Примерный перечень тем рефератов

- Извлечение именованных сущностей и отношений.
- Дистрибутивная семантика.
- Тематическое моделирование.
- Стилеметрия.
- К проблеме анализа лингвистических данных при помощи компьютерных методов.
- Анализ текстов с использованием открытого программного обеспечения Voyant tools.
- Цифровые технологии в гуманитарном знании.
- Metadata в структуре объекта гуманитарных исследований. Opendata в гуманитарных исследованиях. Bigdata в гуманитарных исследованиях.
- Базы данных в гуманитарных исследованиях. Цифровые архивы.
- Системы машинного перевода: история разработок, нерешенные проблемы, перспективы. История создания систем машинного перевода
- Языковые корпусы: возможности практического применения в лингвистических исследованиях. Типы языковых корпусов
- Прикладные и теоретические лингвистические модели
- Применение систем автоматического распознавания речи
- Обработка речевого сигнала в системах распознавания речи.
- Применение систем автоматического анализа текста
- Моделирование семантики высказывания как прикладная лингвистическая проблема
- Онтологии. Семантические сети
- Автоматическое распознавание текста: основные принципы, проблемы и способы их разрешения
- Основы NLP.
- Text Mining.
- Требования, предъявляемые к системам представления и обработки знаний
- Семантические сети и графы. Фреймы
- Приобретение и формализация знаний. Трудности построения баз знаний
- Методы моделирования и обучения нейронный сетей
- Семантический анализ целого текста. Анализ тональности

Примеры практикоориентированных заданий:

1) провести мини-исследование по интересующей теме, связанной с возможностями применения инструментов компьютерной лингвистики

2) сравнить системы машинного перевода
3) исследовать предложенные тексты количественными методами
4) визуализировать предложенные тексты изученными методами.
5) построить граф по заданным критериям при помощи программы сетевого анализа Gephi.
6) провести стилеметрическое исследование.
7) найти в НКРЯ слова, изменившие смысл
8) проанализировать популярность хэштегов
9) собрать digital born данные с использованием инструмента Web scraper. Каждый студент получает задание на создание парсера для интересующего его сайта. Перед созданием парсера следует изучить возможности Web scraper, понять пайплайн работы, разобраться в структуре интересующего сайта.
10) провести тематическое моделирование отзывов игроков, проанализировать их, и попытаться составить портрет игрового опыта, связав это с особенностями разных игр
11) защита проекта исследования или ресурса по компьютерной лингвистике. В процессе защиты студент должен показать глубокое знакомство с теми из основных современных методов компьютерной лингвистики, которые применяются в её/его проекте.

Требования к выполнению заданий (или шкалы и критерии оценивания):

Для оценивания результатов обучения на зачете используется 4-балльная шкала: «отлично», «хорошо», «удовлетворительно» и «неудовлетворительно».
Соотношение показателей, критериев и шкалы оценивания результатов обучения.

Критерии оценивания компетенций	Уровень сформированности компетенций	Шкала оценок
Обучающийся в полной мере демонстрирует знание информационно-лингвистических технологий, технологий автоматической обработки естественного языка и искусственного интеллекта; принципов работы лингвистически ориентированных программных продуктов; основных типов задач обработки и анализа естественно-языковых текстов; основных видов автоматизированных систем обработки и анализа естественно-языковых текстов умение использовать лингвистически-ориентированные программные системы; подбирать информационно-коммуникационные технологии для наиболее эффективного решения профессиональных задач; применять информационно-лингвистические технологии, технологии автоматической обработки естественного языка и искусственного интеллекта в соответствии с решаемой профессиональной задачей; проводить	*Повышенный уровень*	*Отлично*

оценку качества и осуществлять выбор автоматизированной технологии обработки текстов в конкретных условиях решения прикладных информационно-аналитических задач; применять автоматизированные технологии обработки текстов при решении прикладных информационно-аналитических задач, владение навыками работы с программными системами, реализующими автоматизированные технологии автоматической обработки естественного языка Обучающимся получено 80-100 баллов по итогам текущей успеваемости.		
Обучающийся имеет знание информационно-лингвистических технологий, технологий автоматической обработки естественного языка и искусственного интеллекта; принципов работы лингвистически ориентированных программных продуктов; основных типов задач обработки и анализа естественно-языковых текстов; основных видов автоматизированных систем обработки и анализа естественно-языковых текстов умение использовать лингвистически-ориентированные программные системы; подбирать информационно-коммуникационные технологии для наиболее эффективного решения профессиональных задач; применять информационно-лингвистические технологии, технологии автоматической обработки естественного языка и искусственного интеллекта в соответствии с решаемой профессиональной задачей; проводить оценку качества и осуществлять выбор автоматизированной технологии обработки текстов в конкретных условиях решения прикладных информационно-аналитических задач; применять автоматизированные технологии обработки текстов при решении прикладных информационно-аналитических задач, владение навыками работы с программными системами, реализующими автоматизированные технологии автоматической обработки естественного языка	*Базовый уровень*	*Хорошо*

Обучающимся получено не менее 70 баллов по итогам текущей успеваемости.		
Обучающийся демонстрирует частичное знание информационно-лингвистических технологий, технологий автоматической обработки естественного языка и искусственного интеллекта; принципов работы лингвистически ориентированных программных продуктов; основных типов задач обработки и анализа естественно-языковых текстов; основных видов автоматизированных систем обработки и анализа естественно-языковых текстов умение использовать лингвистически-ориентированные программные системы; подбирать информационно-коммуникационные технологии для наиболее эффективного решения профессиональных задач; применять информационно-лингвистические технологии, технологии автоматической обработки естественного языка и искусственного интеллекта в соответствии с решаемой профессиональной задачей; проводить оценку качества и осуществлять выбор автоматизированной технологии обработки текстов в конкретных условиях решения прикладных информационно-аналитических задач; применять автоматизированные технологии обработки текстов при решении прикладных информационно-аналитических задач, владение навыками работы с программными системами, реализующими автоматизированные технологии автоматической обработки естественного языка Обучающимся получено не менее 60 баллов по итогам текущей успеваемости.	*Пороговый уровень*	*Удовлетворительно*
Ответ на контрольно-измерительный материал не соответствует любым четырем из перечисленных показателей. Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки. Обучающимся получено менее 60 баллов по итогам текущей успеваемости.	–	*Не удовлетворительно*

20.2 Промежуточная аттестация

Промежуточная аттестация по дисциплине осуществляется с помощью следующих оценочных средств: *Собеседование по билетам к зачету, Практико-ориентированные задания*

Перечень вопросов к зачету

1. Автоматический анализ естественного языка. Общие проблемы автоматической обработки естественного языка.
2. Автоматическая обработка письменного текста. Автоматическое распознавание текста.
3. Текст как объект интеллектуального анализа. Классификация и кластеризация текстов.
4. Самообучающиеся системы: нейронные сети. Машинное обучение в лингвистике.
5. Визуализация данных. Современные автоматизированные средства визуализации данных.
6. Информационный поиск. Основные понятия информационного поиска. Лингвистические аспекты информационного поиска.
7. Современные информационно-поисковые системы: принципы организации. Фильтрация и ранжирование.
8. Применение статистических методов в лингвистических исследованиях. Частотность.
9. Электронные лингвистические ресурсы. Лингвистические базы данных. Построение и применение лингвистических ресурсов.
10. Цифровые библиотеки. Информационные системы в филологических задачах.
11. Открытые данные и проблема их обработки. Верификация данных. Большие данные в филологии и языкознании.
12. Корпусная лингвистика. Принципы организации корпуса.
13. Основные свойства корпуса. Классификации корпусов.
14. Современные корпуса текстов. Интернет как корпус. Национальный корпус русского языка и другие русскоязычные корпусы. Корпусы иных языков.
15. Обработка корпусных данных. Работа с параллельными корпусами. Структура и назначение параллельных корпусов в решении задач ИИ.
16. Компьютерная лексикография. Прикладные аспекты лексикографии.
17. Электронные словари: основные типы словарей и принципы их организации. Структура словарной статьи электронного словаря. Электронные словари, доступные в сети.
18. Частотные словари. Иноязычные словари. Лексическая информация в системах ИИ.
19. Автоматический перевод. История создания систем автоматического перевода. Современное состояние отрасли.
20. Машинный перевод. Машинный vs автоматизированный перевод.
21. Современные системы машинного перевода. Перспективы развития машинного перевода.
22. Методы оценки качества перевода. Сравнение систем машинного перевода. Программные средства помощи переводчику.
23. Методика контент-анализа. Проведение анкетирования.
24. Мультимодальная лингвистика. Работа с видео- и аудиоматериалами.
25. Экспертные системы. Стилеметрия в переводоведении.

Примеры практикоориентированных заданий:

1) провести мини-исследование по интересующей теме, связанной с возможностями применения инструментов компьютерной лингвистики
2) сравнить системы машинного перевода
3) исследовать предложенные тексты количественными методами

4) визуализировать предложенные тексты изученными методами.
5) построить граф по заданным критериям при помощи программы сетевого анализа Gephi.
6) провести стилеметрическое исследование.
7) найти в НКРЯ слова, изменившие смысл
8) проанализировать популярность хэштегов
9) собрать digital born данные с использованием инструмента Web scraper. Каждый студент получает задание на создание парсера для интересующего его сайта. Перед созданием парсера следует изучить возможности Web scraper, понять пайплайн работы, разобраться в структуре интересующего сайта.
10) провести тематическое моделирование отзывов игроков, проанализировать их, и попытаться составить портрет игрового опыта, связав это с особенностями разных игр
11) защита проекта исследования или ресурса по компьютерной лингвистике. В процессе защиты студент должен показать глубокое знакомство с теми из основных современных методов компьютерной лингвистики, которые применяются в её/его проекте.

Оценка знаний, умений и навыков, характеризующая этапы формирования компетенций в рамках изучения дисциплины осуществляется в ходе текущей и промежуточной аттестаций.

Текущая аттестация проводится в соответствии с Положением о текущей аттестации обучающихся по программам высшего образования Воронежского государственного университета. Текущая аттестация проводится в формах устного опроса (индивидуальный опрос, фронтальная беседа, доклады); тестирования, выполнения индивидуальных заданий. Критерии оценивания приведены выше.

Промежуточная аттестация проводится в соответствии с Положением о промежуточной аттестации обучающихся по программам высшего образования.

Контрольно-измерительные материалы промежуточной аттестации включают в себя теоретические вопросы, позволяющие оценить уровень полученных знаний, а также индивидуальные практические задания. При оценивании используются качественные шкалы оценок. Критерии оценивания приведены выше.

20.3 Фонд оценочных средств сформированности компетенций студентов, рекомендуемый для проведения диагностических работ

Перечень заданий для оценки сформированности компетенции:

1) закрытые задания (тестовые, средний уровень сложности):

1. Что из перечисленного является корпусом?
    a. Topic Modelling Tool
    b. Google forms
    c. GeoJSON
    d. НКРЯ
2. Какую программу Вы будете использовать для сетевого анализа?
    a. Gephi
    b. Linguee
    c. YouAlign
    d. Data Wrapper
3. Какой ресурс Вы будете использовать для проведения стилеметрии?

a. R
b. SketchEngine
c. N-gram Viewer
d. Gephi

4. Какую программу Вы будете использовать для анализа речи?
    a. OPUS
    b. Praat
    c. Hunalign
    d. Voyant tools
5. Какую программу Вы будете использовать для распознавания текста (OCR)?
    a. PROMT
    b. ABBYY Fine Reader Online
    c. AntConc
    d. ELAN
6. Какую программу Вы будете использовать для анализа собственного корпуса?
    a. Elibrary
    b. GitHub
    c. AntConc
    d. Sublime
7. Какую программу Вы будете использовать для построения облаков слов?
    a. Skuuper Cleaner
    b. Praat
    c. Voyant tools
    d. Zotero
8. Каким ресурсом Вы воспользуетесь, чтобы сравнить частоту использования существительного *a book* и глагола *to book* в английском языке с 1900 по 2010 гг.?
    a. Gephi
    b. N-gram Viewer
    c. Convertio
    d. Hunalign
9. Метод, создающий базу для последующего морфологического и синтаксического анализа на основе выделения слов, цифровых комплексов, формул и т. д.
    a. Семантический анализ
    b. Классификация
    c. Графематический анализ
    d. Кластеризация
10. Какая технология компьютерной лингвистики поможет сосчитать в тексте количество глаголов и прилагательных?
    a. морфологический анализатор
    b. распознавание символов
    c. анализ тональности
    d. извлечение ключевых слов
11. Какая технология компьютерной лингвистики поможет сократить текст, убрав из него все причастные и деепричастные обороты?
    a. синтаксический анализ
    b. извлечение именованных сущностей
    c. анализ жанров
    d. поиск плагиата

12. Какая технология компьютерной лингвистики поможет проанализировать комментарии к одной из нашумевших новостей и оценить отношение к событию?
    a. морфологический анализ
    b. синтаксический анализ
    c. реферирование текстов
    d. анализ тональности
13. Какая технология компьютерной лингвистики поможет проанализировать массив текстов и понять, о чем они?
    a. поиск плагиата
    b. извлечение именованных сущностей
    c. распознавание символов
    d. извлечение ключевых слов
14. Библиотека для получения векторных представлений слов на основе их совместной встречаемости в текстах.
    a. TEI
    b. word2vec
    c. tf-idf
    d. regex
15. Модель, часто используемая при обработке текстов, представляющая собой неупорядоченный набор слов, входящих в обрабатываемый текст.
    a. word2vec
    b. tf-idf
    c. doc2vec
    d. Bag of Words
16. Какое онлайн-приложение вы выберите для визуализации текста в виде облаков слов?
    a. AntConc
    b. ELAN
    c. Wordle
    d. Praat
17. Автоматическое нахождение основы слова:
    a. Лемматизация
    b. Стемминг
    c. Кластеризация
    d. Анализ тональности
18. Автоматическое приведение словоформ к начальной форме:
    a. Стемминг
    b. Разметка
    c. Лемматизация
    d. Tf-idf
19. Метрика в задачах машинного обучения, являющаяся средним гармоническим между полнотой и точностью.
    a. F-Score
    b. Precision
    c. Recall
    d. Accuracy
20. Какую программу Вы будете использовать для парсинга:
    a. SQL
    b. AntConc

c. ParseHub
d. Topic Modelling Tool

21. Преимущества языковых корпусов – это:
    a. возможность хранения неограниченного объема текстовых данных
    b. возможность многократного использования информации
    c. возможность быстрого поиска и сбора данных
    d. все перечисленное выше
22. Автоматический анализ структуры текстовых данных, который позволяет осуществлять процедуру сбора синтаксических конструкций (предложений) на естественном языке из Интернет-источников – это:
    a. Парсинг
    b. Коллокация
    c. Онтология
    d. Тезаурус
23. Сведения, которые не являются частью текста, но содержат информацию о нем:
    a. Метаданные
    b. Коллокации
    c. Стоп-слова
    d. N-граммы
24. Какую программу Вы будете использовать для выравнивания параллельных текстов?
    a. ABBYY Aligner
    b. Online OCR
    c. WIX
    d. Mendeley
25. Какой ресурс вы будете использовать для проведения тематического моделирования?
    a. YouAlign
    b. OPUS
    c. Topic Modelling Tool
    d. Linguee
26. Семантический словарь, т. е. словарь, в котором представлены смысловые связи слов — синонимические, отношения Род-Вид (иногда называемые отношением Выше-Ниже), Часть-Целое, ассоциации, пр.
    a. База данных
    b. Тезаурус
    c. Корпус
    d. Онтология
27. Набор понятий, сущностей определенной области знаний, ориентированный на многократное использование для различных задач.
    a. Онтология
    b. Корпус
    c. Лингвистическая разметка
    d. Словарь
28. Что из перечисленного является примером онтологии?
    a. GloWbE
    b. РуТез
    c. НКРЯ
    d. WordNet

2) открытые задания (тестовые, повышенный уровень сложности):

1. _____ системы – это программные системы для хранения, поиска и выдачи интересующей пользователя информации.
2. Синтаксический _____ – процедура приписывания грамматических характеристик цепочке слов.
3. _____ – это теория и практика разработки словарей.
4. Самостоятельное направление в прикладной лингвистике, ориентированное на использование компьютеров для решения задач, связанных с использованием естественного языка – это _____ лингвистика.
5. _____ поиска – отношение между количеством выданных релевантных текстов к общему количеству выданных системой текстов.
6. _____ поиска – соотношение между количеством выданных релевантных текстов или документов к общему количеству релевантных документов, имеющихся в данной информационной системе.
7. _____ лингвистика – это междисциплинарное прикладное направление, в котором объектом изучения является язык или речь, а инструментом анализа – количественные или статистические методы.
8. _____ – компьютерная коллекция текстов, специально подобранная и подготовленная для научных исследований.
9. История машинного перевода начинается с «_____ эксперимента» в январе 1954 г.
10. _____-слова – слова, знаки и символы, которые не обладают значением сами по себе, например, предлоги, частицы и местоимения.
11. _____ выражения – формальный язык поиска, основанный на использовании специальных метасимволов.
12. N-_____ – последовательности из N слов
13. _____ – этап, в рамках которого происходит разделение текста на более мелкие единицы — на предложения и слова.
14. Тематическое _____ – построение модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов.
15. _____ – это список всех употреблений заданного языкового выражения в контексте.
16. Text _____ – извлечения информации из неструктурированных текстовых данных.
17. Корпусный _____ – поисковая система по текстовым данным, которая предоставляет статистику о языковых единицах и приводит информацию в удобный для анализа вид.
18. Закон _____ гласит следующее: Если все слова языка (или просто относительно длинного текста) упорядочить по убыванию частоты их использования, то частота n-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру n (так называемому рангу этого слова).
19. _____ — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов.
20. Мера _____ – метрика, показывающая все ли правильные примеры нашел классификатор и определяющаяся как количество найденных правильных примеров к общему количеству релевантных примеров в корпусе.
21. Мера _____ – метрика, отражающая количество релевантных категории примеров к общему количеству классифицированных как правильные примеров.

22. _____ текста – определение подлинности или подложности рукописного текста и установление его автора.
23. _____ – исследование и измерение стилевых характеристик текста с целью установления авторства или получения каких-либо сведений об авторе и условиях создания текстового документа.
24. Natural Language _____ – обработка естественного языка с помощью компьютерных технологий.
25. _____ (или аннотация) – это приписывание грамматической информации о входящих в тексты словоформах.
26. _____ текст (битекст) – текст на одном языке вместе с его переводом на другой язык.
27. _____ битекста – это сопоставление соответствующих друг другу предложений одного языка с синтаксическими единицами другого.

Критерии и шкалы оценивания заданий ФОС:

Для оценивания выполнения заданий используется балльная шкала:

1) закрытые задания (тестовые, средний уровень сложности):

- 1 балл – указан верный ответ;
- 0 баллов – указан неверный ответ (полностью или частично неверный).

2) открытые задания (тестовые, повышенный уровень сложности):

- 2 балла – указан верный ответ;
- 0 баллов – указан неверный ответ (полностью или частично неверный).

Задания раздела 20.3 рекомендуются к использованию при проведении диагностических работ с целью оценки остаточных результатов освоения данной дисциплины (знаний, умений, навыков).