

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.



10.06.2023 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Б1.О.13 Инструменты искусственного интеллекта для анализа языка и текста

1. Код и наименование направления подготовки/специальности:

45.05.01 Перевод и переводоведение

2. Профиль подготовки/специализация:

Перевод и переводоведение. Английский и немецкий языки

Перевод и переводоведение. Английский и испанский языки

Перевод и переводоведение. Английский и французский языки

Перевод и переводоведение. Английский и китайский языки

Перевод и переводоведение. Английский и итальянский языки

Перевод и переводоведение. Итальянский и английский языки

Перевод и переводоведение. Немецкий и английский языки

Перевод и переводоведение. Французский и английский языки

3. Квалификация выпускника: лингвист-переводчик

4. Форма обучения: очная

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

6. Составители программы: Подтележникова Е.Н., кандидат филол. наук, доцент кафедры теоретической и прикладной лингвистики, Стародубцева Ю.А., кандидат филол. наук

7. Рекомендована: НМС факультета РГФ, протокол № 7 от 28 апреля 2023 г.

8. Учебный год: 2023/2024

Семестр(ы): 2

9. Цели и задачи учебной дисциплины

Целью изучения дисциплины является овладение с технологиями автоматической обработки естественного языка.

Задачи:

- 1) овладение знаниями в области лингвистического обеспечения различных компьютерных систем и баз данных (автоматическая обработка языка и речи, языковые корпуса, информационно-поисковые, экспертные, лексикографические системы, системы машинного перевода и др.), а также знаниями о различных источниках информации и принципах работы современных информационных технологий при решении задач профессиональной деятельности;
- 2) формирование информационно-поисковых умений (поиск, сбор, хранение, обработка и представление информации при решении задач профессиональной деятельности);
- 3) формирование навыков использования технологии автоматической обработки естественного языка и искусственного интеллекта в различных сферах деятельности для решения познавательных задач.

10. Место учебной дисциплины в структуре ООП: блок Б1. Дисциплины (модули). Обязательная часть.

Для успешного освоения дисциплины требуются базовые навыки работы с компьютером, владение лингвистической терминологией.

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
ОПК –4	Способен работать с электронными словарями, различными источниками информации, осуществлять поиск, хранение, обработку и анализ информации, представлять ее в требуемом формате с использованием информационн	ОПК-4.1	владеет знаниями о технологиях автоматической обработки естественного языка и искусственного интеллекта в различных сферах деятельности для решения познавательных задач, в том числе для анализа языка и текста	знать: принципы работы лингвистически ориентированных программных продуктов уметь: работать с электронными словарями, различными источниками информации, осуществлять поиск, хранение, обработку и анализ информации, представлять ее в требуемом формате с использованием информационных, компьютерных и сетевых технологий владеть: навыками использования различных справочных ресурсов, печатных и электронных словарей, сетевых, иных электронных и печатных источников в

	ых, компьютерных и сетевых технологий			профессиональной деятельности для поиска необходимой лингвистической и экстралингвистической информации.
ОПК-5	Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности	ОПК-5.1	ОПК-5.1 Критически оценивает современные информационные технологии	знать: основные методы использования современных информационных технологий уметь: критически оценивать информационные технологии при решении задач профессиональной деятельности владеть: навыком выбора оптимальных информационных технологий для решения задач профессиональной деятельности
ОПК-5	Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности	ОПК-5.2	Подбирает оптимальные информационные технологии при решении задач профессиональной деятельности	знать: инструментарий для поиска, сбора, хранения, обработки, представления информации при решении задач профессиональной деятельности уметь: подбирать и использовать информационные технологии при решении задач профессиональной деятельности владеть: навыком выбора оптимальных информационных технологий для решения задач профессиональной деятельности

12. Объем дисциплины в зачетных единицах/час. (в соответствии с учебным планом) — 2 з.е. / 72 ч.

Форма промежуточной аттестации: зачет.

13. Трудоемкость по видам учебной работы

Вид учебной работы		Трудоемкость		
		Всего	По семестрам	
			2 семестр	№ семестра
Контактная работа		34	34	...
в том	лекции	-	-	

числе:	практические	-	-		
	лабораторные	34	34		
	курсовая работа	-	-		
	др. виды(при наличии)				
Самостоятельная работа		38	38		
Итого:		34	34		

13.1. Содержание дисциплины

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК *
1.	Искусственный интеллект в гуманитарной сфере	Проблемы искусственного интеллекта и типы решаемых задач. Введение в цифровую филологию. Направления цифровой филологии. Digital Humanities.	
2.	Лингвистические и филологические ресурсы и программы	Электронные лингвистические ресурсы. Лингвистические базы данных. Построение и применение лингвистических ресурсов. Цифровые библиотеки. Информационные системы в филологических задачах.	
3.	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. Тезаурусы.	
4.	Корпусная лингвистика	Корпусная лингвистика. Принципы организации корпуса. Основные свойства корпуса. Классификации корпусов. Современные корпуса текстов. Интернет как корпус. Национальный корпус русского языка и другие русскоязычные корпуса. Корпусы иных языков. Обработка корпусных данных. Работа с параллельными корпусами. Структура и назначение параллельных корпусов в решении задач ИИ. Google books Ngram Viewer.	
5.	Системы автоматической	Автоматический анализ естественного языка. Общие проблемы автоматической	

	обработки письменного текста.	обработки естественного языка. Автоматическая обработка письменного текста. Автоматическое распознавание текста. Ввод, редактирование и форматирование текста (в среде Microsoft Word). Распознавание и графематический анализ текста (OCR).	
6.	Аспекты интеллектуального анализа текстов. Text Mining	Текст как объект интеллектуального анализа. Классификация и кластеризация текстов. Самообучающиеся системы: нейронные сети. Машинное обучение в лингвистике.	
7.	Визуализация текстовых данных	Визуализация данных (диаграммы, гистограммы, графики; облака слов). Современные автоматизированные средства визуализации данных.	
8.	Задачи информационного поиска с точки зрения лингвиста	Информационный поиск. Основные понятия информационного поиска. Лингвистические аспекты информационного поиска. Современные информационно-поисковые системы: принципы организации. Поиск научной литературы. Фильтрация и ранжирование.	
9.	Квантитативная лингвистика	Применение статистических методов в лингвистических исследованиях. Анализ данных в Microsoft Excel. Частотность.	
10.			
11.	ИИ в переводческой деятельности. Автоматизированный перевод. Машинный перевод	Автоматический перевод. История создания систем автоматического перевода. Современное состояние отрасли. Машинный перевод. Машинный vs автоматизированный перевод. Методы оценки качества перевода. Современные системы машинного перевода. Перспективы развития машинного перевода. Сравнение систем машинного перевода. Программные средства помощи переводчику.	
12.	ИИ в филологическом образовании	Проблемы компьютерной лингводидактики. Использование элементов ИИ в обучении языку. Компьютерные обучающие программы. Справочные системы для изучающих язык.	
13.	ИИ и социолингвистика.	Методика контент-анализа. Мультимодальная лингвистика. Проведение анкетирования. Работа с видео- и аудиоматериалами. Экспертные системы. Стилеметрия в переводе.	

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела) дисциплины	Виды занятий (количество часов)				
		Лек ции	Практи ческие	Лабораторные	Самостоятельная работа	Всего
1.1	Искусственный интеллект в гуманитарной сфере			2	2	4
1.2	Системы автоматической обработки письменного текста.			4	4	8
1.3	Аспекты интеллектуального анализа текстов. Text Mining			2	2	4
1.4	Визуализация текстовых данных			2	3	5
1.5	Задачи информационного поиска с точки зрения лингвиста			4	4	8
1.6	Квантитативная лингвистика			2	2	4
1.7	Лингвистические и филологические ресурсы и программы. Big Data			2	3	5
1.8	Корпусная лингвистика			6	6	12
1.9	Компьютерная лексикография.			2	3	5
1.10	ИИ в переводческой деятельности. Автоматизированный перевод. Машинный перевод			4	5	9
1.11	ИИ в филологическом образовании			2	2	4
1.12	ИИ и социолингвистика.			2	2	4
	Итого:			34	38	72

14. Методические указания для обучающихся по освоению дисциплины

Приступая к изучению дисциплины, обучающемуся необходимо внимательно ознакомиться с тематическим планом занятий, списком рекомендованной учебной литературы, уяснить последовательность выполнения индивидуальных (самостоятельных) учебных заданий.

Лабораторные занятия сопровождаются демонстрацией преподавателем презентационных материалов, раскрывающих основные вопросы темы. Обучающимся рекомендуется вести записи для лучшего усвоения материала.

При изучении учебной дисциплины особое внимание следует уделить приобретению навыков решения профессионально-ориентированных задач. Для этого, изучив материал данной темы, необходимо разобраться в решениях соответствующих задач, которые рассматривались на практических занятиях, приведены в учебно-методических материалах, пособиях, учебниках, обратив особое внимание на методические указания по их решению.

В качестве текущей аттестации обучающимся предлагается выполнение практических работ, обобщающих полученные знания, умения и навыки.

15. Перечень литературы, ресурсов интернет, необходимых для освоения дисциплины

№ п/п	Источник
1	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).
2	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

3	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
4	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 .
5	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).
6	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).
7	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).

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

№ п/п	Ресурс
1.	ЭБС «Университетская библиотека онлайн». – Режим доступа: по подписке. – URL: ЭБС "Университетская библиотека онлайн" читать электронные книги (biblioclub.ru)
2.	
3.	Филологический портал www.philology.ru
4.	Международный лингвистический портал The LinguistList. URL: www.linguistlist.org
5.	База данных языков мира SIL International. URL: http://www.sil.org/
6.	Электронные словари Мультитран. URL: www.multitran.ru
7.	Национальный корпус русского языка. URL: www.ruscorpora.ru
8.	The Corpus of Contemporary American English. URL: http://corpus.byu.edu
9.	Лингвистика Интернета: формирование дисциплинарной парадигмы (http://www.textology.ru/article.aspx?ald=76)

10.	Новые возможности лингвистических исследований по исторической семантике с применением электронных ресурсов (http://textualheritage.org/content/view/74/68/lang,ru..)
11.	Языковые ресурсы: традиции и инновации (http://elib.grsu.by/katalog/161659-346552.pdf)
12.	Контроль использования интернет-ресурсов (http://alexott.net/ru/writings/cf/JI200502.pdf)
13.	Интернет в работе переводчика (http://samlib.ru/w/wagapow_a_s/tissen.shtml)
14.	Компьютерная лингвистика (http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvi..)
15.	ICT (http://iite.unesco.org/pics/publications/en/files/321..)
16.	Создание автоматического оглавления в MS Word 2007. Режим доступа: http://www.mbooom.ru/news/urok_23_sozdanie_avtomaticheskogo_oglavlenija_v_ms_word_2007/2011-07-25-104
17.	Текстовое поле. Как убрать или добавить. Режим доступа: http://informatio.ru/news/programma_ms_word_2007/tekstovoe_pole_kak_ubrat_ili_dobavit_podlozhku_v_word_2007/
18.	Инструкция по использованию программ конкорданс. Режим доступа: https://eflnotes.wordpress.com/2013/03/06/building-your-own-corpus-textstat-antconc/
19.	Использование ABBY Lingvo. Режим доступа: http://www.ciao-italy.ru/materialy-i-sredstva/abbyy-lingvo-tutor-poshagovaya-instrukciya.html

16. Перечень учебно-методического обеспечения для самостоятельной работы

№ п/п	Источник
1.	Баранов А. Н. Лингвистическая экспертиза текста. Теоретические основания и практика : учебное пособие / А. Н. - 2-е изд. - М. : Флинта : Наука, 2009. - 591 с.
2.	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.
3.	Всеволодова А.В. Компьютерная обработка лингвистических баз данных: учебное пособие: для студентов, аспирантов, преподавателей-филологов / А.В. Всеволодова .— 2-е изд., испр. — М. : Флинта : Наука, 2007 .— 90 с.
4.	Ляшевская О. Н. Корпусные инструменты в грамматических исследованиях русского языка. / О.Н. Ляшевская. - Москва : Издательский Дом ЯСК : Рукописные памятники Древней Руси, 2016. - 520 с.
5.	Маннинг Кристофер Д. Введение в информационный поиск = Introduction to Information retrieval / Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце ; [пер. с англ. Д.А. Ключина] .— М. ; СПб. ; Киев : Вильямс, 2011. - 520 с.
6.	Марчук Ю. Н. Компьютерная лингвистика : учебное пособие для студентов вузов, специализирующихся по направлению и специальности "Филология" / Ю.Н. Марчук. - М. : Восток-Запад, 2007. - 317 с.
7.	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.
8.	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.
9.	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.

17. Образовательные технологии, используемые при реализации учебной дисциплины, включая дистанционные образовательные технологии (ДОТ), электронное обучение (ЭО), смешанное обучение):

При реализации дисциплины могут применяться дистанционные образовательные технологии в части освоения материала, проведения текущей аттестации, самостоятельной работы по дисциплине или отдельным ее разделам и т.д. Для освоения материала на практических занятиях используются задания, предназначенные как для индивидуального решения задач, так и для коллективного обсуждения стратегии решения той или иной задачи.

Также используется следующее программное обеспечение:

1. Неисключительные права на ПО Dr. Web Enterprise Security Suite
Комплексная защита Dr. Web Desktop Security Suite.
2. Программная система для обнаружения текстовых заимствований в учебных и научных работах Антиплагиат.ВУЗ
3. Программное обеспечение Microsoft Windows

18. Материально-техническое обеспечение дисциплины:

/ауд. 12/ - компьютерный класс: Компьютер Arbyte Tempo/AOC (12 шт.), Проектор Benq MW523 (1 шт.), Экран проекционный (1 шт.)	г.Воронеж, пл.Ленина 10, ауд.12
--	------------------------------------

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
--------------	--	----------------	-------------------------------------	--------------------

Искусственный интеллект в гуманитарной сфере Системы автоматической обработки письменного текста. Аспекты интеллектуального анализа текстов. Text Mining Визуализация текстовых данных Задачи информационного поиска с точки зрения лингвиста Квантитативная лингвистика Лингвистические и филологические ресурсы и программы. Big Data Корпусная лингвистика Компьютерная лексикография. ИИ в переводческой деятельности. Автоматизированный перевод. Машинный перевод ИИ в филологическом образовании ИИ и социолингвистика.	ОПК–4 ОПК–5	ОПК–4.1 ОПК–5.1 ОПК-5.2	Практикоориентированные задания
Промежуточная аттестация форма контроля – зачет			Теоретический вопрос, Практикоориентированное задание

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

Примеры практикоориентированных заданий:

1. Лингвистические корпуса текстов. Освоение навигации по НКРЯ

1. Найдите примеры с выбранной вами лексической единицей в параллельном подкорпусе НКРЯ. Определите, как можно применить полученную информацию при обучении переводу.
2. Автоматически выгрузите из НКРЯ в Excel корпус параллельных предложений с лексической единицей объемом 50 пар предложений. Проанализируйте переводные эквиваленты, выделите переводческие трансформации. Классифицируйте лексический материал, отсортируйте по группам. Проведите статистический анализ полученных групп, результаты представьте в виде диаграммы.

2. Машинный перевод

Сравните системы машинного перевода. Результаты анализа представьте в виде реферата в текстовом процессоре Microsoft Word, оформив его в соответствии с требованиями:

- a. Создайте титульный лист, содержание и введение в соответствии с темой вашей работы.
- b. Текст работы должен соответствовать следующим требованиям:
 - шрифт Times New Roman, размер 14, через 1,5 интервал, стиль "Обычный" (весь текст);
 - абзац - 0,5 см; выравнивание по ширине;
 - все инициалы в тексте набираются через неразрывный пробел ([Ctrl + Shift + Пробел] или команда "Вставка" - "Символ" - "Специальные символы" - "Неразрывный пробел");
 - используйте список
 - вставьте номера страниц
- c. Список литературы дается по алфавиту в конце статьи с обязательной нумерацией и озаглавляется ЛИТЕРАТУРА (выравнивание по центру);
- d. Ссылки на литературу оформляются следующим образом: Л.В. Щерба отмечал: "... [Щерба 1990, с. 1-5].

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

Для оценивания результатов обучения на зачете используется 2-балльная шкала: «зачтено», «не зачтено».

Соотношение показателей, критериев и шкалы оценивания результатов обучения.

Критерии оценивания компетенций	Уровень сформированности компетенций	Шкала оценок
Обучающийся в полной мере владеет понятийным аппаратом данной области науки (теоретическими основами дисциплины), способен иллюстрировать ответ примерами, фактами, данными научных исследований, применять теоретические знания для решения практических задач в области прикладной лингвистики.	<i>Повышенный уровень</i>	<i>Зачтено</i>
Обучающийся владеет понятийным аппаратом данной области науки (теоретическими основами дисциплины), способен иллюстрировать ответ примерами, но допускает ошибки при применении теоретических знаний для решения практических задач в области прикладной лингвистики.	<i>Базовый уровень</i>	<i>Зачтено</i>
Обучающийся владеет частично теоретическими основами дисциплины, фрагментарно способен иллюстрировать ответ примерами, фактами, данными научных исследований, не умеет применять теоретические знания для решения практических задач в области прикладной лингвистики.	<i>Пороговый уровень</i>	<i>Зачтено</i>
Ответ на контрольно-измерительный материал не соответствует любым четырем из перечисленных	—	Не зачтено

показателей. Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки.		
---	--	--

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. Проблемы искусственного интеллекта и типы решаемых задач. Гуманитарная экспертиза технологических изменений. Этические аспекты функционирования ИИ.
2. Введение в цифровую филологию. Направления цифровой филологии. Digital Humanities.
3. Автоматический анализ естественного языка. Общие проблемы автоматической обработки естественного языка.
4. Автоматическая обработка письменного текста. Автоматическое распознавание текста.
5. Текст как объект интеллектуального анализа. Классификация и кластеризация текстов.
6. Самообучающиеся системы: нейронные сети. Машинное обучение в лингвистике.
7. Визуализация данных. Современные автоматизированные средства визуализации данных.
8. Информационный поиск. Основные понятия информационного поиска. Лингвистические аспекты информационного поиска.
9. Современные информационно-поисковые системы: принципы организации. Фильтрация и ранжирование.
10. Применение статистических методов в лингвистических исследованиях. Частотность.
11. Электронные лингвистические ресурсы. Лингвистические базы данных. Построение и применение лингвистических ресурсов.
12. Цифровые библиотеки. Информационные системы в филологических задачах.
13. Открытые данные и проблема их обработки. Верификация данных. Большие данные в филологии и языкознании.
14. Корпусная лингвистика. Принципы организации корпуса.
15. Основные свойства корпуса. Классификации корпусов.
16. Современные корпуса текстов. Интернет как корпус. Национальный корпус русского языка и другие русскоязычные корпуса. Корпусы иных языков.
17. Обработка корпусных данных. Работа с параллельными корпусами. Структура и назначение параллельных корпусов в решении задач ИИ.
18. Компьютерная лексикография. Прикладные аспекты лексикографии.
19. Электронные словари: основные типы словарей и принципы их организации. Структура словарной статьи электронного словаря. Электронные словари, доступные в сети.
20. Частотные словари. Иноязычные словари. Лексическая информация в системах ИИ.
21. Автоматический перевод. История создания систем автоматического перевода. Современное состояние отрасли.

22. Машинный перевод. Машинный vs автоматизированный перевод.
23. Современные системы машинного перевода. Перспективы развития машинного перевода.
24. Методы оценки качества перевода. Сравнение систем машинного перевода. Программные средства помощи переводчику.
25. Проблемы компьютерной лингводидактики. Использование элементов ИИ в обучении языку.
26. Компьютерные обучающие программы. Справочные системы для изучающих язык.
27. Методика контент-анализа. Проведение анкетирования.
28. Мультимодальная лингвистика. Работа с видео- и аудиоматериалами.
29. Экспертные системы. Стилеметрия в переводе.

Примеры практикоориентированных заданий:

Лингвистические корпусы текстов. Освоение навигации по НКРЯ

На примере выбранной вами лексической единицы покажите организацию различных видов разметки в НКРЯ.

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 форме выполнения практикоориентированных 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 теоретический вопрос, позволяющий оценить уровень полученных знаний, а также практикоориентированное задание. При оценивании используются качественные шкалы оценок. Критерии оценивания приведены выше.

ОПК-4

Перечень заданий для проверки сформированности компетенции:

Тип I. Закрытые (мн. выбор; подбор соответствий) (*базовый уровень сложности*)

1. Совокупность специально отобранных текстов, размеченных по различным лингвистическим параметрам и обеспеченных системой поиска, называется ...

- a) базой данных;
- b) словарем;
- c) информационным массивом;
- d) корпусом**

2. Разметка бывает ...

- a) морфологической; синтаксической; семантической и просодической;**
- b) полнотекстовой и фрагментной;
- c) синхронической и диахронической;
- d) звуковой, письменной, смешанной.

3. Символ * в поисковом запросе заменяет

- a) один символ
- b) одно слово
- c) любое количество символов в словоформе**
- d) любое количество символов в предложении.

4. С появлением корпусов лингвисты получили возможность исследовать:

- a) языковую норму
- b) реальное употребление языковых единиц**
- c) ошибки в речи носителей языка
- d) частотность грамматических конструкций

5. Самый первый языковой корпус содержал

- a) 100 млн. словоупотреблений
- b) 1 млн. текстов
- c) 1 млн. словоупотреблений**
- d) 1 млн. размеченных предложений

6. Что такое метаразметка корпуса?

- a) грамматический анализ предложения
- b) информация о свойствах словоформ
- c) информация о характеристиках текста**
- d) информация о частях речи

7. Морфологическая и синтаксическая разметка обеспечивают

- a) возможность грамматического анализа предложения
- b) автоматический поиск грамматической информации**
- c) поиск точных форм слов
- d) поиск начальных форм слов

8. Специальные веб-страницы, позволяющие находить веб-ресурсы, текстовое содержание которых соответствует запросу пользователя это:

- a) тезаурусы
- b) конкордансы
- c) поисковые системы**
- d) электронные словари

9. Процесс представления содержания документа на информационно-поисковом языке называется

- a) лемматизация
- b) парсинг
- c) стемминг
- d) индексирование**

10. Процесс нахождения основы слова для заданного исходного слова называется:

- a) лемматизация
- b) парсинг
- c) стемминг**
- d) индексирование

11. Действительное соответствие содержания выданного документа содержанию запроса:

- a) смысловая релевантность**
- b) формальная релевантность
- c) поисковое предписание
- d) поисковый образ

12. Соотнесите источники данных лингвистической информации с их недостатками

1. Носители языка а) Представляют информацию о языковом явлении вне контекста
2. Интуиция лингвиста б) Не всегда дают точную информацию
3. Грамматики и словари с) Трудоемкий поиск информации
4. Тексты д) Не позволяет видеть вариативность узуса

б); 2. д); 3. а); 4. с)

13. Лексико-грамматический поиск в НКРЯ дает возможность искать информацию о
а) словоформе

б) лемме

с) нет верного варианта ответа

14. Какой корпус имеет возможность постоянного пополнения

а) специализированный

б) закрытый

с) одноязычный

д) открытый

15. Каким должен быть корпус, чтобы выводы, полученные на основе корпусного анализа, могли распространяться на использование языка в определенном языковом сообществе в конкретный период времени:

а) размеченным

б) однородным

с) репрезентативным

д) современным

16. Какие виды разметки невозможно выполнить автоматически?

а) лемматизация

б) морфологическая

с) дискурсивная

д) семантическая

е) синтаксическая

17. Лемма – это:

а) одна из возможных словоформ лексемы

б) информация о грамматических свойствах словоформы

с) информация о частеречной принадлежности слова

д) начальная форма слова

18. Точность информационного поиска – это:

а) запрос пользователя.

б) сравнение содержания документа и запроса.

в) быстрота работы информационно-поисковой системы.

г) отношение между количеством выданных релевантных текстов к общему количеству выданных системой текстов.

19. Преимущества языковых корпусов – это:

а) возможность хранения неограниченного объема текстовых данных;

б) возможность многократного использования информации;

с) возможность быстрого поиска и сбора данных;

д) все вышеперечисленное

20. Структурный компонент словаря, в который включаются все единицы, входящие в область описания словаря, называется

а) Вводная статья.

б) Лексический вход.

с) Таксон.

d) Словник.

21. Цель дескриптивного толкового словаря:

a. Показать норму употребления слова.

b. Дать наиболее полное описание лексики и все имеющиеся релевантные случаи употребления.

c. Показать норму употребления слова и дать наиболее полное описание лексики и все имеющиеся релевантные случаи употребления.

Тип II. Открытые (с кратким ответом: в формулировке задания отсутствуют варианты ответа; на заполнение пропуска) (*повышенный уровень сложности*)

1. Чему равняется X в формуле:

X = количество выданных релевантных текстов/общее количество релевантных документов в ИПС
Полнота поиска,
полноте поиска

2. Чему равняется Y в формуле:

Y = количество релевантных документов / количество всех найденных документов
Точность поиска,
точности поиска

_____ - множество документов, выдаваемых в процессе информационного поиска, формально соответствующие запросу (релевантных), но не являющиеся релевантными по смыслу.
Информационный шум

Особая разновидность словарей общей или специальной лексики, в которых указаны семантические отношения между лексическими единицами называется _____.
тезаурус, тезаурусом

Назовите типы информационно-поисковых систем (по способу поиска информации):
ручные,
автоматизированные, механизированные

Результаты поиска информации в ИПС оценивается с точки зрения _____.
Полноты, точности

Назовите модели представления данных в базах данных:
иерархическая, сетевая,
реляционная

_____ - это определенным образом организованное собрание слов с комментариями к ним.
Словарь

Отрасль прикладной лингвистики, занимающаяся созданием электронных словарей, называется _____.
компьютерная
лексикография, компьютерной лексикографией

Для улучшения возможностей поиска и отбора информации тексты, включенные в корпус, снабжаются морфологической, синтаксической и семантической _____.
разметкой

_____ - упорядоченная коллекция книг, произведений печати и письменности, снабженная средствами навигации и поиска.
Национальный корпус
языка, Языковой корпус

Анализ языковых данных корпуса _____ методом позволяет получать информацию о частотности различных языковых явлений или оценивать вероятность появления слов в определенных типах контекстов. **статистическим**

_____ разметка предполагает описание слов с точки зрения их вхождения в различные смысловые группы. **Семантическая**

В _____ языковых корпусах представлено все жанровое и хронологическое разнообразие текстов. **национальных**

_____ корпус включает в себя либо тексты определенных жанров, либо тексты, функционирующие в определенной сфере. **специализированный**

Список всех употреблений заданной словоформы в контексте, состоящем из нескольких словоформ слева и справа от заданной словоформы называется _____. **конкорданс,**

конкордансом

Какой подкорпус НКРЯ необходимо использовать для исследования перевода русского прилагательного *сердобольный* на другие языки? **параллельный**

_____ - необходимо-достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т.п. **репрезентативность**

Тип III. Открытые (с полным ответом) (*высокий уровень сложности*)

Какой подкорпус НКРЯ вы будете использовать, чтобы исследовать (а) перевод лексемы «гадкий» на изучаемый язык; (Б) употребления второго винительного падежа; (в) употребления прилагательных кОпусный и корпуснОй.

Расскажите о Национальном корпусе русского языка: его состав, структура, поисковые возможности, доступность.

ОПК 5.1

Задания	Ответы (обязательно !)
Множественный выбор, верно/неверно, на соответствие (средний уровень сложности) Стародубцева Ю.А.,	
1. Какое из следующих утверждений не относится к компьютерным обучающим программам? а) Компьютерные обучающие программы заменяют преподавателя; б) Компьютерные обучающие программы организуют и выполняют рутинную работу; с) Компьютерные обучающие программы способствуют повышению активности обучаемого; д) Компьютерные обучающие программы создают возможности для самообразования.	а) Компьютерные обучающие программы заменяют преподавателя
2. Требования к системам МП включают ... а) устойчивость, тиражируемость, адаптируемость, оптимальность временных параметров, комфорт пользователя; б) полнота, адекватность, актуальность, достоверность;	а) устойчивость, тиражируем

с) репрезентативность, полнота, экономичность, адекватность, компьютерная поддержка; д) репрезентативность, полнота, экономичность, структуризация, компьютерная поддержка.	ость, адаптируемость, оптимальность временных параметров, комфорт пользователя
3. Что не относится к процессу и понятию машинного перевода? а) междисциплинарность; б) использование машинных средств; с) принципиальное сходство этапов понимания и синтеза текста; д) учет языковых и экстралингвистических знаний.	д) учет языковых и экстралингвистических знаний
4. Что включает в себя понятие АСПОТ? а) словарь в специальном машинном формате, предназначенный для применения на ЭВМ пользователем; б) компьютерные версии хорошо известных словарей (Вебстер, Коллинз, Ожегов...); с) словарь в специальном машинном формате, предназначенный для применения на ЭВМ компьютерной программой; д) словари, предназначенные для обычного пользователя.	с) словарь в специальном машинном формате, предназначенный для применения на ЭВМ компьютерной программой
5. Электронный словарь — это ... а) введенный в компьютер бумажный словарь, снабженный средствами поиска и отображения информации; б) организованное собрание слов с комментариями, в которых описываются особенности структуры и/или функционирования этих слов; с) организованное собрание слов с описанием их значения, особенностей употребления, структурных свойств, сочетаемости, соотношения с лексическими системами других языков и т.д.; д) словарь в специальном машинном формате, предназначенный для применения на ЭВМ пользователем или компьютерной программой.	д) словарь в специальном машинном формате, предназначенный для применения на ЭВМ пользователем или компьютерной программой.
6. Одна из основных проблем компьютерного анализа речи состоит в том, что ... а) невозможно создать искусственный интеллект; б) компьютер не умеет работать со смыслом; с) у компьютера нет дополнительных источников информации (ситуация, контекст, прошлый опыт в данной области и т.п.); д) разработчики не желают делиться своими профессиональными секретами.	б) компьютер не умеет работать со смыслом;
7. Корпусный менеджер ... а) обеспечивает сортировку результатов поиска, статистические подсчеты, составление списков слов на основе корпуса; б) это специальная программа поиска по корпусу; с) это человек, составляющий корпуса и управляющий ими; д) это специальная программа подготовки текстов к их включению в корпус.	а) обеспечивает сортировку результатов поиска, статистические подсчеты, составление списков слов на основе корпуса;

8. OCR — это... a) система автоматического распознавания символов; b) система переводческой памяти; c) система машинного перевода; d) функция текстового процессора.	a) система автоматического распознавания символов;
9. К устройствам ввода данных не относится: a) сканер; b) принтер; c) клавиатура; d) цифровой фотоаппарат.	b) принтер
10. Какие из следующих приложений не являются текстовыми редакторами? a) MS Excel; b) Corel WordPerfect; c) Microsoft Word; d) Adobe InCopy.	a) MS Excel
11. Программа <i>Microsoft Word</i> позволяет: a) создавать только графические изображения; b) создавать только текстовые документы; c) создавать только текстовые документы с элементами графики; d) создавать текстовые документы с таблицами и элементами графики.	d) создавать текстовые документы с таблицами и элементами графики
12. Красную строку заданного размера можно установить: a) клавишей Tab; b) клавишей «Пробел»; c) пользуясь элементами управления на линейке; d) командой Формат=>Абзац; e) правильные варианты (в) и (г).	е) правильные варианты (в) и (г).
13. Межстрочный интервал в тексте можно изменять с помощью команды: a) таблица =>Свойства таблицы; b) Главная => Шрифт; c) Главная => Абзац; d) Найти => Заменить; e) нажимая клавишу Enter.	с) Главная => Абзац;
14. Используя буфер обмена можно: a) вставлять рисунки из графического редактора в текстовый редактор; b) дублировать фрагменты текста или графики; c) осуществлять все перечисленные действия; d) невозможно ни одно из выше перечисленных действий.	с) осуществлять все перечисленные действия;
15. В ячейке <i>Microsoft Excel</i> C1 необходимо рассчитать произведение содержимого ячеек A1 и B1 для этого в ячейке C1 нужно указать: a) A1*B1; b) =A1*B1; c) ПРОИЗВЕДИ (A1:B1); d) =ПРОИЗВЕД(A1*B1); e) ни одно из выше перечисленного.	b) =A1*B1;
16. В электронных таблицах выделена группа ячеек A1:B4. Сколько ячеек входит в эту группу? a) 4; b) 12; c) 6; d) 8; e) 13.	d) 8;
17. Начало работ по автоматическому переводу было положено a) в начале XIX века b) в 70-х годах XX века c) в начале XX века d) в 50-х годах XX века	d) в 50-х годах XX века

18. В чем заключается особенность стратегии перевода через язык посредник а) стратегия предлагает ориентироваться на узкие тематические сферы текстов. б) стратегия предлагает один или несколько промежуточных языков между структурами входного и выходного языка, на которые по соответствующим правилам переписываются выражения входного языка с) стратегия позволяет на этапе постредактирования или предредактирования включать в технологическую цепочку человека. д) стратегия предлагает замену слова входного языка на его словарный эквивалент на выходном языке.	б) стратегия предлагает один или несколько промежуточных языков между структурами входного и выходного языка, на которые по соответствующим правилам переписываются выражения входного языка
19. Использование родного языка при обучении иностранному не допускалось сторонниками а) коммуникативного метода б) аудиовизуального метода с) прямого метода обучения иностранному языку д) грамматико-переводного метода	с) прямого метода обучения иностранному языку
20. Обучение чтению и переводу текстов – это цель а) грамматико-переводного метода б) коммуникативного метода с) прямого метода обучения иностранному языку д) аудиовизуального метода	а) грамматико-переводного метода
Короткий ответ (в формулировке задания отсутствуют варианты ответа) (повышенный уровень сложности) Стародубцева Ю.А.,	
Отрасль прикладной лингвистики, занимающаяся созданием электронных словарей, называется _____.	компьютерной лексикографией
Перечислите этапы автоматического анализа текста:	графематический, морфологический, синтаксический, семантический
_____ - специальная компьютерная программа для сбора и систематизации информации, размещенной на различных интернет-ресурсах.	Парсер
_____ - процесс перевода текстов с одного естественного языка на другой полностью специальной компьютерной программой.	Машинный перевод
_____ - процесс перевода текстов человеком на компьютере с использованием компьютерных технологий.	Автоматизированный перевод
База данных, содержащая набор ранее переведенных текстов, называется _____.	Память переводов
_____ - исследование и измерение стиливых характеристик текста с целью установления авторства или получения каких-либо сведений об авторе и условиях создания текстового документа.	Стилеметрия
История машинного перевода начинается с «_____ эксперимента» в январе 1954 г.	Джорджтаунского

_____ - этап, в рамках которого происходит разделение текста на более мелкие единицы – предложения и слова.	Токенизация
_____ данных – совокупность данных, организованных по определенным правилам, предусматривающим общие принципы описания, хранения и манипулирования данными.	База
_____ речи – процесс преобразования лексемно-морфологических, синтаксических и семантических компьютерных представлений в текст на естественном языке.	Синтез
_____ система – упорядоченная совокупность документов и информационных технологий, предназначенных для хранения и поиска информации, представленной в виде текстов или их частей.	Информационно-поисковая
База данных, которую можно представить в виде таблицы с очень большим количеством строк и столбцов называется _____.	реляционная
Базы данных по своей структуре разделяются на _____.	иерархические, сетевые, реляционные
_____ программы — программы, которые борются с компьютерными вирусами и возобновляют зараженные файлы.	Антивирусные
Информационная _____ – это сохранение и защита информации, а также ее важнейших элементов, в том числе системы и оборудование, предназначенные для использования, сбережения и передачи этой информации.	безопасность
Угрозы информационной безопасности можно разделить на: _____.	естественные, искусственные, внутренние, внешние
Искусственные угрозы информационной безопасности делятся на _____ (совершаются людьми по незнанию или неосторожности) и _____ (хакерские атаки).	Непреднамеренные, преднамеренные
<i>Практико-ориентированная задача (средний уровень сложности)</i> Стародубцева Ю.А.,	
1. Расскажите о этапах автоматического анализа текста.	
2. Раскройте суть трех главных принципов, необходимых для успешного внедрения систем информационной безопасности (конфиденциальность, целостность, доступность).	

ОПК-5.2

Задания	Ответы (обязательно!)
<i>Множественный выбор, верно/неверно, на соответствие (средний уровень сложности)</i> Стародубцева Ю.А.,	
1. Компьютерный учебник — это ... а) программа, предлагающая пользователю вопрос и несколько вариантов ответов на него; б) программа формирования автоматического навыка выполнения определенных коммуникативных действий путем многочисленных повторов; в) программы, предназначенные для представления учебного материала; г) программно-методический комплекс, позволяющий самостоятельно освоить учебный курс или его большой раздел.	г) программно-методический комплекс, позволяющий самостоятельно освоить учебный курс или его большой раздел.
2. Метод автоматического аннотирования, при котором важные слова выделяются в заголовке, подзаголовке, начале и конце текста, называется ... а) статистическим;	с) позиционным

б) логико-семантическим; с) позиционным; д) функциональным.	
3. Слово, относящееся к основному содержанию текста и повторяющееся в нем несколько раз, в автоматическом реферировании называется ... а) важным словом; б) термином; с) главным словом; д) ключевым словом.	д) ключевым словом
4. Реферат — это... а) связный текст, который кратко выражает тему, предмет, цель, методы и результаты исследования; б) процесс составления содержания документа (книги, статьи, патента на изобретение и др.); с) краткое изложение содержания документа, дающее общее представление о его теме; д) краткий текст, выполняющий сигнальную функцию (информирует о том, что есть публикация на определенную тему).	а) связный текст, который кратко выражает тему, предмет, цель, методы и результаты исследования
5. Терминосистема – это... а) термины в рамках одной теории б) совокупность терминов данной отрасли производства, знания (терминология) с) все термины естественного языка (терминология).	а) термины в рамках одной теории
6. Какой вариант ответа содержит допустимые сокращения? а) т. е., т. к., т. наз., и др., и пр. (в середине предложения) б) и т. д., и т. п. (только в конце предложения), и др., и пр. (только в конце предложения), и др., и пр. (в середине предложения) с) и т. д., и т. п. (в середине предложения), и др., и пр. (в середине предложения) д) т. е., т. к., и т. д., и т. п. (только в конце предложения), и др., и пр. (только в конце предложения)	д) т. е., т. к., и т. д., и т. п. (только в конце предложения), и др., и пр. (только в конце предложения)
7. Понятие библиографирование соответствует процессу: а) доведения библиографической информации до потребителя; б) библиографического поиска с) подготовки библиографической информации.	с) подготовки библиографической информации.
8. Выделите виды библиографической систематизации: а) библиографическая классификация; б) библиографирование; с) библиографическая группировка; д) реферирование; е) рецензирование; ф) библиографическая предметизация.	а) библиографическая классификация; с) библиографическая группировка; ф) библиографическая предметизация.
9. Термин библиография используется для обозначения: а) науки; б) пособия; с) дисциплины; д) деятельности.	д) деятельности.
10. Термин библиографоведение используется для обозначения: а) науки; б) пособия; с) дисциплины; д) деятельности.	а) науки
Короткий ответ (в формулировке задания отсутствуют варианты ответа) (повышенный уровень сложности) Стародубцева Ю.А.,	
1. Процесс перехода термина из одной науки в другую называется _____.	ретерминологизацией
2. Согласно критерию числа ссылок на один и тот же документ выделяют ссылки первичные и _____.	повторные
3. По месту расположения выделяют следующие виды ссылок – затекстовая,	внутритекстовая

подстрочная и _____.	
4. _____ обзор – это связанное повествование о нескольких документах.	библиографический
5. _____ - минимальная структура библиографического описания.	элемент
6. Область _____ - это последняя область в библиографическом описании.	примечания
7. _____ библиографической записи – это элемент, расположенной перед библиографическим описанием и предназначенный для упорядочения и поиска библиографических записей.	заголовок
8. Область _____ - это область, где указывается имя издателя, распространителя и т.п.	выходных данных

Критерии и шкалы оценивания заданий для оценки сформированности компетенций:

Тип I. Закрытые (мн. выбор; подбор соответствий) (базовый уровень сложности) (0 – 1)

- 1 балл – указан верный ответ;
- 0 баллов – указан неверный ответ, в том числе частично.

Тип II. Открытые (с кратким ответом: в формулировке задания отсутствуют варианты ответа; на заполнение пропуска) (повышенный уровень сложности) (0 – 2)

- 2 балла – указан верный ответ;
- 0 баллов – указан неверный ответ, в том числе частично

Тип III. Открытые (с полным ответом) (высокий уровень сложности)

Вариант критериев и шкалы оценивания (0 – 3 – 5)

- 5 баллов – получен верный ответ; ответ полный и аргументированный;
- 3 балла – получен верный, но неполный ответ (более 50% от запрашиваемой информации предоставлено), или ответ недостаточно аргументирован (обоснован);
- 0 баллов – получен неверный ответ, или ответ содержит грубые ошибки в обосновании/аргументации

Вариант критериев и шкалы оценивания (0 – 3 – 5 – 10)

- 10 баллов – получен верный ответ; ответ полный и аргументированный;
- 5 балла – получен верный, но неполный ответ (более 75% от запрашиваемой информации предоставлено), или ответ недостаточно аргументирован (обоснован);
- 5 балла – получен верный, но неполный ответ (получено 50-75% от запрашиваемой информации предоставлено), или ответ не аргументирован (обоснован);
- 0 баллов – получен неверный ответ, или ответ содержит грубые ошибки в обосновании/аргументации