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.
10.06.2022 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Б1.В.04 Технологии корпусной лингвистики

1. Код и наименование направления подготовки/специальности:

45.03.03 Фундаментальная и прикладная лингвистика

2. Профиль подготовки/специализация:

Экспертно-аналитическая деятельность

3. Квалификация выпускника: бакалавр

4. Форма обучения: очная

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

6. Составители программы: Шилихина Ксения Михайловна, доктор филол. наук, доцент кафедры теоретической и прикладной лингвистики

7. Рекомендована: Научно-методическим советом факультета РГФ, протокол № 8 от 23.05.2022 г.

8. Учебный год: 2024/2025 **Семестр:** 5

9. **Цели и задачи учебной дисциплины**

Целями освоения учебной дисциплины являются:

- формирование у студентов навыков практического использования корпусных данных в лингвистических исследованиях, а также умения создавать языковые корпуса, осуществлять различные виды разметки (морфологическую, синтаксическую, семантическую, дискурсивную) с помощью компьютера и вручную.

Задачи учебной дисциплины:

- обучение работе с различными компьютерными программами, которые используются при создании корпусов,
- ознакомление со статистическими методами и приемами обработки корпусных данных, а также способами лингвистической интерпретации числовых данных.

10. Место учебной дисциплины в структуре ООП: дисциплина Б1.В.04 Технологии корпусной лингвистики относится к блоку «Дисциплины (модули)» Федерального государственного образовательного стандарта высшего образования по направлению подготовки 45.03.03 Фундаментальная и прикладная лингвистика и входит в часть, формируемую участниками образовательных отношений. Для ее успешного освоения необходимы базовые знания, умения и навыки, сформированные в процессе изучения дисциплин Б1.О.11 «Алгебра и начала анализа», Б1.О.12 «Математическая логика», Б1.О.13 Теория вероятностей, Б1.О.14 Математическая статистика. Изучение данной дисциплины предшествует освоению дисциплин Б1.О.27 Основные проблемы современной лингвистики, Б1.В.01 Проектирование баз данных, Б1.В.02 Автоматическая обработка естественного языка, Б1.В.05 Анализ данных для лингвиста, Б1.В.06 Лингвистическая экспертиза текста, Б1.В.ДВ.05.01 Общая и компьютерная лексикография, Б1.В.ДВ.06.01 Компьютерная лингвистика, Б1.В.ДВ.06.02 Квантитативная лингвистика, ФТД.03 Анализ медиатекстов, ФТД.04 Компьютерный анализ звучащей речи.

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	Коды	Индикаторы	Планируемые результаты обучения
ПК-1	Способен спланировать и провести лингвистический эксперимент, описать его результаты и сформулировать выводы	ПК-1.1	Собирает, обрабатывает, анализирует и обобщает результаты экспериментов и исследований в соответствующей области знаний	Знать: общие положения о специфике и архитектуре эксперимента; базовые принципы проведения экспериментов с носителями языка; принципы составления анкет и вопросников; методы создания лингвистических экспериментов с привлечением достижений современных технологий; имеет базовые представления о методах математической статистики, используемых при обработке результатов эксперимента. Уметь: разработать и провести базовый лингвистический эксперимент; отобрать испытуемых; обобщить и проанализировать полученные данные; сформулировать результат. Владеть: опытом проведения базовых лингвистических экспериментов; разработки их архитектуры; поиска

				испытуемых; обработки результатов.
ПК-2	Владеет основными способами описания и формальной репрезентации денотативной, концептуальной, коммуникативной и прагматической информации, содержащейся в тексте на естественном языке	ПК-2.1	Представляет денотативную, концептуальную, коммуникативную и прагматическую информацию, содержащуюся в тексте на ЕЯ, формальными методами	Знать: основы формального моделирования лингвистических объектов, компьютерной обработки информации и создания компьютерных лингвистических приложений. Уметь: различать основные типы формальных моделей описания естественного языка, структурировать и моделировать базовые явления языка с использованием математического аппарата и представлять в алгоритмическом виде процессы анализа и синтеза текста/дискурса; работать с существующими системами представления знаний. Владеть: методами формализации и алгоритмизации, применяемыми в лингвистике; корпусными методами работы с языковым материалом, гипертекстовыми технологиями, подходами к построению различных словарей и лингвистических баз данных
ПК-4	Владеет базовыми навыками доработки и обработки (например, корректура, редактирование, комментирование, реферирование, информационно-словарное описание) различных типов текстов, навыками сбора, мониторинга и предоставления информации	ПК-4.3	Составляет информационно-словарное описание языковых единиц	Знать: основы стилистики, корректирования и редактирования; имеет представление о словарях и справочниках в избранной сфере профессиональной деятельности. Уметь: вести редактуру и корректуру текста, осуществлять первичный реальный комментарий к тексту, собирать и интерпретировать информацию из различных источников, редактировать материалы для СМИ и веб-сайтов, материалы и документы, обеспечивающие работу руководителя. Владеть: Навыком создания различных типов текста; сбора, обработки и систематизации информации.
ПК-14	Владеет принципами создания электронных языковых ресурсов (текстовых, речевых и мультимодальных корпусов; словарей, тезаурусов, онтологий; фонетических, лексических, грамматических и	ПК-14.2	Пользуется электронными языковыми ресурсами для решения прикладных задач	Знать: методики поиска, анализа и обработки материала исследования Уметь: работать с различными источниками информации Владеть: навыками реферирования, формулирования целей, задач, методов, выводов научного исследования

	иных баз данных и баз знаний) и умеет пользоваться такими ресурсами			

12. Объем дисциплины в зачетных единицах/час. — 4 з.е. /144 ч.

Форма промежуточной аттестации: экзамен.

13. Трудоемкость по видам учебной работы

Вид учебной работы		Трудоемкость	
		Всего	По семестрам
			5 семестр
Аудиторные занятия		**54**	**54**
в том числе:	лекции	18	18
	практические	-	-
	лабораторные	36	36
Самостоятельная работа		**54**	**54**
в том числе: курсовая работа (проект)		-	-
Форма промежуточной аттестации *(экзамен – 36 час.)*		**36**	Экзамен
Итого:		**144**	**144**

13.1. Содержание дисциплины

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК *
		1. Лекции	
1.1	Корпусная лингвистика как прикладная дисциплина	Корпус как источник лингвистической информации. Преимущества и недостатки корпусов по сравнению с другими источниками данных. Прикладной характер корпусной лингвистики.	
1.2	Типология языковых корпусов.	Типы корпусов. Состав и структура корпусов различных типов.	
1.3	Анализ корпусных данных	Конкордансеры. AntConc. Sketch Engine. Информативные возможности конкордансеров.	
1.4	Статистические методы в корпусных исследованиях	Частотность употребления языковых единиц. Абсолютная и относительная частота. Индекс взаимной информации. Другие статистические методы исследования корпусных данных	
1.5	Параллельные корпуса в сопоставительных исследованиях	Параллельный корпус как источник данных для межъязыковых сопоставлений. Лексические и грамматические данные.	
1.6	Корпусные данные в исследованиях неологизмов и заимствований	Орфографическая и грамматическая вариативность новой лексики. Мониторинг использования новых слов в текстах разных жанров.	
1.7	Корпуса в лексикографии	Возможности использования корпусов в описании слов. Выделение значений на основе корпусных данных.	
1.8	Грамматические исследования на базе корпусов	Corpus-driven и corpus-based подходы к изучению грамматики. Синтаксические корпуса (treebanks). Анализ конкретных грамматических исследований на базе корпусных данных.	

1.9	Изучение исторических изменений языка на базе корпусов	Исторический корпус. Примеры исторических корпусов. Хронологический анализ языковых данных.	
2. Практические занятия			
2.1			
2.2			
3. Лабораторные занятия			
3.1	Корпусная лингвистика как прикладная дисциплина	Корпус как источник лингвистической информации. Преимущества и недостатки корпусов по сравнению с другими источниками данных. Прикладной характер корпусной лингвистики.	
3.2	Краткая история корпусной лингвистики	Брауновский корпус. LOB. Поисково-информационные возможности первых корпусов.	
3.3	Основные корпуса русского языка	Уппсальский корпус. Тюбингенский корпус. ХАНКО. Национальный корпус русского языка. RuSKELL	
3.4	Основные корпуса английского и др. языков	Британский национальный корпус. Корпуса М. Дэвиса. «Срезовые» (snapshot) корпуса.	
3.5	Типология языковых корпусов.	Типы корпусов. Состав и структура корпусов различных типов.	
3.6	Принципы формирования корпусов	Сбалансированность. Репрезентативность. Критерии отбора текстов в корпуса разных типов.	
3.7	Метаразметка	Метаразметка как элемент корпуса. Стандарты метаразметки.	
3.8	Разметка текстов	Виды разметки. Морфологическая разметка. Синтаксическая разметка. Семантическая разметка. Международные стандарты разметки.	
3.9	Анализ корпусных данных	Конкордансеры. AntConc. Sketch Engine. Информативные возможности конкордансеров.	
3.10	Статистические методы в корпусных исследованиях	Частотность употребления языковых единиц. Абсолютная и относительная частота. Индекс взаимной информации. Другие статистические методы исследования корпусных данных	
3.11	Параллельные корпуса в сопоставительных исследованиях	Параллельный корпус как источник данных для межъязыковых сопоставлений. Лексические и грамматические данные.	
3.12	Корпусные данные в исследованиях неологизмов и заимствований	Орфографическая и грамматическая вариативность новой лексики. Мониторинг использования новых слов в текстах разных жанров.	
3.13	Корпуса в лексикографии	Возможности использования корпусов в описании слов. Выделение значений на основе корпусных данных.	
3.14	Грамматические исследования на базе корпусов	Corpus-driven и corpus-based подходы к изучению грамматики. Синтаксические корпуса (treebanks). Анализ конкретных грамматических исследований на базе корпусных данных.	
3.15	Изучение исторических изменений языка на базе корпусов	Исторический корпус. Примеры исторических корпусов. Хронологический анализ языковых данных.	
3.16	Использование корпусов в преподавании иностранных языков	Учебные корпуса. Анализ лексической сочетаемости, грамматических свойств слов изучаемого языка как способ активного постижения ИЯ. 2	

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	Корпусная лингвистика как прикладная дисциплина	2		2	2	6
2	Краткая история	-		2	2	4

	корпусной лингвистики					
3	Основные корпуса русского языка	-		2	2	4
4	Основные корпуса английского и др. языков	-		2	2	4
5	Типология языковых корпусов.	2		2	3	7
6	Принципы формирования корпусов	-		2	2	4
7	Метаразметка	-		2	3	5
8	Разметка текстов	-		2	3	5
9	Анализ корпусных данных	2		4	4	10
10	Статистические методы в корпусных исследованиях	2		4	3	9
11	Параллельные корпуса в сопоставительных исследованиях	2		2	3	7
12	Корпусные данные в исследованиях неологизмов и заимствований	2		2	3	7
13	Корпуса в лексикографии	2		2	5	9
14	Грамматические исследования на базе корпусов	2		2	5	9
15	Изучение исторических изменений языка на базе корпусов	2		2	5	9
16	Использование корпусов в преподавании иностранных языков	-		2	7	9
	Итого:	18		36	54	108

14. Методические указания для обучающихся по освоению дисциплины

Для изучения разделов данной учебной дисциплины необходимо вспомнить и систематизировать знания, полученные ранее по лингвистике.

При изучении материала учебной дисциплины по учебнику нужно, прежде всего, уяснить существо каждого излагаемого там вопроса. Главное - это понять изложенное в учебнике, а не «заучить».

Изучать материал рекомендуется по темам конспекта лекций и по главам (параграфам) учебника (учебного пособия). Сначала следует прочитать весь материал темы (параграфа), особенно не задерживаясь на том, что показалось не совсем понятным: часто это становится понятным из последующего. Затем надо вернуться к местам, вызвавшим затруднения и внимательно разобраться в том, что было неясно.

Особое внимание при повторном чтении необходимо обратить на формулировки соответствующих определений, формулы и т.п. (они обычно бывают набраны в учебнике курсивом); в точных формулировках, как правило, существенно каждое слово и очень полезно понять, почему данное положение сформулировано именно так. Однако не следует стараться заучивать формулировки; важно понять их смысл и уметь изложить результат своими словами.

Закончив изучение раздела, полезно составить краткий конспект, по возможности, не заглядывая в учебник (учебное пособие).

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

по их решению. Затем необходимо самостоятельно решить несколько аналогичных задач из сборников задач, приводимых в разделах рабочей программы, и после этого решать соответствующие задачи из сборников тестовых заданий и контрольных работ.

Закончив изучение раздела, нужно проверить умение ответить на все вопросы программы курса по этой теме (осуществить самопроверку).

Все вопросы, которые должны быть изучены и усвоены, в программе перечислены достаточно подробно. Однако очень полезно составить перечень таких вопросов самостоятельно (в отдельной тетради) следующим образом:

– начав изучение очередной темы программы, выписать сначала в тетради последовательно все перечисленные в программе вопросы этой темы, оставив справа широкую колонку;

– по мере изучения материала раздела (чтения учебника, учебно-методических пособий, конспекта лекций) следует в правой колонке указать страницу учебного издания (конспекта лекции), на которой излагается соответствующий вопрос, а также номер формулы, которые выражают ответ на данный вопрос.

В результате в этой тетради будет полный перечень вопросов для самопроверки, который можно использовать и при подготовке к экзамену. Кроме того, ответив на вопрос или написав соответствующую формулу (уравнение), можете по учебнику (конспекту лекций) быстро проверить, правильно ли это сделано, если в правильности своего ответа Вы сомневаетесь. Наконец, по тетради с такими вопросами Вы можете установить, весь ли материал, предусмотренный программой, Вами изучен.

Следует иметь в виду, что в различных учебных изданиях материал может излагаться в разной последовательности. Поэтому ответ на какой-нибудь вопрос программы может оказаться в другой главе, но на изучении курса в целом это, конечно, никак не скажется.

Указания по выполнению тестовых заданий и контрольных работ приводятся в учебно-методической литературе, в которых к каждой задаче даются конкретные методические указания по ее решению и приводится пример решения.

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

а) основная литература:

№ п/п	Источник
1	Учебник по лексикологии / Е. А. Лукьянова, И.В. Толочин, М.Н. Коновалова, М.В. Сорокина ; под ред. И.В. Толочина. – Санкт-Петербург : Антология, 2014. – 352 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=257920
2	Ляшевская О. Н. Корпусные инструменты в грамматических исследованиях русского языка. / О.Н. Ляшевская. - Москва : Издательский Дом ЯСК : Рукописные памятники Древней Руси, 2016. - 520 с. - URL: http://biblioclub.ru/index.php?page=book&id=473302

б) дополнительная литература:

№ п/п	Источник
3	Копотев, М. Введение в корпусную лингвистику / М. Копотев. - Прага : Animedia Company, 2014. - 195 с. : ил., табл. - ISBN 978-80-7499-067-0 ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=375463
4	Грудева, Е.В. Корпусная лингвистика : учебное пособие / Е.В. Грудева ; науч. ред. Л.Н. Чурилина. – 3-е изд., стер. – Москва : ФЛИНТА, 2017. – 166 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=364207
5	Ляшевская, О.Н. Корпусные инструменты в грамматических исследованиях русского языка / О.Н. Ляшевская. – Москва : Языки славянской культуры (ЯСК) : Рукописные памятники Древней Руси, 2016. – 520 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=473302
6	Щипицина, Л.Ю. Информационные технологии в лингвистике : учебное пособие : [16+] / Л.Ю. Щипицина. – Москва : ФЛИНТА, 2013. – 127 с. : табл. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=375745
7	Формализация исследовательских процедур анализа семантики языковых единиц / М.В. Каменский, Т.Н. Ломтева, Н.С. Кабылкина и др. ; под общ. ред. М.В. Каменского ; Северо-Кавказский федеральный университет. – Ставрополь : Северо-Кавказский Федеральный университет (СКФУ), 2016. – 170 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=466913

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

№ п/п	Ресурс
8	ЭБС Лань. – Режим доступа: по подписке. – URL: ЭБС Лань (lanbook.com)
9	ЭБС «Университетская библиотека онлайн». – Режим доступа: по подписке. – URL: ЭБС "Университетская библиотека онлайн" читать электронные книги (biblioclub.ru)
10	ЭБС ЮРАЙТ.– Режим доступа: по подписке. – URL: Образовательная платформа Юрайт. Для вузов и ссузов. (urait.ru)
11	Филологический портал www.philology.ru

16. Перечень учебно-методического обеспечения для самостоятельной работы

№ п/п	Источник
1	Хроленко, А.Т. Современные информационные технологии для гуманитария : [16+] / А.Т. Хроленко, А.В. Денисов. – 5-е изд., стер. – Москва : ФЛИНТА, 2018. – 129 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=363413
2	Норман, Б.Ю. Лингвистические задачи : учебное пособие / Б.Ю. Норман. – 5-е изд., стер. – Москва : ФЛИНТА, 2017. – 273 с. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=69155

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

При реализации дисциплины могут проводиться различные типы лекций (вводная, обзорная и т.д.). При проведении лабораторных работ предпочтение отдается применению классических технологий: обсуждение со студентами заранее подготовленных ими тем и разбор практических задач.

18. Материально-техническое обеспечение дисциплины:

/ауд. 12/ - компьютерный класс: Компьютер Arbyte Tempo/AOC (12 шт.), Проектор Benq MW523 (1 шт.), Сканер Canon Canoscan LiDE 120 (5 шт.) Экран проекционный (1 шт.)

---

19. Оценочные средства для проведения текущей и промежуточной аттестации

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

№ п/п	Наименование раздела дисциплины (модуля)	Компетенция	Индикаторы достижения компетенции	Оценочные средства
1	1. Корпусная лингвистика как прикладная дисциплина 2. Краткая история корпусной лингвистики 3. Основные корпуса русского языка 4. Основные корпуса английского и др. языков 5. Типология языковых корпусов 6. Принципы формирования корпусов 7. Метаразметка 8. Разметка текстов 9. Анализ корпусных данных	ПК-1	Собирает, обрабатывает, анализирует и обобщает результаты экспериментов и исследований в соответствующей области знаний (ПК-1.1)	Тесты № 1, 2 Реферат
		ПК-2	Представляет денотативную, концептуальную, коммуникативную и прагматическую информацию, содержащуюся в тексте на ЕЯ, формальными методами (ПК-2.1)	
		ПК-4	Составляет информационно-словарное описание языковых единиц (ПК-4.3)	

№ п/п	Наименование раздела дисциплины (модуля)	Компетенция	Индикаторы достижения компетенции	Оценочные средства
	10. Статистические методы в корпусных исследованиях 11. Параллельные корпуса в сопоставительных исследованиях 12. Корпусные данные в исследованиях неологизмов и заимствований 13. Корпуса в лексикографии 14. Грамматические исследования на базе корпусов 15. Изучение исторических изменений языка на базе корпусов 16. Использование корпусов в преподавании иностранных языков	ПК-14	Пользуется электронными языковыми ресурсами для решения прикладных задач (ПК-14.2)	
				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:

практические задания, в том числе домашние задания

тестовые задания

Тестовые задания

Тест № 1

1. Какие источники данных являются традиционными для лингвистики?

а) _____

_____

б)_____

в) _____

2. Языковой корпус – это

а) электронная коллекция текстов, которые отобраны и обработаны по определенным критериям;

б) электронная библиотека художественных текстов;

в) набор файлов с текстами разных жанров; г) всё перечисленное выше.

3. С появлением корпусов лингвисты получили возможность исследовать

а) языковую норму;

б) реальное употребление языковых единиц;

в) ошибки в речи носителей языка;

г) частотность грамматических конструкций.

4. Самый первый корпус содержал

а) 100 млн. словоупотреблений;

б) 1 млн. словоупотреблений;

в) 1 млн. текстов;

г) 1 млн. предложений.

5. Чтобы выводы, полученные на основе корпусного анализа, могли распространяться на использование языка в определенном языковом сообществе в конкретный период времени, корпус должен быть

а) размеченным;

б) репрезентативным;

с) однородным;

д) современным.

6. Метаразметка – это

а) грамматический анализ предложения;

б) информация о свойствах словоформ;

с) информация о свойствах текста;

д) информация о частях речи.

7. Морфологическая и синтаксическая разметка обеспечивает

а) возможность грамматического анализа предложения;

б) автоматический поиск грамматической информации;

с) поиск точных форм слов;

д) все перечисленное выше.

8. Лемма – это а) начальная форма слова;

б) одна из возможных словоформ лексемы;

с) информация о грамматических свойствах словоформы;

д) информация о частеречной принадлежности слова.

9. Символ * в поисковом запросе заменяет

а) один символ;

б) одну морфему;

с) любое количество символов в словоформе;

д) два символа.

10. Сбалансированность – это

а) равномерная представленность в корпусе текстов разных жанров;

б) наличие в корпусе текстов разных авторов;

с) наличие в корпусе текстов одинаковой длины;

д) наличие в корпусе параллельных текстов.

11. Благодаря корпусам лингвисты впервые смогли получать

а) качественную информацию о функционировании языка

б) количественную информацию о функционировании языка

в) данные об используемых грамматических конструкциях

г) данные о новых словах и выражениях

12. Какой из перечисленных ниже корпусов НЕ является корпусом русского языка?

а) НКРЯ

б) Тюбингенский корпус

в) ХАНКО

г) DWDS

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

а) лемме

б) словоформе

14. Сбалансированный и репрезентативный корпус может дать пользователю следующую информацию:

а) _____

б) _____

в) _____

15. В _____ корпусах представлено все жанровое / хронологическое разнообразие текстов. _____ корпусы включают в себя либо тексты определенных жанров, либо тексты, функционирующие в определенной сфере.

16. Имеет возможность постоянного пополнения

а) одноязычный корпус

б) специализированный корпус

в) открытый корпус

г) закрытый корпус

17. Назовите критерии, значимые для отбора текстов в корпус:

а) _____

б) _____

в) _____

г) _____

д) _____

е) _____

18. Назовите функции метаразметки

а) _____

б) _____

в) _____

19. Назовите три подвида метаразметки

а) _____

б) _____

в) _____

20. В стандартах TEI и EAGLES критерии метаразметки делятся на

а) _____

б)_____

Тест № 2

1. Разметка – это _____
_____
_____

2. Назовите основные виды разметки:

а)_____

б)_____

в)_____

г)_____

д)_____

е)_____

ж)_____

з)_____

3. Для каких целей необходима разметка?
_____
_____
_____
_____

4. Какие способы создания разметки существуют в корпусной лингвистике?

а)_____

б)_____

в)_____

5. Приписывание грамматических характеристик каждой словоформе – это _____ разметка.

6. Программы для создания морфологической разметки называются _____ _____

7. Для повышения качества работы программ, с помощью которых делается морфологическая разметка, используют не только лингвистические правила, но и _____.

8. Перечислите основные проблемы, с которыми сталкиваются при морфологической разметке:

а)_____

б)_____

в)_____

г)_____

д)_____

9. Синтаксическая разметка – это _____ _____ _____

10. Чаще всего для синтаксической разметки используются _____ и _____

11. В современных корпусах семантическая разметка – это _____ _____ _____

12. _____ частота показывает, насколько часто встречается в некотором заранее определенном объеме текстового материала.

13. Основная проблема использования статистических методов в корпусных исследованиях заключается в том, что они плохо применимы к _____

14. Коллокация – это _____ _____ _____

15. Индекс взаимной информации показывает _____ _____ _____

16. Напишите формулу, по которой вычисляется индекс взаимной информации:

17. Назовите другие методы статистического изучения корпусных данных:

а)_____

б)_____

в)_____

г)_____

18. Назовите сферы лингвистических исследований, в которых корпусные данные оказываются очень полезными:

а)_____

б)_____

в)_____

г)_____

д)_____

19. Коллигация – это _____

_____

20. Термином семантическая просодия обозначают _____

_____

Описание технологии проведения

Тест-задания выдаются студенту на электронном или бумажном носителе. Время выполнения теста – 25 мин. Каждое правильно выполненное задание оценивается в 1 балл. Максимально возможная сумма баллов за все правильно выполненные задания в тесте – 20 баллов.

Требования к выполнению заданий (или шкалы и критерии оценивания)

Выполнение теста оценивается по двухбалльной шкале: зачтено или не зачтено. Оценка «зачтено» ставится при правильном выполнении не менее 60 % заданий, что соответствует 12 баллам. Оценка «не зачтено» ставится в том случае, если студент набрал менее 12 баллов, т.е. выполнил менее 60 % заданий теста.

Примерный перечень тем рефератов

1. Национальный корпус русского языка: состав, структура, поисковые возможности (тема для 2 докладчиков).
2. Корпуса русского языка: Уппсальский корпус, Машинный фонд русского языка и др. История создания, возможности использования (тема для 2 докладчиков).
3. Национальный корпус чешского языка: состав, структура, поисковые возможности
4. Национальный корпус польского языка: состав, структура, поисковые возможности
5. The Corpus of Contemporary American English (тема для 2 докладчиков).
6. The Global Corpus of Web-Based English
7. Британский национальный корпус: история создания и возможности использования в лингвистических исследованиях (тема для 2 докладчиков).
8. The Michigan Corpus of Academic English как специализированный корпус
9. Корпус текстов журнала TIME
10. The Lancaster-Oslo-Bergen (LOB) Corpus: принципы формирования, объем, жанровая принадлежность текстов, поисковые возможности
11. The Brown Corpus: принципы формирования, объем, жанровая принадлежность текстов, поисковые возможности
12. Проект “Один речевой день” как корпус разговорного русского языка
13. Параллельный подкорпус Национального корпуса русского языка
14. Das Digitale Wörterbuch der deutschen Sprache
15. RLC: Russian Learner Corpus
16. The Hansard Corpus (British Parliament)
17. Проект «Рассказы о сновидениях»
18.ХАНКО – Хельсинский аннотированный корпус
19. The Corpus of American Soap Operas
20. Google Books как корпус
21. Метаразметка в Национальном корпусе русского языка
22. Проект TEI (Text Encoding Initiative)
23.Рекомендации EAGLES (Expert Advisory Group on Language Engineering Standards)
24. Стандарт CES (Corpus Encoding Standard)
25. Стандарт XCES (Corpus Encoding Standard for XML)
26. Проект ISLE (International Standard for Language Engineering)
27. Стандарт CDIF (Corpus Document Interchange Format)

28.Частеречная разметка (POS-tagging)
29. Синтаксическая разметка в НКРЯ
30. Семантическая разметка в НКРЯ

20.2 Промежуточная аттестация

Промежуточная аттестация по дисциплине осуществляется с помощью следующих оценочных средств: собеседование по экзаменационным билетам

Перечень вопросов к экзамену

1. Корпус как источник лингвистической информации. Виды корпусов.
2. Корпус и другие источники лингвистической информации: достоинства и недостатки.
3. История корпусной лингвистики.
4. Национальный корпус русского языка: состав, виды разметки и поисковые возможности.
5. Корпуса русского языка (Ханко, Уппсальский корпус, Тюбингенский корпус): состав, поисковые возможности, история разработки.
6. Корпуса английского языка (The Brown Corpus, LOB, BNC, COCA и др.).
7. Критерии отбора текстов в корпуса разных типов.
8. Сбалансированность и репрезентативность как основные требования к созданию корпуса.
9. Метаразметка: виды и функции.
10. Стандарты метаразметки (TEI и EAGLES)
11. Морфологическая разметка. Программы морфологической разметки.
12. Синтаксическая разметка. Программы синтаксической разметки.
13. Семантическая разметка (на примере НКРЯ и COCA).
14. Использование корпусов в лексико-семантических исследованиях.
15. Изучение коллокаций, коллострукций и семантической просодии с помощью корпусов.
16. Корпусные данные в исследованиях грамматики.
17. Использование корпусов в лексикографии.
18. Контрастивные исследования языков на основе корпусов.
19. Оценка частотности языковых явлений. Абсолютная и относительная частота. Индекс взаимной информации.
20. Параллельный корпус. Проблема выравнивания текстов. Программы выравнивания.
21. Программа составления конкордансов AntConc.
22. Лингвистическая теория и корпусные данные: corpus-based и corpus-driven варианты исследований.
23. Изучение заимствований на основе корпусных данных.
24. Изучение исторических изменений языка методами корпусной лингвистики. Исторические корпуса. Проблемы создания исторических корпусов.

Описание технологии проведения

Экзамен проводится по билетам, каждый из которых содержит 2 вопроса. На подготовку ответа отводится 30 минут. Правильный ответ на каждый вопрос в билете оценивается в 10 баллов. Максимальное количество набранных баллов – 20.

Требования к выполнению заданий, шкалы и критерии оценивания

Для оценивания результатов обучения на экзамене используются следующие показатели:

1. знание понятийного аппарата корпусной лингвистики;

2. способность иллюстрировать ответ примерами, фактами, данными научных исследований;

3. владение терминологическим аппаратом изучаемой дисциплины; программными средствами для проведения анализа корпусных данных.

Для оценивания результатов обучения на экзамене используется 4-х балльная шкала: «Отлично», «Хорошо», «Удовлетворительно», «Неудовлетворительно».

1. Оценка «Отлично» ставится в случае, если студент набрал 18-20 баллов.
2. Оценка «Хорошо» ставится в случае, если студент набрал 15-17 баллов.
3. Оценка «Удовлетворительно» ставится в случае, если студент набрал 12-14 баллов.
4. Оценка «Неудовлетворительно» ставится в случае, если студент набрал менее 12 баллов.

Соотношение показателей, критериев и шкалы оценивания результатов обучения.

Критерии оценивания компетенций	Уровень сформированности компетенций	Шкала оценок
Полное соответствие ответа обучающегося всем перечисленным критериям. Продемонстрировано знание понятийного аппарата корпусной лингвистики, способность иллюстрировать ответ примерами, фактами, данными научных исследований, владение терминологическим аппаратом изучаемой дисциплины, программными средствами для проведения анализа корпусных данных	Повышенный уровень	Отлично
Ответ на контрольно-измерительный материал не соответствует одному (двум) из перечисленных показателей, но обучающийся дает правильные ответы на дополнительные вопросы. Недостаточно продемонстрировано знание базовых понятий корпусной лингвистики, способность иллюстрировать ответ примерами, фактами, данными научных исследований, владение терминологическим аппаратом изучаемой дисциплины, программными средствами для проведения анализа корпусных данных	Базовый уровень	Хорошо
Ответ на контрольно-измерительный материал не соответствует любым двум (трем) из перечисленных показателей, обучающийся дает неполные ответы на дополнительные вопросы. Демонстрирует частичные знание базовых понятий корпусной лингвистики, способность иллюстрировать ответ примерами, фактами, данными научных исследований, владение терминологическим аппаратом изучаемой дисциплины, программными средствами для проведения анализа корпусных данных	Пороговый уровень	Удовлетворительно
Ответ на контрольно-измерительный материал не соответствует любым трем (четырем) из перечисленных показателей. Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки при	–	Неудовлетворительно

практическом применении приобретенных знаний; не может использовать программные средства для проведения анализа корпусных данных.		

20.3 Материалы для диагностической работы

Закрытые вопросы

1. Какой критерий будет использоваться в качестве основного при отборе текстов для исторического корпуса?

а) жанр текста
б) модус текста
в) время создания текста
г) место создания текста

2. Что является основным недостатком языковой интуиции лингвиста как источника данных?
а) она не успевает за языковыми изменениями
б) она не позволяет видеть вариативность в использовании языка
в) сконструированные лингвистом высказывания не соответствуют языковой норме

3. Языковой корпус – это
а) электронная коллекция текстов, отобранных и обработанных по определенным критериям
б) электронная библиотека художественных текстов
в) набор файлов с текстами разных жанров
11. С появлением корпусов лингвисты получили возможность исследовать
а) языковую норму;
б) реальное употребление языковых единиц;
в) ошибки в речи носителей языка

4. Лемма – это
а) начальная форма слова;
б) одна из возможных словоформ лексемы;
в) информация о грамматических свойствах словоформы;
г) информация о частеречной принадлежности слова.

5. Символ * в поисковом запросе заменяет
а) один символ;
б) одну морфему;
в) любое количество символов в словоформе;
г) два символа.

6. Из приведенных ниже корпусов закрытым является
а) Национальный корпус русского языка
б) The Corpus of Contemporary American English
в) The British National Corpus
г) Генеральный интернет-корпус русского языка

7. Имеет возможность постоянного пополнения
а) одноязычный корпус
б) специализированный корпус
в) открытый корпус
г) закрытый корпус

8. Внешняя метаразметка описывает жанр, объем, время создания текста, а также включают в себя характеристики автора.

Ответ: верно.

9. Внутренняя метаразметка включает информацию о структуре текста (деление на главы, абзацы, предложения, словоформы).

Ответ: верно.

10. Языковая интуиция лингвиста не является надежным источником данных, поскольку она не позволяет увидеть вариативность в использовании языка.

Ответ: верно.

11. Отбор текстов в корпус осуществляется на основе четких эксплицитных критериев.

Ответ: верно.

12. Тенденция слова встречаться в определенном синтаксическом окружении называется
а) коллокация
б) коллигация
в) семантическая просодия

13. Английский глагол *to happen* часто сочетается с существительными *accident, disaster, incident, tragedy, crash, explosion.* Также в его ближайшем окружении встречаются прилагательные *terrible, strange, horrible, awful.* Как можно охарактеризовать его семантическую просодию?
а) она является негативной
б) она является положительной
в) она является нейтральной

14. Какой вид метаразметки содержит информацию о структуре текста и дает возможность работать с частью текста?
а) внешняя метаразметка
б) внутренняя метаразметка
в) технико-технологическая метаразметка

15. Принцип, в соответствии с которым любое слово может быть употреблено практически в любом контексте, называется
а) принципом открытого выбора
б) принципом идиоматичности
в) принципом семантической просодии

Открытые вопросы

1. Лексикографы анализируют употребление слов, используя _____.

Ответ: конкорданс / конкордансы

2. _____ позволяет пользователю задавать подкорпус текстов с определенными свойствами и формулировать запросы применительно к отобранным текстам.

Ответ: Метаразметка / метаразметка

3. Дополнительная информация о лингвистических свойствах словоформ, которая приписывается в виде тэгов, называется _____.

Ответ: разметка / разметкой

4. Приведение словоформ к начальной форме называется _____.

Ответ: лемматизация / лемматизацией

5. Корпуса, в которых представлено все жанровое и хронологическое разнообразие текстов на определенном языке, называются _____.

Ответ: национальный / национальный корпус / национальными / национальными корпусами

6. Если корпус включает в себя только тексты определенного жанра или тексты, функционирующие в определенной сфере, то такой корпус называется _____.

Ответ: специализированный / специализированный корпус / специализированным / специализированным корпусом

7. Применение _____ методов позволяет получить количественные данные о функционировании языка.

Ответ: статистических / статистические

8. Приведение словоформ к их начальной форме называется _____.

Ответ: лемматизация / лемматизацией

9. Если в языковом корпусе равномерно представлены все типы и жанры письменных и устных текстов, представленные в данном языке, то такой корпус является _____.

Ответ: сбалансированным

10. Для сопоставительных исследований языков используются корпуса, содержащие предложения одного языка и их переводы на другой язык или другие языки. Такие корпуса называются _____.

Ответ: параллельные / параллельными / параллельный

11. Корпус, содержащий тексты, относящиеся к разным хронологическим периодам, позволяет получать информацию об _____ изменениях в употреблении языковых единиц различных уровней.

Ответ: исторических

Практикоориентированные вопросы (эссе)

1. Перечислите критерии, которые должны учитываться при отборе текстов для национального языкового корпуса.

(Ответ: автор, время, место, язык, жанр, тип текста)

2. Для каких целей могут использоваться учебные корпуса? Приведите два примера практического применения учебных языковых корпусов.

(Ответ: для выявления наиболее типичных ошибок, для обучения иностранному языку)

3. Перечислите источники языковых данных, которыми пользовались лингвисты до появления корпусов.

(Ответ: традиционные письменные тексты, словари и грамматики, интуиция лингвиста, общение с носителями языка)

4. Какие виды разметки существуют в современных корпусах?

(Ответ: морфологическая, синтаксическая, семантическая, акцентологическая и др.)

5. Как языковой корпус может помочь в исследованиях заимствований и неологизмов?

(Ответ: можно установить время появления слова в языке, варианты написания, лексическую сочетаемость, проследить развитие новых значений у слова и др.)