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

УТВЕРЖДАЮ

Заведующий кафедрой
теоретической и прикладной лингвистики

проф. А.А. Кретов
02.07.2018 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
Б1.Б.26 Технологии корпусной лингвистики

1. Код и наименование направления подготовки/специальности:

45.03.03 Фундаментальная и прикладная лингвистика

2. Профиль подготовки/специализации: -

3. Квалификация (степень) выпускника: бакалавр

4. Форма образования: 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 № 10 от 19.06.2018 г.

8. Учебный год: 2018/2019 **Семестр(-ы):** 6-й

9.**Цели и задачи учебной дисциплины:** основной целью дисциплины является формирование у студентов умений и 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 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.Б.26 – Технологии корпусной лингвистики входит в базовый цикл и является обязательной для освоения в рамках направления «Фундаментальная и прикладная лингвистика». Изучение данной дисциплины предшествует освоению следующих дисциплин: Б1.Б.28 – Технологии обработки текста и звучащей речи, Б1.В.ОД.1 – Основные проблемы современной лингвистики.

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

Компетенция		Планируемые результаты обучения
Код	Название	
ПК-3	Владение методами сбора и документации лингвистических данных	знать: методики поиска, анализа и обработки материала исследования уметь: работать с различными источниками информации владеть (иметь навык(и)): навыками реферирования, формулирования целей, задач, методов, выводов научного исследования

12. Объем дисциплины в зачетных единицах/час.(*в соответствии с учебным планом)* **—** 3 ЗЕТ/ 108 часов.

Форма промежуточной аттестации: экзамен, курсовая работа

13. Виды учебной работы

Вид учебной работы	Трудоемкость	
	Всего	По семестрам
		6 семестр
Аудиторные занятия	48	48
в том числе: лекции	30	30
практические	16	16
лабораторные	0	0
Самостоятельная работа	26	26
Экзамен, курсовая работа	36	36
Итого:	108	108

13.1. Содержание дисциплины

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины
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	Краткая история корпусной лингвистики	Брауновский корпус. LOB. Поисково-информационные возможности первых корпусов.
1.3	Основные корпуса русского языка	Уппсальский корпус. Тюбингенский корпус. ХАНКО. Национальный корпус русского языка. RuSKELL
1.4	Основные корпуса английского и др. языков	Британский национальный корпус. Корпуса М. Дэвиса. «Срезовые» (snapshot) корпуса.
1.5	Типология языковых корпусов.	Типы корпусов. Состав и структура корпусов различных типов.
1.6	Принципы формирования корпусов	Сбалансированность. Репрезентативность. Критерии отбора текстов в корпуса разных типов.
1.7	Метаразметка	Метаразметка как элемент корпуса. Стандарты метаразметки.
1.8	Разметка текстов	Виды разметки. Морфологическая разметка. Синтаксическая разметка. Семантическая разметка. Международные стандарты разметки.
1.9	Анализ корпусных данных	Конкордансеры. AntConc. Sketch Engine. Информативные возможности конкордансеров.
1.10	Статистические методы в корпусных исследованиях	Частотность употребления языковых единиц. Абсолютная и относительная частота. Индекс взаимной информации. Другие статистические методы исследования корпусных данных
1.11	Параллельные корпуса в сопоставительных исследованиях	Параллельный корпус как источник данных для межъязыковых сопоставлений. Лексические и грамматические данные.
1.12	Корпусные данные в исследованиях неологизмов и заимствований	Орфографическая и грамматическая вариативность новой лексики. Мониторинг использования новых слов в текстах разных жанров.
1.13	Корпуса в лексикографии	Возможности использования корпусов в описании слов. Выделение значений на основе корпусных данных.
1.14	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.
1.15	Изучение исторических изменений языка на базе корпусов	Исторический корпус. Примеры исторических корпусов. Хронологический анализ языковых данных.
1.16	Использование корпусов в преподавании иностранных языков	Учебные корпуса. Анализ лексической сочетаемости, грамматических свойств слов изучаемого языка как способ активного постижения ИЯ.
2. Практические занятия		
2.1	Основные корпуса русского языка	Национальный корпус русского языка. RuSKELL. Типы запросов.
2.2	Основные корпуса английского и др. языков	Британский национальный корпус. Корпуса М. Дэвиса. Типы запросов.
2.3	Типология языковых корпусов.	Типы корпусов. Состав и структура корпусов различных типов.
2.4	Метаразметка	Метаразметка как элемент корпуса. Стандарты метаразметки.

2.5	Разметка текстов	Виды разметки. Морфологическая разметка. Синтаксическая разметка. Семантическая разметка. Международные стандарты разметки.
2.6	Анализ корпусных данных	Конкордансеры. AntConc. Sketch Engine. Информативные возможности конкордансеров.
2.7	Статистические методы в корпусных исследованиях	Частотность употребления языковых единиц. Абсолютная и относительная частота. Индекс взаимной информации. Другие статистические методы исследования корпусных данных
2.8	Параллельные корпуса в сопоставительных исследованиях	Параллельный корпус как источник данных для межъязыковых сопоставлений. Лексические и грамматические данные.

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела) дисциплины	Виды занятий (часов)			
		Лекции	Практические	Самостоятельная работа	Всего
1.1	Корпусная лингвистика как прикладная дисциплина	2		1	3
1.2	Краткая история корпусной лингвистики	2		1	3
1.3	Основные корпуса русского языка	2	2	2	6
1.4	Основные корпуса английского и др. языков	2	2	2	6
1.5	Типология языковых корпусов.	2	2	2	6
1.6	Принципы формирования корпусов	2		1	3
1.7	Метаразметка	2	2	2	6
1.8	Разметка текстов	2	2	2	6
1.9	Анализ корпусных данных	2	2	2	6
1.10	Статистические методы в корпусных исследованиях	2	2	2	6
1.11	Параллельные корпуса в сопоставительных исследованиях	2	2	2	5
1.12	Корпусные данные в исследованиях неологизмов и 2заимствований	2		1	3
1.13	Корпуса в лексикографии	2		1	3
1.14	Грамматические исследования на базе корпусов	2		1	3
1.15	Изучение исторических изменений языка на базе корпусов	2		1	3
1.16	Использование корпусов в преподавании иностранных языков	2		1	3
	Итого:	32	18	24	108 (36 ч. – экзамен)

14. Методические указания для обучающихся по освоению дисциплины

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

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

а) основная литература:

№ п/п	Источник
1.	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
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. : ил. - Библиогр.: с. 480-513. URL: http://biblioclub.ru/index.php?page=book&id=473302

б) дополнительная литература:

№ п/п	Источник
3.	Гальперин И. Р. Текст как объект лингвистического исследования / И. Р. Гальперин. – Москва : УРСС, 2005. – 144 с.
4.	Тураева З. Я. Лингвистика текста / З. Я. Тураева. – Москва: Либроком, 2009. – 144 с.
5.	Филиппов К. А. Лингвистика текста : курс лекций / К. А. Филиппов. – Санкт-Петербург: Изд-во СПбГУ, 2005. – 336 с.

в) базы данных, информационно-справочные и поисковые системы**:**

№ п/п	Источник
27.	ЭБС «Университетская библиотека online. URL: https://biblioclub.lib.vsu.ru

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

№ п/п	Источник

17. Информационные технологии, используемые для реализации учебной дисциплины, включая программное обеспечение и информационно-справочные системы (при необходимости)

Программы UAM CorpusTool, Sketch Engine, CLAWS (online-версия), AntConc

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.)	г.Воронеж, пл.Ленина 10, ауд.12

19. Фонд оценочных средств:

19.1. Перечень компетенций с указанием этапов формирования и планируемых результатов обучения

Код и	Планируемые результаты обучения	Этапы	

содержание компетенции (или ее части)	(показатели достижения заданного уровня освоения компетенции посредством формирования знаний, умений, навыков)	формирования компетенции (разделы (темы) дисциплины или модуля и их наименование)	ФОС* (средства оценивания)
ПК-3	знать: основные принципы документации лингвистических данных, стандарты метаразметки и аннотации	Разделы 1.1-1.8	Устный опрос, реферат, тест 1
	уметь: отбирать и систематизировать данные для корпусов различных типов:	Разделы 1.1-1.8	Устный опрос, реферат, тест 1
	владеть: навыками создания лингвистических корпусов различных типов	Разделы 1.1-1.8	Устный опрос, реферат, тест 1
Промежуточная аттестация			КИМ

19.2 Описание критериев и шкалы оценивания компетенций (результатов обучения) при промежуточной аттестации

Для оценивания результатов обучения на экзамене используются следующие показатели:

знать: основные принципы документации лингвистических данных, стандарты метаразметки и аннотации

уметь: отбирать и систематизировать данные для корпусов различных типов:

владеть: навыками создания лингвистических корпусов различных типов

Для оценивания результатов обучения на экзамене используется 4-балльная шала: «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Соотношение показателей, критериев и шкалы оценивания результатов обучения.

Критерии оценивания компетенций	Уровень сформированности компетенций	Шкала оценок
Обучающийся в полной мере владеет понятийным аппаратом корпусной лингвистики, способен иллюстрировать ответ примерами, фактами, данными научных исследований, использовать программные средства для создания анализа корпусных данных.	*Повышенный уровень*	*Отлично*
Обучающийся владеет понятийным аппаратом корпусной лингвистики, способен иллюстрировать ответ примерами, фактами, данными научных исследований, использовать программные средства для создания анализа корпусных данных, однако допускает ошибки в использовании терминологии и/или анализе конкретных языковых явлений различных уровней.	*Базовый уровень*	*Хорошо*
Обучающийся владеет частично теоретическими основами дисциплины, фрагментарно способен иллюстрировать ответ примерами, фактами, данными научных исследований, применять использовать программные средства для создания анализа корпусных данных, допускает ошибки в использовании терминологии и/или анализе конкретных дискурсивных явлений различных уровней.	*Пороговый уровень*	*Удовлетвори-тельно*
Ответ на контрольно-измерительный материал не соответствует любым четырем из перечисленных	–	*Неудовлетвори-тельно*

показателей. Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки в анализе корпусных данных.		

19.3 Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующие этапы формирования компетенций в процессе освоения образовательной программы

19.3.1 Перечень вопросов к экзамену (зачету): (нужное выбрать)

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.

19.3.2 Перечень практических заданий

19.3.4 Тестовые задания
Комплект заданий для теста № 1

1. Какие источники данных являются традиционными для лингвистики?

а) _____
б) _____
в) _____
г) _____

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 стандантах TEI и EAGLES критерии метаразметки делятся на
а) _____
б)_____

Комплект заданий для теста № 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 _____
_____

19.3.4 Перечень заданий для контрольных работ

19.3.5 Темы курсовых работ

19.3.6 Темы 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

19.4. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций

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. 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.

При оценивании используются качественные шкалы оценок. Критерии оценивания приведены выше.