МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Российский государственный гуманитарный университет»
(ФГБОУ ВО «РГГУ»)

ИНСТИТУТ ЛИНГВИСТИКИ

Учебно-научный центр лингвистической типологии

КОРПУСНАЯ ЛИНГВИСТИКА

Рабочая программа дисциплины

Направление 45.03.02 Лингвистика
Профиль Язык и коммуникация

Уровень высшего образования: бакалавриат

Форма обучения очная

РПД адаптирована для лиц
с ограниченными возможностями
здоровья и инвалидов

Москва 2023

Корпусная лингвистика

Рабочая программа дисциплины

Составитель:
к. филол. н., доцент УНЦ компьютерной лингвистики А.Ч. Пиперски

Ответственный редактор:
д. филол. н., проф. Я.Г. Тестелец

УТВЕРЖДЕНО
Протокол заседания УНЦ КЛ № 5 от 31.03.22

УТВЕРЖДЕНО
Протокол заседания кафедры ТиПЛ
№ 10 от 02.04.2024

ОГЛАВЛЕНИЕ

1. Пояснительная записка

1.1. Цели и задачи курса

Предмет курса – современные представления лингвистики о компьютерных корпусах текстов и основные методы корпусного анализа.

Цель курса – освоение студентами базовых понятий и методов лингвистически ориентированного корпусного анализа, классификации корпусов, принципов описания и составления корпусов и вопросов соотнесения знаний, полученных с применением лингвистических корпусов, с лингвистическими знаниями, полученными иным способом.

Задачи курса:

Курс нацелен на **формирование** у студентов следующих профессиональных **компетенций**:

- владение основными понятиями и категориями современной лингвистики
- владение основными методами фонологического, морфологического, синтаксического, дискурсивного и семантического анализа с учетом корпусных данных
- владение основными способами описания и формальной репрезентации денотативной, концептуальной, коммуникативной и прагматической информации, содержащейся в корпусах текстов на естественном языке
- способность определять тип корпуса с учетом специфики входящих в него жанров и функционально-стилевых разновидностей

1.2. Перечень планируемых результатов обучения по дисциплине, соотнесенных с индикаторами достижения компетенций

Компетенция (код и наименование)	**Индикаторы компетенций** (код и наименование)	**Результаты обучения**
УК-4 Способен осуществлять деловую коммуникацию в устной и письменной формах на государственном зыке Российской Федерации и иностранном(ых) языке(ах)	4.3	Использует информационно-коммуникационные технологии при поиске необходимой информации в процессе решения стандартных коммуникативных задач для достижения профессиональных целей на государственном м иностранном (ых) языке (ах)
ПК-3 Способен к научно-исследовательской деятельности	3.1	Владеет основами методов научного исследования, информационной и библиографической культурой
	3.2	Владеет стандартными методиками поиска, анализа и обработки материала исследования

	3.3	Умеет логично и последовательно представить результаты своего исследования

1.3. Место дисциплины в структуре образовательной программы

Дисциплина относится к части, формируемой участниками образовательного процесса, блока Б1.В дисциплин учебного плана, читается в 7 семестре преподавателями УНЦ компьютерной лингвистики Института лингвистики.
Для освоения дисциплины необходимы знания, умения и владения, сформированные в ходе изучения следующих дисциплин и прохождения практик: «Практическая семантика и лексикография», «Технологии искусственного интеллекта в гуманитарных сферах».
В результате освоения дисциплины формируются знания, умения и владения, необходимые для изучения следующих дисциплин и прохождения практик: преддипломная практика, ГИА.

2. Структура дисциплины

Общая трудоёмкость дисциплины составляет 3 з.е., 108 ч., в том числе контактная работа обучающихся с преподавателем 42 ч., промежуточная аттестация 18 ч., самостоятельная работа обучающихся 48 ч.

Тематический календарный план курса

№ № раздела	Раздел курса	Семестр: недели	Виды учебной работы и трудоемкость (в часах)			Формы контроля успеваемости
			лекции	практ.з.	СРС	
1.	*Введение. Понятие лингвистического корпуса. Сбалансированность и репрезентативность.*	*2: 1*	*2*	*2*	*4*	*Контроль посещаемости студентов*
2.	*Основные корпуса русского и английского языков.*	*2: 2-4*	*4*	*4*	*8*	*Контроль посещаемости студентов. Мини-тесты по пройденным темам. Обсуждение прочитанной научной литературы и практический анализ корпусных данных*
3.	*Принципы разметки корпусов.*	*2: 5*	*2*	*2*	*6*	*Контроль посещаемости студентов. Мини-тесты по пройденным темам. Обсуждение прочитанной научной литературы и практический анализ корпусных данных*
4.	*Подсчёт частотности языковых единиц по корпусам. Частотные словари.*	*2: 6-9*	*2*	*4*	*6*	*Контроль посещаемости студентов. Мини-тесты по пройденным темам. Обсуждение прочитанной научной литературы и практический анализ корпусных данных*
	Промежуточная контрольная работа и ее обсуждение	*2: 10-11*	*2*	*2*	*6*	*Письменная контрольная работа по разделам 1-4*
5.	*Извлечение ключевых слов и устойчивых сочетаний из корпусов. Методы и оценки.*	*2: 12-13*	*2*	*4*	*6*	*Контроль посещаемости студентов. Мини-тесты по пройденным темам. Обсуждение прочитанной научной литературы и практический анализ корпусных данных*
6.	*Автоматическое сравнение корпусов.*	*2: 14-15*	*2*	*4*	*6*	*Контроль посещаемости студентов. Мини-тесты по пройденным темам. Обсуждение прочитанной научной литературы и практический анализ корпусных*

№ № раздела	Раздел курса	Семестр: недели	Виды учебной работы и трудоемкость (в часах)			Формы контроля успеваемости
			лекции	практ.з.	СРС	
						данных
7.	*Обсуждение индивидуальных проектов*	*2: 16-18*	*2*	*2*	*6*	*Представление промежуточных результатов индивидуальных проектов. Участие в коллективном обсуждении*
	Экзамен				*18*	*Защита индивидуальных исследовательских проектов. Ответы на контрольные вопросы по курсу*
	Итого:		***20***	***22***	***66***	

3. Содержание дисциплины

Раздел I. Введение. Понятие лингвистического корпуса. Сбалансированность и репрезентативность.

Корпус как лингвистический объект. Ключевые характеристики корпуса как объекта изучения: существование в реальном мире, потенциально неограниченный объем, разнообразие типов. Сбалансированность и репрезентативность корпуса.

Раздел II. Основные корпуса русского и английского языков

Корпуса русского языка (обзор): 1. Национальный корпус русского языка (НКРЯ); 2. ruWac; 3. ruTenTen; 4. Хельсинкский аннотированный корпус (ХАНКО); 5. Интегрум; 6. Открытый корпус (OpenCorpora); 7. Генеральный Интернет-корпус русского языка (ГИКРЯ). Корпуса английского языка (обзор): 1. British National Corpus (BNC); 2. Corpus of Contemporary American English (COCA); 3. Corpus of Global Web-Based English (GloWbe); 4. Brown Corpus; 5. Google Books: Google Ngrams Viewer и поисковый интерфейс на сайте Brigham Young University.

Раздел III. Принципы разметки корпусов.

Уровни разметки лингвистического корпуса. Метязыковая и собственно лингвистическая разметка. Токенизация. Морфологическая, синтаксическая и семантическая разметки. Трудности при автоматической и ручной разметке. Оценка качества автоматической разметки корпусов.

Раздел IV. Подсчёт частотности языковых единиц по корпусам. Частотные словари.

Абсолютная и относительная частотность. ipm (instances per million) как мера относительной частотности: $f_{ipm}(w) = \frac{c(w)}{N}$*, где* $f_{ipm}(w)$ *— относительная частотность слова,* $c(w)$ *— абсолютное количество вхождений слова w, N — общий объём корпуса в словах*

Закон Ципфа: $f(w) = \frac{C}{r^a}$*, где* $f(w)$ *— частотность слова (абсолютная или относительная), C — константа, a — параметр, определяющий скорость убывания частот.*

Смещённость частот при разбиении корпуса на части: смещение в сторону выше среднего арифметческого — обычно в небольшом количестве подкорпусов. Медиана, мода и среднее арифметическое как меры центральной тенденции для частотности по подкорпусам.

Раздел V. Извлечение ключевых слов и устойчивых сочетаний из корпусов. Методы и оценки

Основные методы и подходы к извлечению ключевых слов. Меры ключёвости: хи-квадрат, логарифмическое правдоподобие, %DIFF Килгаррифа. Извлечение ключевых слов с помощью существующих электронных ресурсов (SketchEngine). Подходы к извлечению ключевых неоднословных выражений.

Раздел VI. Автоматическое сравнение корпусов

Частотные списки как основа для сравнения корпусов. Мера tf-idf. Меры измерения расстояния между частотными списками. Способы оценки сравнения корпусов (корпуса известной степени сходства).

4. Образовательные технологии

Дисциплина «Корпусная лингвистика» реализуется преимущественно интерактивно – в форме интерактивных лекций, семинарских занятий и в различных видах коллективной и самостоятельной работы студента.

Наименование раздела	Виды учебной работы	Информационные и образовательные технологии
Раздел I. Введение. Понятие лингвистического корпуса. Сбалансированность и ре-презентативность.		
Раздел II. Основные корпуса русского и английского языков	**Семинар 1.** 1. Национальный корпус русского языка (НКРЯ); 2. ruWac; 3. ruTenTen; 4. Хельсинкский аннотированный корпус (ХАНКО); 5. Интегрум.	Анализ корпусов с выявлением их особенностей
	Семинар 2. 6. Открытый корпус (OpenCorpora); 7. Генеральный Интернет-корпус русского языка (ГИК-РЯ). Корпуса английского языка (обзор).	Анализ корпусов с выявлением их особенностей
	Семинар 3. 1. British National Corpus (BNC); 2. Corpus of Contemporary American English (COCA); 3. Cor-pus of Global Web-Based English (GloWbe); 4. Brown Corpus; 5. Google Books: Google Ngrams Viewer и поисковый интерфейс на сайте Brigham Young University.	Анализ корпусов с выявлением их особенностей
Раздел III. Принципы разметки корпусов	**Семинар 4.** Токенизация. Морфологическая, синтаксическая и семантическая разметки. Трудности при автоматической и ручной разметке. Оценка качества автоматической разметки корпусов.	Практическая работа с корпусами разноструктурных языков.
Раздел IV. Подсчёт частотности языковых единиц по корпусам. Частотные словари	**Семинар 5.** Абсолютная и относительная частотность. ipm (instances per million) как мера относительной частотности.	Практическая работа с корпусами разноструктурных языков.
	Семинар 6. Закон Ципфа	Доклад по прочитанной литературе. Коллективное обсуждение

Наименование раздела	Виды учебной работы	Информационные и образовательные технологии
		разбираемой темы
	Семинар 7. Смещённость частот при разбиении корпуса на части.	Практическая работа с корпусами разноструктурных языков
	Семинар 8. Медиана, мода и среднее арифметическое как меры центральной тенденции для частотности по подкорпусам.	Самостоятельный анализ корпусных данных с последующим разбором на семинаре
Контрольная работа	**Семинары 9-10.** Письменная контрольная работа	Самостоятельный анализ материала
	Семинар 11. Обсуждение контрольной работы	Коллективное обсуждение разбираемой темы
Раздел V. Извлечение ключевых слов и устойчивых сочетаний из корпусов. Методы и оценки	**Семинар 12.** Основные методы и подходы к извлечению ключевых слов. Меры ключёвости: хи-квадрат, логарифмическое правдоподобие, %DIFF Килгарриффа.	Коллективное обсуждение разбираемой темы
	Семинары 13-14. Извлечение ключевых слов с помощью существующих электронных ресурсов (SketchEngine). Подходы к извлечению ключевых неоднословных выражений.	Доклад по прочитанной литературе. Коллективное обсуждение разбираемой темы
Раздел VI. Автоматическое сравнение корпусов	**Семинар 15.** Частотные списки как основа для сравнения корпусов. Мера tf-idf.	Коллективное обсуждение разбираемой темы
	Семинар 16. Меры измерения расстояния между частотными списками. Способы оценки сравнения корпусов (корпуса известной степени сходства).	Доклад по прочитанной литературе. Коллективное обсуждение разбираемой темы
Обсуждение индивидуальных проектов	**Семинары 17-19.** Обсуждение индивидуальных курсовых проектов	Индивидуальные презентации с дальнейшим обсуждением текущих результатов исследований
Промежуточная аттестация	**Экзамен**	Представление итоговых результатов индивидуальных исследований. Ответы на контрольные вопросы по курсу

5. Оценка планируемых результатов обучения

5.1. Система оценивания

Контролируемые разделы дисциплины (модуля)	**Оценочные средства**
Раздел II. Основные корпуса русского и английского языков	Мини-тесты по пройденным темам. Проверка и выполнение практических заданий дома и на занятиях, сообщения о прочитанной литературе
Раздел III. Принципы разметки корпусов	Мини-тесты по пройденным темам. Проверка и выполнение практических заданий дома и на занятиях, сообщения о прочитанной литературе
Раздел IV. Подсчёт частотности языковых единиц по корпусам. Частотные словари	Мини-тесты по пройденным темам. Проверка и выполнение практических заданий дома и на занятиях, сообщения о прочитанной литературе. Промежуточная контрольная по разделам II-IV.
Раздел V. Извлечение ключевых слов и устойчивых сочетаний из корпусов. Методы и оценки	Мини-тесты по пройденным темам. Проверка и выполнение практических заданий дома и на занятиях, сообщения о прочитанной литературе.
Раздел VI. Автоматическое сравнение корпусов	Мини-тесты по пройденным темам. Проверка и выполнение практических заданий дома и на занятиях, сообщения о прочитанной литературе.

Оценка знаний студента производится по 100-балльной шкале и учитывает результаты текущего контроля успеваемости (до 60 баллов) и результаты промежуточной аттестации (до 40 баллов).

Оценка «удовлетворительно» выставляется, если студент набрал в сумме не менее 50 баллов. При выставлении оценки в ведомость и в зачетную книжку преподаватель должен указать результат в соответствии с традиционной шкалой оценок и со шкалой оценок Европейской системы переноса и накопления кредитов (European Credit Transfer System; далее – ECTS) в соответствии с таблицей:

100-балльная шкала	**Традиционная шкала**		**Шкала ECTS**
95 – 100	отлично	зачтено	A
83 – 94			B
68 – 82	хорошо		C
56 – 67	удовлетворительно		D
50 – 55			E
20 – 49	неудовлетворительно	не зачтено	FX
0 – 19			F

Распределение баллов по видам учебной деятельности таково:

1) уровень активности студента на семинарах (выполнение домашних заданий и участие в их обсуждении; выступления по прочитанной литературе; участие в об-

суждении и выполнении коллективных заданий; представление промежуточных результатов индивидуальных исследований и участие в их обсуждении) — всего до 20 баллов;
2) написание мини-тестов по темам предшествующих занятий — до 15 баллов;
3) написание промежуточной письменной контрольной — до 25 баллов;
4) защита индивидуального исследовательского проекта — до 30 баллов;
5) ответы на контрольные вопросы по курсу — до 10 баллов

Если студент не набрал 50 баллов, он проходит пересдачу в форме устного тестирования по всей программе курса с обязательным выполнением практических заданий.

Баллы за участие в семинарах. Преподавание дисциплины строится на параллельном обсуждении теоретических вопросов и выполнении индивидуальных и коллективных практических заданий. Все эти задания в обязательном порядке обсуждаются на семинарах. Активное и разумное участие в обсуждении этих заданий, а также доклады о прочитанной литературе и обсуждение индивидуальных исследовательских проектов может принести студенту до 20 баллов.

Баллы за выполнение текущих контрольных работ. В начале большинства семинаров проводятся мини-тесты, включающие в себя базовые вопросы по темам предшествующих занятий. Суммарно за семестр выполнение этих тестов может принести студенту до 15 баллов. По завершении раздела IV проводится промежуточная письменная контрольная, предполагающая выполнение практических заданий с опорой на теоретические положения курса (до 25 баллов).

Баллы за промежуточную аттестацию. Экзамен состоит из двух частей. В первой части студенты представляют (в форме устного доклада, подкрепленного визуальной презентацией) результаты индивидуальных исследований по выбранным ими темам. Оценивается степень овладения тематикой, корректность выбранного метода исследования, наличие содержательных результатов и умение их представить аудитории (в сумме — до 30 баллов). Во второй части студенты устно отвечают на контрольные вопросы по курсу (до 10 баллов).

5.2.Критерии выставления оценки по дисциплине

Баллы/ Шкала ECTS	**Оценка по дисциплине**	**Критерии оценки результатов обучения по дисциплине**
100-83/ A,B	«отлично»/ «зачтено (отлично)»/ «зачтено»	Выставляется обучающемуся, если он глубоко и прочно усвоил теоретический и практический материал, может продемонстрировать это на занятиях и в ходе промежуточной аттестации. Обучающийся исчерпывающе и логически стройно излагает учебный материал, умеет увязывать теорию с практикой, справляется с решением задач профессиональной направленности высокого уровня сложности, правильно обосновывает

Баллы/ Шкала ECTS	Оценка по дисциплине	Критерии оценки результатов обучения по дисциплине
		принятые решения. Свободно ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «высокий».
82-68/ C	«хорошо»/ «зачтено (хорошо)»/ «зачтено»	Выставляется обучающемуся, если он знает теоретический и практический материал, грамотно и по существу излагает его на занятиях и в ходе промежуточной аттестации, не допуская существенных неточностей. Обучающийся правильно применяет теоретические положения при решении практических задач профессиональной направленности разного уровня сложности, владеет необходимыми для этого навыками и приёмами. Достаточно хорошо ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «хороший».
67-50/ D,E	«удовлетвори-тельно»/ «зачтено (удовлетворительно)»/ «зачтено»	Выставляется обучающемуся, если он знает на базовом уровне теоретический и практический материал, допускает отдельные ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает определённые затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, владеет необходимыми для этого базовыми навыками и приёмами. Демонстрирует достаточный уровень знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «достаточный».
49-0/ F,FX	«неудовлетворительно»/ не зачтено	Выставляется обучающемуся, если он не знает на базовом уровне теоретический и практический материал, допускает грубые ошибки при его изложении на занятиях и в ходе промежуточной аттестации.

Баллы/ Шкала ECTS	Оценка по дисциплине	Критерии оценки результатов обучения по дисциплине
		Обучающийся испытывает серьёзные затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, не владеет необходимыми для этого навыками и приёмами. Демонстрирует фрагментарные знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции на уровне «достаточный», закреплённые за дисциплиной, не сформированы.

5.3. Оценочные средства (материалы) для текущего контроля успеваемости, промежуточной аттестации обучающихся по дисциплине

Контрольные вопросы

Основные методы лингвистического исследования: интроспекция, эксперимент и наблюдение над реальностью. Место корпусной лингвистики в этом противопоставлении.

1. Лингвистические корпуса: определение и примеры применения в лингвистических исследованиях.
2. Корпуса русского языка
3. Корпуса английского языка
4. Типы разметки в корпусах
5. Стандарты морфологической разметки для русского и английского языка. Омонимия и её разрешение.
6. Количественные исследования на корпусном материале. Базовые методы статистики в корпусных исследованиях.
7. Нормирование частотности языковых единиц в корпусах различного объёма.
8. Частотные словари. Закон Ципфа.
9. Исследование сочетаемости слов при помощи корпусов. Коллокации и меры их оценки. Лексические функции и их корпусное исследование.
10. Дифференциальные исследования на корпусном материале и приспособленность различных корпусов русского и английского языка для их проведения.
11. Проблема отбора текстов в корпус, репрезентативности и сбалансированности корпуса.
12. Многоязычные корпуса и их использование в лексикографии и в преподавании иностранных языков.
13. Интернет как корпус. Поисковые системы как заменитель корпусов («Googleology»), Яндекс.Блоги.
14. Создание пользовательских корпусов:
15. Применение корпусных методов в различных областях лингвистики.

Образцы вопросов для мини-тестов

1. *Вопрос из мини-теста по теме «Абсолютная и относительная частотность. ipm (instances per million) как мера относительной частотности» (тест проводится в начале семинара №6)*
Вычислите относительную частотность леммы «социалистический» в НКРЯ по текстам 1961–2019 годов и отдельно по десятилетиям (1961–1970, 1971–1980, …, 2011–2019). Для шести значений по десятилетиям определите среднее арифметическое, медиану, стандартное отклонение и вычислите D Жюйана.

1. *Вопрос из мини-теста по теме «Основные методы и подходы к извлече-нию ключевых слов. Меры ключёвости» (тест проводится в начале семинара №13)*

Укажите, для каких значений константы, прибавляемой к числителю и знаменателю в простой формуле ключёвости, слово «облако» является более ключевым для поэтического корпуса НКРЯ в сравнении с основным, чем слово «длань».

2. *Вопрос из мини-теста по теме «Частотные списки как основа для сравнения корпусов» (тест проводится в начале семинара №15)*

Какое максимальное значение R в частотном словаре Ляшевской и Шарова может иметь слово, которое встретилось в корпусе, легшем в основу словаря, 500 раз?
(А) 0; (Б) 1; (В) 100; (Г) 500

Образец домашнего задания

Задание выполняется в качестве подготовки к семинару №6 «Закон Ципфа»

Один лингвист подсчитал частотность слов в языке L. Готовясь к докладу, он записал на листочек, сколько раз встретились в доступных ему текстах пятнадцать самых частотных слов этого языка. К сожалению, лингвист пролил на этот листочек кофе, так что некоторые цифры теперь не читаются. Его записи выглядят так:

domin	6749
dotem	8998
dun	3001
ga	4503
grimun	2697
grumid	2075
kugrum	1801
kumun	27005
letun	3374
led	?854
mat	2249

mig	2454
mugun	??01
mulunt	1930
munt	13497

Помогите лингвисту восстановить цифры, залитые кофе. Поясните ваше решение.

Образцы коллективных практических заданий

1. Образец №1 (выполнение и обсуждение задания происходит на семинаре №4)
Оцените качество морфологической разметки словоформы «третью» в корпусе Araneum Russicum Minus.

2. Образец № 2 (выполнение и обсуждение задания происходит на семинаре №14)

Выберите пять любых существительных, которые достаточно часто встречаются в НКРЯ после леммы «зелёный», и оцените, насколько сильную коллокацию с этим словом они образуют при помощи любой известной вам меры.

Примеры заданий из промежуточной письменной контрольной

3. Посмотрите шуточную сцену из телепередачи *A bit of Fry and Laurie* и ознакомьтесь с ее расшифровкой. Выполните следующие задания.
А. Разработайте формат корпусной разметки для данной сцены.
Б. Проведите автоматический морфологический анализ данного текста с помощью таггера CLAWS7. Найдите ошибки в морфологической разметке и объясните их происхождение.

- Every day in Britain, more than 10 million people are mad. That's the disturbing conclusion in a report just published called Is Britain Turning into a Nation of Mad People? Dr Mijory Marjorie is with me now.
Dr Marjorie, just how serious is this problem?
- It's very serious.
- Wait a minute, I haven't finished yet.
- Sorry.
- In real terms.
- Okay?
-Yes, go on. Yes.
- It's very serious indeed. In 1957, when records began, we were, I think, the sixth maddest country in Europe.
Whereas last year's figures show that now Britain, I'm afraid, leads the European community
- It is a community, isn't it?
- Yes. Britain now leads Europe in terms of being mad.
- Well, that's a worrying trend, certainly.
-You're very kind.
- Right. Right. Now, for those viewers who may just have switched on this minute, would you mind having this conversation with me all over again?
- Fine with me.

- Right. Is Britain turning into a nation of mad people? Dr Mijory Marjorie is with me now. Dr Marjorie, just how serious is this problem in real terms?
- Not particularly.
- Not particularly what?
- Serious.
- Isn't it?
- No, no, no.
- Right, when we talk about Britain's increasing madness, what sort of madness are we really discussing?
- Well, all sorts really. From the kind of madness that makes people want to put on a hat when they get into a car, to the really extreme madness that prompts people to go to the theatre.
- Right. So that's quite a broad basket of madness, really, isn't it?
- Well, we've tried to be pretty thorough.
- Right, right. Now, for those people who've just tuned in right now. Could I suggest that you invest in a copy of Radio Times? That way you can plan your viewing properly and stop butting into programmes five minutes after they've begun. I mean, you wouldn't after all start a novel at chapter 5, would you?
- Well, you would if the first four chapters were rubbish.
- Oh, be quiet. Um, now Dr Marjorie, examining the causes beneath and behind and, to some extent, to one side of Britain's underlying and increasing madness, what exactly are they?
- Well, we examined...
- Sorry, sorry, who's "we"?
- My mother and I. And a woman called Alice.
- Fine, fine.
- And we came up with some pretty interesting results. You see, essentially, madness is like charity. It begins at home.
- Oh, that's interesting.

2. В Брауновском корпусе английского языка 1 161 192 слова (считая словами в том числе и знаки препинания). Дана таблица, в которой указаны частоты некоторых слов и их сочетаний (без учёта регистра; START и END — начало и конец предложения соответственно):

	Всего	START	*he*	*him*	*men*	*saw*	*the*	.	END
START	57340	0	2983	0	17	0	6857	54	0
he	9548	0	0	0	0	93	3	3	0
him	2619	0	19	0	0	0	47	494	0
men	763	0	7	0	0	0	2	62	2
saw	352	0	0	20	1	0	64	7	0
the	69971	0	0	0	97	0	0	1	9
.	49346	0	0	0	0	0	0	0	49346
END	57340	0	0	0	0	0	0	0	0

а) Сравните вероятности предложений *The men saw him*. и *Him the saw men.* в предположении, что каждое слово зависит только от своего соседа слева.

б) Как выглядит наиболее вероятное 5-словное английское предложение, составленное из этих слов (START и END не входят в счёт слов, а точка входит)?

6. Учебно-методическое и информационное обеспечение дисциплины

6.1. Список источников и литературы

Литература

Грудева, Е.В. **Корпусная лингвистика** : учеб. пособие / Е.В. Грудева. - 3-е изд., стер. - Москва : ФЛИНТА, 2017. - 165 с. - ISBN 978-5-9765-1497-3. - Текст : электронный. - URL: https://znanium.com/catalog/product/1032488 (дата обращения 23.03.24)

Перечень электронных ресурсов и программных средств

Национальный корпус русского языка (НКРЯ): http://ruscorpora.ru/
Мультимедийный корпус в составе НКРЯ: http://ruscorpora.ru/mycorpora-murco.html
ruTenTen: http://the.sketchengine.co.uk
Открытый корпус (OpenCorpora): http://opencorpora.org
Генеральный Интернет-корпус русского языка (ГИКРЯ): http://webcorpora.ru

7. Материально-техническое обеспечение дисциплины

Занятия по курсу можно проводить с максимальной эффективностью в компьютерном классе или аудитории с доступом в Интернет, проектором и экраном для презентаций. Необходимо также наличие доски или флипчарта, чтобы преподаватель мог разбирать примеры по ходу объяснения и записывать задания. Для самостоятельной работы студентам необходимо рабочее место, оборудованное персональным компьютером с доступом в Интернет, аудио- и видеоплеером (Windows Media Player, MPC, WinAmp, VLC и т.п.) а также офисными программами (Microsoft Office, OpenOffice, LibreOffice, Zoho Office и т.п.).

8. Обеспечение образовательного процесса для лиц с ограниченными возможностями здоровья и инвалидов

В ходе реализации дисциплины используются следующие дополнительные методы обучения, текущего контроля успеваемости и промежуточной аттестации обучающихся в зависимости от их индивидуальных особенностей:

- для слепых и слабовидящих:

- лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением;
- письменные задания выполняются на компьютере со специализированным программным обеспечением, или могут быть заменены устным ответом;
- обеспечивается индивидуальное равномерное освещение не менее 300 люкс;
- для выполнения задания при необходимости предоставляется увеличивающее устройство; возможно также использование собственных увеличивающих устройств;
- письменные задания оформляются увеличенным шрифтом;
- экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.

- для глухих и слабослышащих:

- лекции оформляются в виде электронного документа, либо предоставляется звукоусиливающая аппаратура индивидуального пользования;

- письменные задания выполняются на компьютере в письменной форме;

- экзамен и зачёт проводятся в письменной форме на компьютере; возможно проведение в форме тестирования.

- для лиц с нарушениями опорно-двигательного аппарата:

- лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением;

- письменные задания выполняются на компьютере со специализированным программным обеспечением;

- экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.

При необходимости предусматривается увеличение времени для подготовки ответа.

Процедура проведения промежуточной аттестации для обучающихся устанавливается с учётом их индивидуальных психофизических особенностей. Промежуточная аттестация может проводиться в несколько этапов.

При проведении процедуры оценивания результатов обучения предусматривается использование технических средств, необходимых в связи с индивидуальными особенностями обучающихся. Эти средства могут быть предоставлены университетом, или могут использоваться собственные технические средства.

Проведение процедуры оценивания результатов обучения допускается с использованием дистанционных образовательных технологий.

Обеспечивается доступ к информационным и библиографическим ресурсам в сети Интернет для каждого обучающегося в формах, адаптированных к ограничениям их здоровья и восприятия информации:

- для слепых и слабовидящих:

- в печатной форме увеличенным шрифтом;

- в форме электронного документа;

- в форме аудиофайла.

- для глухих и слабослышащих:

- в печатной форме;

- в форме электронного документа.

- для обучающихся с нарушениями опорно-двигательного аппарата:

- в печатной форме;

- в форме электронного документа;

- в форме аудиофайла.

Учебные аудитории для всех видов контактной и самостоятельной работы, научная библиотека и иные помещения для обучения оснащены специальным оборудованием и учебными местами с техническими средствами обучения:

- для слепых и слабовидящих:

  - устройством для сканирования и чтения с камерой SARA CE;

  - дисплеем Брайля PAC Mate 20;

  - принтером Брайля EmBraille ViewPlus;

- для глухих и слабослышащих:

  - автоматизированным рабочим местом для людей с нарушением слуха и слабослышащих;

  - акустический усилитель и колонки;

- для обучающихся с нарушениями опорно-двигательного аппарата:

  - передвижными, регулируемыми эргономическими партами СИ-1;

  - компьютерной техникой со специальным программным обеспечением.

9.1. Методические материалы

Семинар 1. 1. Национальный корпус русского языка (НКРЯ); 2. ruWac; 3. ruTenTen; 4. Хельсинкский аннотированный корпус (ХАНКО); 5. Интегрум.

Цель семинара: закрепить изложенные в курсе знания о лингвистических корпусах. Практическое применение конкретных корпусов.

Контрольные вопросы:

1. Лингвистические корпуса: определение и примеры применения в лингвистических исследованиях.
2. Корпуса русского языка

Семинар 2. Модус, стиль, регистр

Цель семинара: закрепить изложенные в лекции представления о фундаментальных противопоставлениях между устным и письменным модусами и обсудить соотношение понятий «модус», «стиль» и «регистр» на примере концепции русской разговорной речи. Обнаружение рефлексов устности vs. письменности в предложенных текстах. Выявление нарушений контакта и вовлеченности в образцах нормированной устной речи.

Контрольные вопросы:

1. Фундаментальные различия между устным и письменным модусами
2. Первичность устного модуса
3. Формальность и функциональный стиль; регистр
4. Основные положения концепции русской разговорной речи по Е. А. Земской
5. Контролируемое сопоставление модусов

Семинар 3. Диалог и монолог

Цель занятия: закрепить изложенные в лекции представления о критериях разграничения диалогической и монологической коммуникативных ситуаций (в частности, идею о нечеткости этого разграничения), базовых ролях участников коммуникации, принципах описания структуры диалога и разделяемых участниками диалога прагматических и социокультурных принципах коммуникации. Сообщения о прочитанной литературе. Выполнение коллективного практического задания — анализ диалогического текста с выявлением правил чередования реплик, случаев иллокутивного (само)вынуждения, (не)предпочтительных реакций в смежных парах, способов реализации принципа кооперации и применения стратегий вежливости.

Контрольные вопросы:

1. Основные роли участников в диалогической коммуникативной ситуации
2. Анализ бытового диалога: принципы чередования, смежные пары
3. Иллокутивное вынуждение в структуре диалога
4. Принцип кооперации и максимы Грайса. Контекстные импликатуры
5. Стратегии вежливости в теории П. Браун и С. Левинсона
6. Критерии противопоставления монолога и диалога

Семинар 4. Структурная и содержательная связность

Цель занятия: закрепить изложенные на лекции представления о соотношении понятий когезии (структурной связности) и когерентности (содержательной связности). Роль имплицитных знаний (в частности, фреймов) в поддержании содержательной связности. Эксперимент на понимание текста с «выключенным» фреймом. Коллективное практическое задание: выявление в предложенном тексте маркеров структурной связности и определение их классификационной принадлежности.

Контрольные вопросы:
1. Подходы к формализации понятия «дискурсивная связность». Когезия и когерентность
2. Классификация маркеров структурной связности (когезии)
3. Ментальная репрезентация как когнитивная модель содержательной связности (когерентности)
4. Способы построения ментальной репрезентации «снизу вверх» и «сверху вниз». Роль фреймов в процессе понимания текста

Семинар 5. Глобальная структура дискурса

Цель занятия: закрепить изложенные на лекции представления об иерархическом характере дискурсивной структуры и подходах к выявлению единиц глобального уровня. Эксперименты по восстановлению границ абзацев в письменных текстах различных жанров. Обсуждение возможных методов регистрации пограничных маркеров в устных текстах. Проверка домашнего задания: получение макроструктуры текста в терминах Т. ван Дейка.

Контрольные вопросы:
1. Дискурсивная структура как иерархия. Локальная и глобальная структура
2. Сверхфразовое единство: стандартная структура и механизмы внутренней организации
3. Абзац как единица глобальной структуры письменного текста
4. Дискурсивный топик: механизмы развития, соотношение с фреймами / сценариями
5. Маркеры границ эпизодов в текстах различных типов
6. Макроструктура текста по Т. ван Дейку

Семинар 6. Локальная структура дискурса

Цель занятия: применить на практике изложенные на лекции принципы сегментации устной речи на элементарные дискурсивные единицы (ЭДЕ). Выявление ЭДЕ в предложенных устных текстах. Проверка домашнего задания: первичная классификация ЭДЕ на основании степени соответствия простой клаузе.

Контрольные вопросы:
1. Элементарная дискурсивная единица: формальные свойства и когнитивная мотивация
2. Статусы активации; данное, новое, доступное. Ограничение одной новой идеи
3. Стандартное синтаксическое наполнение ЭДЕ. Роль клаузы в локальной структуре
4. Субклаузальные ЭДЕ: классификация и когнитивная мотивация

Семинар 7. Между локальной и глобальной структурой

Цель занятия: закрепить изложенные на лекции представления о подходах к описанию связи между локальным и глобальным уровнями дискурсивной структуры. Выступления по прочитанной литературе. Проверка домашнего задания: выявление статистически значимых результатов в серии работ, посвященных комбинированию единиц локальной структуры при

построении более крупных комплексов. Обсуждение применимости понятия «предложение» к данным устной речи.

Контрольные вопросы:
1. Предложение в письменной и устной речи
2. Техники комбинирования ЭДЕ в устной речи
3. Коммуникативное членение высказывания как фактор дискурсивной структуры
4. Коммуникативная динамика текста

Семинар 8. Теория риторической структуры

Цель занятия: на практике закрепить основные положения теории риторической структуры, введенные на лекции. Проверка домашнего задания: выполнение РРС-анализа предложенного текста. Применение классификационных параметров к встречающимся в анализе риторическим отношениям. Выявление способов маркирования риторических отношений.

Контрольные вопросы:
1. Основные положения теории риторической структуры (РРС). Параметры классификации риторических отношений
2. Способы маркирования риторических отношений в текстах различных жанров
3. Расширение аппарата РРС для анализа устных рассказов
4. Практические и теоретические применения РРС. Корпуса с РРС-разметкой

Семинары 9-11

На семинарах 9-10 студенты пишут промежуточную контрольную работу, в которой оцениваются их знания по разделам II-IV курса. Для проверки полученных знаний и навыков студентам предлагается выполнить практические задания, предполагающие, помимо прочего, просмотр видеофрагментов. На семинаре 11 происходит обсуждение результатов контрольной работы и разбор заданий.

Материально-техническое обеспечение занятия:
Доска, проектор, акустическая система.

Семинар 12. Референция в дискурсе

Цель занятия: обсудить со студентами дискурсивно ориентированные подходы к описанию референции. Выступления по прочитанной литературе. Демонстрация важности понятия активации для моделирования референциального выбора в дискурсе. Обсуждение описанных в литературе факторов активации; анализ различий между понятиями линейного и риторического расстояния. Предварительное тестирование модели референциального выбора А. А. Кибрика.

Контрольные вопросы:
1. Проблема референциального выбора и степень активации референта
2. Факторы активации. Линейное и риторическое расстояние. Референциальный конфликт
3. Модель референциального выбора А. А. Кибрика

Семинары 13-14. Дискурсивные основания грамматических категорий

Цель занятий: обсудить со студентами содержательные принципы дискурсивно ориентированных подходов к описанию грамматических категорий. Выступления по прочи-

танной литературе. Обсуждение конкретных примеров влияния дискурсивных факторов на грамматическую форму.

Контрольные вопросы:
1. Соотношение дискурса и грамматики. Примеры влияния дискурса на грамматику. Emergent grammar как радикальная ветвь функционализма
2. Дискурсивные мотивации основных стратегий падежного маркирования
3. Роль клаузы в порождении и восприятии дискурса. Совместное построение реплик в диалоге
4. Теория риторической структуры и сложный синтаксис
5. Дискурсивные основания переходности. Противопоставление основной линии и фона
6. Когнитивная мотивация подлежащего: данные лингвистических экспериментов

Семинар 15. Общие вопросы мультиканальной лингвистики

Цель занятия: обсудить со студентами принципиальную несводимость естественной коммуникации к обмену вербальными сигналами; предложить схему мультиканальной коммуникации и наметить основные проблемы, решаемые в мультиканальных исследованиях. Выступления по прочитанной литературе. Работа с электронными ресурсами, посвященными мультиканальности. Классификация жестов по признаку конвенциональности (континуум Кендона). Анализ внутренней структуры иллюстративных жестов.

Контрольные вопросы:
1. Мультиканальная коммуникация и ее составляющие
2. Регистрация мультиканального поведения. Методология мультиканального исследования
3. Степень конвенциональности невербальных единиц. Континуум Кендона
4. Классификация и внутренняя структура иллюстративных жестов.

Семинар 16. Координация речи и невербальных коммуникативных каналов

Цель занятия: обсудить со студентами некоторые подходы к анализу координации речи и других коммуникативных каналов. Выступления по прочитанной литературе. «Единая точка роста» по Д. Макнилу и проверка гипотезы об опережающем характере жестикуляции. Роль движения взгляда в коммуникации.

Контрольные вопросы:
1. Центральный статус речи в мультиканальной коммуникации
2. Анализ координации речи и других коммуникативных каналов. Гипотезы компенсации и сотрудничества
3. «Единая точка роста» по Д. Макнилу. Гипотеза опережающего характера жестикуляции
4. Роль движения взгляда в естественной коммуникации

Семинары 17-19

Семинары 17-19 посвящены обсуждению предварительных результатов индивидуальных исследований студентов, окончательные результаты которых они должны будут представить во время экзамена. Оцениваются как выступления со своими исследованиями, так и участие в дискуссии.

9.2. Методические рекомендации для студентов по освоению дисциплины

Курс «Технологии корпусной лингвистики» нацелен как на получение студентами обширных теоретических знаний, так и на овладение ими базовых методов практического анализа текстов различных корпусов. Теоретические основы излагаются на лекциях, а также во время семинарских занятий 12-16. Проверка усвоения базовых понятий курса осуществляется в начале каждого семинара при помощи мини-теста. Практические навыки анализа текстов совершенствуются на семинарах, подготовка к которым обычно требует выполнения того или иного домашнего задания. Проверка того, насколько полно студенты овладели комплексом теоретических знаний и практических умений, проводится (i) в формате письменной контрольной работы посредине семестра, (ii) при подготовке и презентации результатов индивидуального курсового проекта.

Для успешного освоения программы студентам необходимо присутствовать на лекциях и семинарских занятиях, внимательно читать и при необходимости конспектировать обязательную и рекомендованную дополнительную литературу, самостоятельно формулировать вопросы и задавать их преподавателю во время занятий. Безусловным требованием к студентам также является их готовность участвовать в коллективной дискуссии и проявлять критический подход к постановке и способам решения конкретных задач. Поскольку выполнение практических заданий может быть связано с использованием тех или иных программных средств, от студентов требуются базовые навыки работы с различными компьютерными редакторами и умение быстро обучаться пользованию новыми программными продуктами.

АННОТАЦИЯ

Курс «Корпусная лингвистика» читается УНЦ компьютерной лингвистики ИЛ РГГУ.

Предмет курса – современные представления лингвистики о компьютерных корпусах текстов и основные методы корпусного анализа.

Цель курса – освоение студентами базовых понятий и методов лингвистически ориентированного корпусного анализа, классификации корпусов, принципов описания и составления корпусов и вопросов соотнесения знаний, полученных с применением лингвистических корпусов, с лингвистическими знаниями, полученными иным способом.

Задачи курса: научить студентов

- владению основными понятиями и категориями современной лингвистики;
- владение основными методами фонологического, морфологического, синтаксического, дискурсивного и семантического анализа с учетом языковых и экстралингвистических факторов;
- владение основными способами описания и формальной репрезентации денотативной, концептуальной, коммуникативной и прагматической информации, содержащейся в тексте на естественном языке;
- способность определять макроструктуру и микроструктуру дискурса с учетом специфики его жанров и функционально-стилевых разновидностей.

Курс нацелен на формирование следующих компетенций:

УК-4 Способен осуществлять деловую коммуникацию в устной и письменной формах на государственном зыке Российской Федерации и иностранном(ых) языке(ах)	4.3	Использует информационно-коммуникационные технологии при поиске необходимой информации в процессе решения стандартных коммуникативных задач для достижения профессиональных целей на государственном м иностранном (ых) языке (ах)
ПК-3 Способен к научно-исследовательской деятельности	3.1	Владеет основами методов научного исследования, информационной и библиографической культурой
	3.2	Владеет стандартными методиками поиска, анализа и обработки материала исследования
	3.3	Умеет логично и последовательно представить результаты своего исследования

По завершении обучения студент, полностью освоивший программу, должен:

знать:

- основные понятия и методы современной корпусной лингвистики;
- базовые принципы создания и использования корпусов различных типов.

уметь:

- определять классификационную принадлежность предложенного для анализа лингвистического корпуса и объяснять свое решение;
- осуществлять самостоятельный выбор лингвистического корпуса, наиболее подходящего для тематики исследования;
- обнаруживать в реальных текстах языковые явления, которые можно изучать при помощи корпусов;
- сопоставлять вклад текстов различных типов и жанров в корпус.

владеть:

- современной терминологией корпусной лингвистики;
- навыками корпусного анализа;
- основными методами статистической обработки языковых данных.

Программой предусмотрены следующие **виды контроля**: промежуточная аттестация в форме экзамена.

Общая трудоемкость освоения дисциплины составляет 3 зачетных единицы.