

МИНОБРНАУКИ РОССИИ



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

ИНСТИТУТ ЛИНГВИСТИКИ
УНЦ компьютерной лингвистики

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ЛИНГВИСТИКЕ

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Направление 45.04.02 «Лингвистика»
Направленность «Иностранные языки»
Уровень квалификации выпускника: магистр

Форма обучения: очная

РПД адаптирована для лиц
с ограниченными возможностями
здоровья и инвалидов

Москва 2019

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ЛИНГВИСТИКЕ
Рабочая программа дисциплины

Составитель:

кандидат филологических наук, доцент УНЦ компьютерной лингвистики

Пиперски Александр Чедович

Ответственный редактор:

заведующий УНЦ компьютерной лингвистики

Селегей Владимир Павлович

УТВЕРЖДЕНО

Протокол заседания УНЦ компьютерной лингвистики
№ 9 от 22.06.2019

ОГЛАВЛЕНИЕ

1. Пояснительная записка

1.1 Цель и задачи дисциплины

1.2. Формируемые компетенции, соотнесённые с планируемыми результатами обучения по дисциплине

1.3. Место дисциплины в структуре образовательной программы

2. Структура дисциплины

3. Содержание дисциплины

4. Образовательные технологии

5. Оценка планируемых результатов обучения

5.1. Система оценивания

5.2. Критерии выставления оценок

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

6. Учебно-методическое и информационное обеспечение дисциплины

6.1. Список источников и литературы

6.2. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

7. Материально-техническое обеспечение дисциплины

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

9. Методические материалы

9.1. Планы практических (семинарских, лабораторных) занятий

9.2. Методические рекомендации по подготовке письменных работ

9.3. Иные материалы

Приложения

Приложение 1. Аннотация дисциплины

Приложение 2. Лист изменений

1. Пояснительная записка

1.1. Цель и задачи дисциплины

Цель дисциплины – познакомить магистрантов с наиболее актуальными современными компьютерными корпусами текстов и лексикографическими ресурсами, программами обработки текста, с технологиями создания собственных исследовательских корпусов, научить применять методы создания собственных исследовательских корпусов, работы с корпусными данными, методы обработки этих данных в собственных научных исследованиях.

Для достижения установленной цели решаются следующие учебные **задачи**:

- познакомить магистрантов с последними трендами в области корпусной лингвистики, с основными типами лингвистических ресурсов, доступными в Интернете и используемыми в профессиональной деятельности: с наиболее актуальными лингвистическими корпусами (национальными и проблемными), специальными программами обработки текстов, лексикографическими ресурсами и т.п.;
- на примерах показать, какие новые возможности в исследовании грамматики и лексики языка дает использование корпусных, а также применение современных методов обработки этих данных;
- дать представления о проблемах создания корпусов, об основных принципах разработки данных ресурсов и об основных требованиях, предъявляемых к ним;
- познакомить с технологиями и проблемами разметки корпусов;
- научить работать с современными пакетами обработки собственных корпусов;
- научить применять специальные методы лингвистических исследований, использующие данные корпусов, в том числе и статистические методы исследования;
- обучить практическим навыкам по применению корпусных методов в своей исследовательской работе.

1.2. Формируемые компетенции, соотнесённые с планируемыми результатами обучения по дисциплине:

Коды компетенции	Содержание компетенций	Перечень планируемых результатов обучения по дисциплине
ПК-28	готовность работать с основными информационно-поисковыми и экспертными системами, системами представления знаний, синтаксического и морфологического анализа, автоматического синтеза, распознавания и понимания речи, обработки лексикографической информации и автоматизированного перевода, автоматизированными системами идентификации и верификации личности	Знать <ul style="list-style-type: none"> ▪ принципы создания собственных исследовательских корпусов; ▪ основные типы исследовательских задач, решаемых с использованием корпусов; ▪ основные применяемые в корпусных исследованиях лексики и грамматики методы ▪ требования, предъявляемые к верификации результатов

		<ul style="list-style-type: none"> ▪ основные методы статистического анализа корпусных данных. <p>Уметь:</p> <ul style="list-style-type: none"> ▪ осуществлять поиск в корпусах в соответствии с исследовательской гипотезой в области грамматики и лексикографических исследований; ▪ создавать и размечать собственные исследовательские и обучающие корпуса; ▪ работать с различными типами программ обработки текстов: конкордансерами, программами для поиска коллокаций, создания частотных списков и т.п., корпусными менеджерами; ▪ разрабатывать методический материал по основным языковым дисциплинам с использованием корпусов; ▪ осуществлять мониторинг и оценку различных типов современных корпусных ресурсов и выбирать ресурсы, подходящие для выполнения тех или иных исследовательских и производственных задач. <p>Владеть:</p> <ul style="list-style-type: none"> ▪ основными методами и средствами профессионального компьютерного инструментария для исследовательской и практической работы; ▪ методами сбора материала с использованием корпусов; ▪ методами анализа корпусных данных, включая статистические методы.
ПК-31	владение современными методиками разработки лингвистического обеспечения в автоматизированных системах различного профиля	<p>Знать:</p> <ul style="list-style-type: none"> ▪ основные принципы создания корпусов и других компьютерных лингвистических ресурсов; ▪ характеристики и особенности современных доступных в Интернете национальных и проблемных корпусов, широко используемых в лингвистических исследованиях, включая недав-

		<p>но вошедшие в лингвистическую практику;</p> <ul style="list-style-type: none"> ■ стандарты, типы и проблемы разметки корпусов, включая такие современные типы разметки, как дискурсивную разметку, интонационную разметку устных корпусов и т.п., применяемые в разметке технологии; <p>Уметь:</p> <ul style="list-style-type: none"> ■ осуществлять поиск в корпусах в соответствии с исследовательской гипотезой в области грамматики и лексикографических исследований; ■ создавать и размечать собственные исследовательские и обучающие корпуса; ■ работать с различными типами программ обработки текстов: конкордансерами, программами для поиска коллокаций, создания частотных списков и т.п., корпусными менеджерами; ■ разрабатывать методический материал по основным языковым дисциплинам с использованием корпусов. <p>Владеть:</p> <ul style="list-style-type: none"> ■ основными методами и средствами профессионального компьютерного инструментария для исследовательской и практической работы.
--	--	---

1.3. Место дисциплины в структуре образовательной программы

Дисциплина «Информационные технологии в лингвистике» является частью Блока 1 учебного плана ОП ВО магистратуры «Иностранные языки» по направлению подготовки 45.04.02 – Лингвистика и имеет статус дисциплины Вариативной части Индекс Б1.В.ДВ.06.01 Дисциплина по выбору.

Для освоения дисциплины необходимы знания, умения и владения, сформированные в ходе изучения «Общее языкознание и основы лингвистических учений» и «Семиотика».

Дисциплина формирует компетенции, необходимые для прохождения преддипломной практики и итоговой аттестации.

2. Структура дисциплины для очной формы обучения

Общая трудоёмкость дисциплины составляет 2 з.е., 72 ч., в том числе контактная работа обучающихся с преподавателем 20 ч., самостоятельная работа обучающихся 52 ч.

Учебная нагрузка студента состоит из посещения и активного восприятия (конспектирования) лекций; активной работы на семинарских занятиях; самостоятельной проработки изучаемых разделов по рекомендованной литературе и материалам лекций; выполнения домашних заданий; подготовки к контрольным работам и промежуточной аттестации. Распределение нагрузки (лекции, семинары и СРС – самостоятельная работа студентов) по изучаемым разделам отражено в приведенной ниже таблице.

№ п/ п	Раздел Дисциплины	Семестр	Неделя семестра	Виды учебной работы, включая самостоятельную работу студентов и трудоёмкость (в часах)				Формы текущего контроля успеваемости (по неделям семестра) Форма промежуточной аттестации (по семестрам)
				лек- ции	прак- тич. заня- тия	само- стоя- тель- ная рабо- та		
1.	Введение. Общее представление о корпусах и корпусной лингвистике. Стандарты разметки. Типы разметки корпусов.	3	1		2	4		ДЗ1. Упр. по теме “Особенности поиска и управления выдачи в НКРЯ”.
2.	Проблемные корпуса (параллельные, диалектные, мультимедийные и др.)	3	2		2	4		ДЗ2. Практическая работа. Разметка мультимедийного / аудиокорпуса
3.	Поиск в корпусе. Использование языка SQP для поиска в корпусе. Составление сложных запросов к корпусу.	3	3		2	4		ДЗ3. Запросы для поиска сложных конструкций с использованием языка регулярных выражений: письменный отчет
4.	Особенности различных типов разметки. Морфологическая разметка	3	4		2	4		ДЗ4. Практическая работа. Тестирование морфологической или синтаксической разметки: письменный отчет

5.	Особенности различных типов разметки. Синтаксическая разметка	3	5		2	2		ДЗ5. Практическая работа. Тестирование морфологической или синтаксической разметки: письменный отчет
6.	Особенности разметки: другие типы разметки	3	6		2	6		ДЗ6. Практическая работа. Семантическая разметка / анафорическая разметка. Контрольная работа по теме: «Корпуса и лингвистические ресурсы»
7.	Методы корпусных исследований. Анализ примеров корпусных исследований	3	7		2	4		ДЗ7. Реферирование статьи из рекомендованного списка. Составление краткого ТЗ к собственному исследовательскому проекту.
8.	Инструменты разметки собственного исследовательского корпуса	3	8		2	4		ДЗ8. Разработка параметров и схемы разметки исследовательского корпуса, разметка корпуса с помощью специальной программы
9.	Составление конкордансов, частотных списков, списков коллокаций с использованием специальных программ	3	9		2	4		ДЗ9. Обработка данных корпуса с использованием специальных программ
10.	Итоговая аттестация				2	16		Защита исследовательского проекта
	.				20	52		

3. Содержание дисциплины

РАЗДЕЛ I. Введение. Общее представление о корпусах и корпусной лингвистике.

- 1.1. Краткая история. Предмет и задачи курса.
- Краткая история корпусной лингвистики. Преимущества современных корпусных исследований. Возможность объединения формального и эмпирического подхода в современной корпусной лингвистике. Компьютерные ресурсы, необходимые лингвистам для решения различных задач. Задачи, решаемые с помощью компьютерных ресурсов.
- 1.2. Основные понятия корпусной лингвистики.
- Корпус. Национальный корпус. Проблемный корпус. Основные единицы. Основные требования, предъявляемые к корпусу. Поиск в корпусе. Основные требования и параметры поиска.
- 1.3. Стандарты и типы разметки. Основные принципы и методы разметки корпусов. Современные технологии разметки корпусов.

РАЗДЕЛ II. Корпуса и инструментарий работы с корпусами

- 2.1. Типы программ обработки текста, методы работы с программами обработки текста.
- 2.2. Программы разметки собственных исследовательских корпусов

РАЗДЕЛ III. Основные методы использования корпусов в грамматике и лексике

- 3.1. Области использования корпусных данных
- 3.2. Методы сбора и статистической обработки корпусных данных. Общие статистические характеристики: меры средней тенденции и изменчивости. Проверка статистических критериев, исследование зависимостей. Корреляционный анализ.

РАЗДЕЛ IV. Примеры корпусных исследований

- 4.1. Примеры использования корпусов в обучении и научных исследованиях: методология создания дидактических материалов с использованием корпусов; методология создания исследовательского корпуса с использованием корпусов общего назначения. Примеры корпусных диалектологических, диахронических, социолингвистических и гендерных исследований, исследований стиля
- 4.3. Использование корпусов в лексикографической работе. Статистические методы в лексикографии
- 4.4. Сравнение корпусов. Стилеметрия.
- 4.5. Примеры применения корпусного анализа в грамматических исследованиях

4. Образовательные технологии

Для данной дисциплины образовательные технологии представлены лекциями, семинарами и самостоятельной работой студентов. На семинарах проводится обсуждение проблем, поднятых на лекциях, осмысление прочитанных студентами работ и материалов, разбираются выполненные дома тренировочные упражнения и задачи. Желающие могут делать небольшие сообщения и презентации на предложенные преподавателем темы. По наиболее значимым темам проводятся контрольные или коллоквиумы.

К рассмотрению и обсуждению привлекается материал на традиционных или электронных носителях. Предлагаются задания на поиск сведений в Интернете, их сопоставление и оценку (иногда перевод с других языков). Это помогает повысить интерактивность даже такой традиционной формы педагогической коммуникации, как лекция.

5. Оценка планируемых результатов обучения

5.1. Система оценивания

Оценка за семестр складывается из следующих составляющих (максимальная сумма 100 баллов):

- 1) оценки за посещение семинаров (всего 10 баллов) и активную работу на них (до 10 баллов) – итого за работу на семинарах до 20 баллов;
- 2) оценка за текущую контрольную работу (до 10 баллов);
- 3) оценка за разработку проекта / доклада по теме (до 20 баллов);
- 4) оценка за презентацию проекта / выступление с докладом (до 10 баллов);
- 5) итоговая контрольная работа (до 20 баллов);
- 6) итоговое собеседование (до 20 баллов).

Для получения высокой оценки студенту необходимо систематически демонстрировать устойчивые результаты обучения.

Полученный совокупный результат конвертируется в традиционную шкалу оценок и в шкалу оценок Европейской системы переноса и накопления кредитов (European Credit Transfer System; далее – ECTS) в соответствии с таблицей:

100-балльная шка- ла	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

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»	<p>Выставляется обучающемуся, если он глубоко и прочно усвоил теоретический и практический материал, может продемонстрировать это на занятиях и в ходе промежуточной аттестации.</p> <p>Обучающийся исчерпывающе и логически стройно излагает учебный материал, умеет увязывать теорию с практикой, справляется с решением задач профессиональной направленности высокого уровня сложности, правильно обосновывает принятые решения.</p> <p>Свободно ориентируется в учебной и профессиональной литературе.</p> <p>Оценка по дисциплине выставляется обучающемуся с учётом результатов текущей и промежуточной аттестации.</p> <p>Компетенции, закреплённые за дисциплиной, сформированы на уровне – «высокий».</p>
82-68/ C	«хорошо»/ «зачтено (хорошо)»/«зачтено»	<p>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.</p> <p>Обучающийся правильно применяет теоретические положения при решении практических задач профессиональной направленности разного уровня сложности, владеет необходимыми для этого навыками и приёмами.</p> <p>Достаточно хорошо ориентируется в учебной и профессиональной литературе.</p> <p>Оценка по дисциплине выставляется обучающемуся с учётом результатов текущей и промежуточной аттестации.</p> <p>Компетенции, закреплённые за дисциплиной, сформированы на уровне – «хороший».</p>
67-50/ D,E	«удовлетворительно»/ «зачтено (удовлетворительно)»/«зачтено»	<p>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.</p> <p>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.</p> <p>Демонстрирует достаточный уровень знания учебной литературы по дисциплине.</p> <p>Оценка по дисциплине выставляется обучающемуся с</p>

Баллы/ Шкала ECTS	Оценка по дисциплине	Критерии оценки результатов обучения по дисциплине
		учёт 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. 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 выставляется 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

Ниже приводятся контрольные вопросы, образцы домашних работ, тестов и контрольных работ, которые могут использоваться для оценивания уровня усвоения материала по данной дисциплине.

Образцы домашних заданий

1. Укажите, сколько раз в Основном корпусе НКРЯ встречается слово *сверхпроводимость* во всех формах. Выразите его частотность в ipm.
2. Сравните частотность употребления двусложных сравнительных союзов *точно, будто* и *словно* у нескольких русских поэтов по Поэтическому корпусу НКРЯ. Как изменяется частотность этих союзов во времени? С какими трудностями вы столкнулись при поиске?
3. Кто из русских писателей по данным НКРЯ реже всего употребляет в своей прозе союз *и* — Ф. М. Достоевский, М. А. Булгаков или М. А. Шолохов? Опишите, как вы получили ответ.
4. Сравните частотность притяжательных местоимений различных лиц и чисел в русских поэтических и прозаических текстах по НКРЯ.
5. Используя НКРЯ, сравните свойства русских конструкций *только и знает / делает / умеет что*. В какой форме в каждой из них чаще употребляется смысловой глагол: в инфинитиве или в форме, дублирующей форму вспомогательного глагола? В каком времени чаще используется каждая из этих конструкций: в прошедшем или в настоящем?

Образцы самостоятельных работ

1. Какой из перечисленных корпусов русско-¹³го языка не имеет морфологической разметки?
(А) Уппсальский корпус; (Б) НКРЯ; (В) ХАНКО; (Г) ruTenTen
2. Если мы ищем формы глагола «стоять» и находим фразу «Этот дом стоит миллион евро», это — ...
(А) True Positive; (Б) True Negative; (В) False Positive; (Г) False Negative
3. Сколько раз в корпусе встречаются *dis legomena*?
(А) 1; (Б) 2; (В) 3; (Г) 4
4. В каком из этих тагсетов число возможных тэгов наибольшее?
(А) CLAWS7; (Б) Penn Treebank Tagset; (В) Brown Corpus Tagset; (Г) MULTEXT-East (Russian)
5. Если вычислить количество вхождений каждого из типов слов в корпус C, медиана этого значения скорее всего будет составлять ...
(А) 2; (Б) 3; (В) 10; (Г) 100
6. Если изобразить на координатной плоскости с обычными (линейными) шкалами график, в котором по оси x отложен ранг слова, а по оси y — его частота, согласно закону Ципфа мы получим ...
(А) окружность; (Б) ветвь гиперболы; (В) параболу; (Г) прямую
7. Средняя частотность отдельного слова (mean word frequency, MWF) обычно ... по мере роста корпуса
(А) увеличивается; (Б) уменьшается; (В) остаётся неизменным; (Г) невозможно сказать
8. Первым корпусом, в названии которого употреблено слово «национальный», стал ...
(А) Болгарский национальный корпус; (Б) Чешский национальный корпус; (В) Британский национальный корпус; (Г) Национальный корпус русского языка
9. Оцените по основному подкорпусу НКРЯ вероятности перехода между тэгами: $P(A|S)$, $P(S|S)$ и $P(V|S)$
10. Какая вероятность перехода больше: $P(S|взмахнул)$ или $P(взмахнул|S)$? Подтвердите свой ответ с помощью НКРЯ.
11. Рассчитайте индекс Герфиндаля—Гиршмана для романа Чарльза Диккенса «Great Expectations» (в оригинале, без лемматизации)
12. Оцените с помощью Araneum Russicum Minus, в каком числе доля творительного падежа от общего количества форм существительных выше: в единственном или во множественно? Опишите сделанные запросы.

13. Сколько типов и сколько токенов насчитывается в следующем мини-корпусе? Кратко опишите возможные проблемы при подсчёте.

We did send your invitation on, but she's travelling so she may not have got it. She sent us a rather vague address in Ibiza, but we haven't heard from her since she was in Paris.

14. Найдите моду и медиану рангово-частотного профиля, а также TTR для романа Джейн Остин «Pride and Prejudice» (в оригинале, без лемматизации).

Образцы проверочных (контрольных) заданий

1. Установите, какое ударение чаще используется в прилагательном *допризывной* / *допризывный*. Какой корпус лучше подходит для такого исследования: НКРЯ или ruTenTen — и почему?
2. Постройте в Excel график распределения частотностей для 1000 наиболее частотных слов по одному из подкорпусов, представленных в словаре [Ляшевская, Шаров 2009], откладывая по оси *x* ранг слова, а по оси *y* — его частотность. Аппроксимируйте полученное распределение при помощи степенной функции и оцените, насколько хорошо оно описывается при помощи закона Ципфа.
3. Сравните частотность сочетаний *A and B* и *B and A* для всех пар английских цветообозначений из множества {*red, black, green, blue, white, pink*}, используя любой достаточно большой корпус английского языка (BNC, COCA и т. п.). Какие цветообозначения тяготеют к первому месту в словосочетании, а какие — ко второму и с какими фонетическими особенностями это может быть связано?
4. Используя подкорпус CHES в Birkbeck Spelling Error Corpus, оцените, в каких словах 10-летние англоязычные дети чаще всего допускают орфографические ошибки. Попробуйте обобщить полученные результаты.
5. Используя COCA, укажите, какие существительные чаще всего сочетаются с прилагательными *independent, free* и *autonomous*. Попробуйте обобщить различия в сочетаемости этих прилагательных.
6. Используя корпус Google Books через интерфейс Brigham Young University, сравните частотность и сочетаемость слов со значением приблизительности (*almost, approximately, nearly* и т. д.) в разные десятилетия XIX и XX веков.

Критерии оценивания самостоятельных и контрольных работ

—результат, содержащий полный правильный ответ, полностью соответствующий требованиям критерия — 85 – 100 %;

—результат, содержащий неполный правильный ответ (степень полноты ответа — более 75%) или ответ, содержащий незначительные неточности, т.е. ответ, имеющий незначительные отступления от требований критерия, — 75 – 84% от максимального количества баллов;

—результат, содержащий неполный правильный ответ (степень полноты ответа — до 75%) или ответ, содержащий незначительные неточности, т.е. ответ, имеющий незначительные отступления от требований критерия — 60 -74 % от максимального количества баллов;

—результат, содержащий неполный правильный ответ, содержащий значительные неточности, ошибки (степень полноты ответа — менее 60%) — до 60 % от максимального количества баллов;

—неправильный ответ (ответ не по существу задания) или отсутствие ответа, т.е. ответ, не соответствующий полностью требованиям критерия, — 0 % от максимального количества баллов.

Тематика рефератов по дисциплине

Студентам предлагаются рефераты по статьям из сборника Biber, Douglas & Randi Reppen (eds.). 2011. *Corpus linguistics*. 4 vols. London: Sage.

Вопросы для промежуточной аттестации

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

1. Корпусные менеджеры: Word-¹⁶ Smith, AntConc и т. д.
 2. Создание пользовательских корпусов в системе SketchEngine и возможности их исследования.
 3. Создание мультимодальных корпусов в программе ELAN.
16. Применение корпусных методов в различных областях лингвистики:
1. Грамматика
 2. Лексикография
 3. Социолингвистика и др.

Критерии оценивания для промежуточной аттестации обучающихся (вопросы к зачету)

- результат, содержащий полный правильный ответ, полностью соответствующий требованиям критерия – 85 – 100 %;
- результат, содержащий неполный правильный ответ (степень полноты ответа – более 75%) или ответ, содержащий незначительные неточности, т.е. ответ, имеющий незначительные отступления от требований критерия, – 75 – 84% от максимального количества баллов;
- результат, содержащий неполный правильный ответ (степень полноты ответа – до 75%) или ответ, содержащий незначительные неточности, т.е. ответ, имеющий незначительные отступления от требований критерия – 60 -74 % от максимального количества баллов;
- результат, содержащий неполный правильный ответ, содержащий значительные неточности, ошибки (степень полноты ответа – менее 60%) – до 60 % от максимального количества баллов;
- неправильный ответ (ответ не по существу задания) или отсутствие ответа, т.е. ответ, не соответствующий полностью требованиям критерия, – 0 % от максимального количества баллов.

6. Учебно-методическое и информационное обеспечение дисциплины

6.1. Список источников и литературы

Литература

Основная

1. Корпусные исследования по русской грамматике : сб. ст. / Рос. акад. наук, Ин-т языкознания ; [ред.-сост. К. Л. Киселева и др.]. - М. : Пробел-2000, 2009. - 513 с. ; 22 см. - Библиогр. в конце ст. - ISBN 978-5-98604-148-3 : 330.00. Ссылка на ресурс: <http://text.lib.rsuh.ru/macro/256.txt>
2. Плунгян Владимир Александрович. Введение в грамматическую семантику: грамматические значения и грамматические системы языков мира : учеб. пособие / В. А. Плунгян ; [М-во образования и науки Рос. Федерации, Гос. образоват. учреждение высш. проф. образования "Рос. гос. гуманитарный ун-т"]. - М. : РГГУ, 2011. - 669 с. ; 22 см. - Библиогр.: с. 498-581. - Указ.: с. 582-664. - ISBN 978-5-7281-1122-1 : 435.00.

6.2. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. Национальный корпус русского языка
2. British National Corpus
3. Corpus of Contemporary American English
4. Sketch Engine

7. Материально-техническое обеспечение дисциплины (модуля)

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.

Состав программного обеспечения (ПО), современных профессиональных баз данных (БД) и информационно-справочных систем (ИСС)

Перечень ПО

№п /п	Наименование ПО	Производитель	Способ распространения (лицензионное или свободно распространяемое)
1.	Microsoft Office 2010	Microsoft	лицензионное
2.	Windows 7 Pro	Microsoft	лицензионное
3.	Microsoft Office 2013	Microsoft	лицензионное
4.	Windows 10 Pro	Microsoft	лицензионное
5.	Kaspersky Endpoint Security	Kaspersky	лицензионное
6.	Microsoft Office 2016	Microsoft	лицензионное

Перечень БД и ИСС

№п /п	Наименование
1	Международные реферативные наукометрические БД, доступные в рамках национальной подписки Web of Science Scopus
2	Профессиональные полнотекстовые БД, доступные в рамках национальной подписки Журналы Cambridge University Press ProQuest Dissertation & Theses Global SAGE Journals Журналы Taylor and Francis
3	Профессиональные полнотекстовые БД JSTOR Издания по общественным и гуманитарным наукам

Состав программного обеспечения с реквизитами документов

Microsoft Office 2010, договор №17/03 от 21.03.2017 с АО «СофтЛайнТрейд»
 Microsoft Office 2013, договор №16 от 13.06.17 с ООО «Софтлайн Проекты»
 Windows 7 Pro, договор №17/03 от 21.03.2017 с АО «СофтЛайнТрейд»
 Windows 10 Pro, договор №16 от 13.06.17 с ООО «Софтлайн Проекты»
 Kaspersky Endpoint Security, договор №594-05-44 от 19.12.18 с АО «СофтЛайнТрейд»
 Microsoft Office 2016, договор №16 от 13.06.2017 с ООО «Софтлайн Проекты»

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 нарушениями¹⁹ 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, регулируемые эргономическими партами СИ-1;
 - компьютерной техникой со специальным программным обеспечением.

9. Методические материалы

9.1. Планы практических занятий

Основные темы курса:

1. Введение. Общее представление о корпусах и корпусной лингвистике. Стандарты разметки. Типы разметки корпусов.
2. Особенности различных типов разметки. Морфологическая разметка
3. Особенности различных типов разметки. Синтаксическая разметка
4. Особенности разметки: другие типы разметки
5. Методы корпусных исследований. Анализ примеров корпусных исследований

ПЛАН СЕМИНАРСКИХ ЗАНЯТИЙ И САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ

В соответствии с учебным планом предусмотрены семинарские занятия. Некоторые из них строго обязательны, а другие допускают рассмотрение той или иной темы с разной степенью подробности: разворачивание и уточнение темы или, напротив, объединение нескольких тем.

№ занятия	Тема семинара	Вопросы для подготовки к семинару и самостоятельной работы
1	Введение. Общее представление о корпусах и корпусной лингвистике.	1. Стандарты разметки. 2. Типы разметки корпусов.
2	Проблемные корпуса (параллельные, диалектные, мультимедийные и др.)	1. Параллельные корпуса. 2. Диалектные корпуса. 3. 3. Мультимедийные корпуса.
3	Поиск в корпусе.	1. Визуальные интерфейсы корпусов.

№ занятия	Тема семинара	Вопросы для подготовки к семинару и самостоятельной работы
1	Введение. Общее представление о корпусах и корпусной лингвистике.	1. Стандарты разметки. 2. Типы разметки корпусов.
	Использование языка SQR для поиска в корпусе. Составление сложных запросов к корпусу.	2. Языки запросов к корпусу.
4-6	Особенности различных типов разметки.	1. Особенности различных типов разметки. Морфологическая разметка 2. Особенности различных типов разметки. Синтаксическая разметка 3. Особенности разметки: другие типы разметки 4. Методы корпусных исследований. Анализ примеров корпусных исследований
7	Промежуточная аттестация: Контрольная работа по теме: «Корпуса и лингвистические ресурсы»	Обобщение пройденного материала
8	Инструменты разметки собственного исследовательского корпуса	1. Система SketchEngine. 2. Программа WebBootCaT.
9-10	Составление конкордансов, частотных списков, списков коллокаций с использованием специальных программ	1. Анализ корпусов с помощью AntConc. 2. Система SketchEngine для анализа корпусов.

9.2. Методические рекомендации по подготовке письменных работ

Самостоятельная работа студентов предполагает:

- повторение материала лекций;
- подготовку письменных и устных домашних заданий;
- внеаудиторную работу студентов (самостоятельное освоение теоретического материала, конспектирование научных статей и монографий, написание рефератов, консультирование в процессе написания рефератов с преподавателем, подготовка докладов, алгоритмов и презентаций, разработка проектов, подготовка к текущему и итоговому контролю).

Самостоятельная работа студента играет большую роль в освоении материала, поскольку она делает восприятие информации не пассивным, а активным процессом. Здесь

важны и самостоятельный поиск материала в²¹ научной литературе и на соответствующих сайтах, и его конспектирование, и при необходимости его трансформация в схемы и алгоритмы.

Общие принципы самостоятельной работы. За редким исключением, СРС нацелена не на запоминание материала, а на его понимание, осмысление, упорядочение и активное практическое владение им. Критерием адекватного понимания является способность объяснить материал своими словами непрофессионалу, умение приводить иллюстративные примеры и практически разрабатывать электронные обучающие материалы, используя предлагаемые инструменты. Необходимо уметь формулировать вопросы и находить ответы на них самостоятельно, в ходе консультации с преподавателем и другими членами группы. Только после этого нужно приступить к выполнению задания.

9.3. Иные материалы

Терминологический словарь-минимум

Baker, Paul, Andrew Hardie & Tony McEnery. 2006. *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.

Рекомендуемая дополнительная литература:

Плунгян, Владимир А., Шестакова Лариса Л. (ред.). 2013. *Корпусный анализ русского стиха : Сборник научных статей*. М.: Издательский центр «Азбуковник».

Aijmer, Karin & Christoph Rühlemann. 2014. *Corpus pragmatics: A handbook*. Cambridge: Cambridge University Press.

Baker, Paul. 2010. *Sociolinguistics and corpus linguistics*. Edinburgh: Edinburgh University Press.

Cheng, Winnie. 2012. *Exploring corpus linguistics: Language in action*. London & New York: Routledge.

Рекомендованная

Киселёва, Ксения Л., Владимир А. Плунгян, Екатерина В. Рахилина, Сергей Г. Татевосов (ред.). *Корпусные исследования по русской грамматике*. М.: Пробел–2009.

Национальный корпус русского языка: 2003—2005. Сборник статей. М.: Индрик, 2005.

Ляшевская, Ольга Н. & Шаров, Сергей А. *Частотный словарь современного русского языка (на материалах Национального корпуса русского языка)*. М.: Азбуковник. 2009.

Грудева, Елена В. 2012. *Корпусная лингвистика*. 2-е изд. М.: ФЛИНТА.

Копотев, Михаил. 2014. *Введение в корпусную лингвистику*. Praha: Animedia Company.

Плунгян, Владимир А., Екатерина В. Рахилина, Татьяна И. Резникова (ред.). *Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы*. СПб.: Нестор-История, 2009.

Biber, Douglas & Randi Reppen (eds.). 2011. *Corpus linguistics*. 4 vols. London: Sage.

Gatto, Maristella. 2014. *The Web as corpus: Theory and practice*. New York : Bloomsbury Academic.

Gries, Stefan Th. 2009. *Quantitative corpus linguistics with R: A practical introduction*. London & New York: Routledge.

Lüdeling, Anke & Merja Kytö. 2008–2009. *Corpus linguistics: An international handbook*. 2 vols. HSK 29.1–2. Berlin & New York: Walter de Gruyter.

Meyer, Charles F. 2002. *English corpus linguistics: An introduction*. Cambridge & New York: Cambridge University Press.

McEnery, Tony & Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge & New York: Cambridge University Press.

O’Keeffe, Anne & Michael McCarthy (eds.). 2010. *The Routledge handbook of corpus linguistics*. London & New York: Routledge.

Sinclair, John. 1991. *Corpus, concordance, collocation*²². Oxford: Oxford University Press.

АННОТАЦИЯ ДИСЦИПЛИНЫ

Дисциплина «Информационные технологии в лингвистике» реализуется УНЦ компьютерной лингвистики Института лингвистики РГГУ.

Цель дисциплины – познакомить магистрантов с наиболее актуальными современными компьютерными корпусами текстов и лексикографическими ресурсами, программами обработки текста, с технологиями создания собственных исследовательских корпусов, научить применять методы создания собственных исследовательских корпусов, работы с корпусными данными, методы обработки этих данных в собственных научных исследованиях.

Для достижения установленной цели решаются следующие учебные **задачи**:

- познакомить магистрантов с последними трендами в области корпусной лингвистики, с основными типами лингвистических ресурсов, доступными в Интернете и используемыми в профессиональной деятельности: с наиболее актуальными лингвистическими корпусами (национальными и проблемными), специальными программами обработки текстов, лексикографическими ресурсами и т.п.;
- на примерах показать, какие новые возможности в исследовании грамматики и лексики языка дает использование корпусных, а также применение современных методов обработки этих данных;
- дать представления о проблемах создания корпусов, об основных принципах разработки данных ресурсов и об основных требованиях, предъявляемых к ним;
- познакомить с технологиями и проблемами разметки корпусов;
- научить работать с современными пакетами обработки собственных корпусов;
- научить применять специальные методы лингвистических исследований, использующие данные корпусов, в том числе и статистические методы исследования;
- обучить практическим навыкам по применению корпусных методов в своей исследовательской работе.

1.2. Формируемые компетенции, соотнесённые с планируемыми результатами обучения по дисциплине:

ПК-29 готовность работать с основными информационно-поисковыми и экспертными системами, системами представления знаний, синтаксического и морфологического анализа, автоматического синтеза, распознавания и понимания речи, обработки лексикографической информации и автоматизированного перевода, автоматизированными системами идентификации и верификации личности,

ПК-31 владение современными методиками разработки лингвистического обеспечения в автоматизированных системах различного профиля.

Знать:

- основные принципы создания корпусов и других компьютерных лингвистических ресурсов;
- характеристики и особенности современных доступных в Интернете национальных и проблемных корпусов, широко используемых в лингвистических исследованиях, включая недавно вошедшие в лингвистическую практику;

- стандарты, типы и проблемы разметки корпусов²⁴, включая такие современные типы разметки, как дискурсивную разметку, интонационную разметку устных корпусов и т.п., применяемые в разметке технологии;
- принципы создания собственных исследовательских корпусов;
- основные типы исследовательских задач, решаемых с использованием корпусов;
- основные применяемые в корпусных исследованиях лексики и грамматики методы
- требования, предъявляемые к верификации результатов
- основные методы статистического анализа корпусных данных.

Уметь:

- применять полученные знания в области корпусной лингвистики в научно-исследовательской и других видах практической деятельности;
- осуществлять мониторинг и оценку различных типов современных корпусных ресурсов и выбирать ресурсы, подходящие для выполнения тех или иных исследовательских и производственных задач;
- осуществлять поиск в корпусах в соответствии с исследовательской гипотезой в области грамматики и лексикографических исследований;
- создавать и размечать собственные исследовательские и обучающие корпуса;
- работать с различными типами программ обработки текстов: конкордансерами, программами для поиска коллокаций, создания частотных списков и т.п., корпусными менеджерами;
- разрабатывать методический материал по основным языковым дисциплинам с использованием корпусов.

Владеть:

- основными методами и средствами профессионального компьютерного инструментария для исследовательской и практической работы;
- методами сбора материала с использованием корпусов;
- методами анализа корпусных данных, включая статистические методы.

Программой дисциплины предусмотрена промежуточная аттестация в форме зачёта. Общая трудоемкость освоения дисциплины составляет 2 зачетные единицы.

ЛИСТ ИЗМЕНЕНИЙ

№	Текст актуализации или прилагаемый к РПД документ, содержащий изменения	Дата	№ протокола
1	Приложение №1	31.08.2020 г.	1

1. Структура дисциплины (к п. 2 РПД на 2020)

Структура дисциплины для очной формы обучения

Общая трудоёмкость дисциплины составляет 2 з.е., 76 ч., в том числе контактная работа обучающихся с преподавателем 20 ч., самостоятельная работа обучающихся 56 ч.

№ п/ п	Раздел Дисциплины	Семестр	Неделя семестра	Виды учебной работы, включая самостоятельную работу студентов и трудоёмкость (в часах)				Формы текущего контроля успеваемости (по неделям семестра) Форма промежуточной аттестации (по семестрам)
				лек- ции	практич. заня- тия	самос- стоя- тель- ная рабо- та		
11.	Введение. Общее представление о корпусах и корпусной лингвистике. Стандарты разметки. Типы разметки корпусов.	3	1		2	4		Д31. Упр. по теме “Особенности поиска и управления выдачи в НКРЯ”.
12.	Проблемные корпуса (параллельные, диалектные, мультимедийные и др.)	3	2		2	4		Д32. Практическая работа. Разметка мультимедийного / аудиокорпуса
13.	Поиск в корпусе. Использование языка SQP для поиска в корпусе. Составление сложных запросов к корпусу.	3	3		2	6		Д33. Запросы для поиска сложных конструкций с использованием языка регулярных выражений: письменный отчет
14.	Особенности различных типов разметки. Морфологическая разметка	3	4		2	6		Д34. Практическая работа. Тестирование морфологической или синтаксической разметки: письменный отчет
15.	Особенности различных типов разметки. Синтаксическая разметка	3	5		2	2		Д35. Практическая работа. Тестирование морфологической или синтаксической разметки: письменный отчет

16.	Особенности разметки: другие типы разметки	3	6		2	6	ДЗ6. Практическая работа. Семантическая разметка / анафорическая разметка. Контрольная работа по теме: «Корпуса и лингвистические ресурсы»
17.	Методы корпусных исследований. Анализ примеров корпусных исследований	3	7		2	4	ДЗ7. Реферирование статьи из рекомендованного списка. Составление краткого ТЗ к собственному исследовательскому проекту.
18.	Инструменты разметки собственного исследовательского корпуса	3	8		2	4	ДЗ8. Разработка параметров и схемы разметки исследовательского корпуса, разметка корпуса с помощью специальной программы
19.	Составление конкордансов, частотных списков, списков коллокаций с использованием специальных программ	3	9		2	4	ДЗ9. Обработка данных корпуса с использованием специальных программ
20.	Итоговая аттестация				2	16	Защита исследовательского проекта
	.				20	56	

2. Образовательные технологии (к п.4 на 2020 г.)

В период временного приостановления посещения обучающимися помещений и территории РГГУ. для организации учебного процесса с применением электронного обучения и дистанционных образовательных технологий могут быть использованы следующие образовательные технологии:

- видео-лекции;
- онлайн-лекции в режиме реального времени;
- электронные учебники, учебные пособия, научные издания в электронном виде и доступ к иным электронным образовательным ресурсам;
- системы для электронного тестирования;
- консультации с использованием телекоммуникационных средств.

3. Состав программного обеспечения (ПО) (к п. 7 на 2020 г.)

№п /п	Наименование ПО	Производитель	Способ распространения (лицензионное или
-------	-----------------	---------------	--

			<i>свободно распространяемое)</i>
7.	Microsoft Office 2010	Microsoft	лицензионное
8.	Windows 7 Pro	Microsoft	лицензионное
9.	Microsoft Office 2013	Microsoft	лицензионное
10.	Windows 10 Pro	Microsoft	лицензионное
11.	Kaspersky Endpoint Security	Kaspersky	лицензионное
12.	Microsoft Office 2016	Microsoft	лицензионное
7.	Zoom	Zoom	лицензионное