МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Российский государственный гуманитарный университет»
(ФГБОУ ВО «РГГУ»)

ИНСТИТУТ ЛИНГВИСТИКИ

УНЦ компьютерной лингвистики

КОМПЬЮТЕРНАЯ И КОРПУСНАЯ ЛИНГВИСТИКА

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

45.03.03 Фундаментальная и прикладная лингвистика

Код и наименование направления подготовки/специальности

Фундаментальная и прикладная лингвистика

Наименование направленности (профиля)/ специализации

Уровень высшего образования: *бакалавриат*

Форма обучения: *Очная*

РПД адаптирована для лиц
с ограниченными возможностями
здоровья и инвалидов

Москва 2022

Компьютерная и корпусная лингвистика

Рабочая программа дисциплины

Составитель:

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

УТВЕРЖДЕНО

Протокол заседания УНЦ компьютерной лингвистики

№_5__ от_____31.03.2022_____

ОГЛАВЛЕНИЕ

1. Пояснительная записка

1.1. Цель и задачи дисциплины

Цель дисциплины — освоение студентами базовых понятий компьютерной и корпусной лингвистики, овладение базовыми принципами автоматической обработки естественного языка.

Задачи

- познакомить студентов с основными задачами, стоящими в современной компьютерной лингвистике, и с методами их решения;

- указать на связь между содержательными характеристиками языковых явлений и способами их автоматической обработки при решении практических задач компьютерной лингвистики;

- научить студентов пользоваться базовыми программными продуктами, разработанными в компьютерной лингвистике, знать их области применения;

- обучить студентов основным понятиям и методам современной корпусной лингвистики;

- познакомить студентов с ключевыми проблемами ручной и автоматической разметки корпусных данных;

- научить студентов пользоваться существующими корпусами, понимать различие в интерфейсах поисковых запросов и форматов выдачи, обоснованно выбирать корпусной ресурс под решение конкретной исследовательской задачи.

В результате освоения дисциплины обучающийся должен:

знать:

- основные понятия и методы современной компьютерной лингвистики
- базовые принципы лингвистической разметки
- основные понятия и методы корпусной лингвистики
- устройство корпусов

уметь:

- анализировать различные уровни языковой структуры
- решать конкретные компьютерно-лингвистические задачи
- анализировать различные уровни языковой структуры
- решать лингвистические задачи с помощью методов корпусной лингвистики

владеть:

- современной терминологией компьютерной лингвистики
- методами решения компьютерно-лингвистических задач
- современной терминологией корпусной лингвистики

методами решения лингвистических задач

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

Компетенция (код и наименование)	**Индикаторы компетенций** (код и наименование)	**Результаты обучения**
ПК-10. Способен пользоваться лингвистически ориентированными программными продуктами	ПК-10.1	Знает: основные типы систем, использующих модули лингвистического анализа; основные принципы и методы компьютерного моделирования лингвистических задач.
	ПК-10.2	Умеет: анализировать работу различных систем обработки текста и звучащей речи для выявления основных лингвистических компонентов и основных типов обработки текста, используемых в данных системах; подбирать необходимые лингвистические ресурсы для различных задач лингвистического обеспечения систем (например, лексикографических, задач морфологического анализа и т.п.).
	ПК-10.3	Имеет практический опыт работы с различными системами автоматической и экспертной обработки текста и звучащей речи.
ПК-11. Владеет принципами создания электронных языковых ресурсов (текстовых, речевых и мультимодальных корпусов; словарей, тезаурусов, онтологий; фонетических, лексических, грамматических и иных баз данных и баз знаний) и умеет пользоваться такими ресурсами	ПК-11.1	Знает: основные принципы обработки информации; базовые принципы корпусной лингвистики, лексикографии, математической статистики; базовые представления о языковом разнообразии; наиболее полные и значимые лингвистические корпуса, электронные словари и базы данных.
	ПК-11.2	Умеет: пользоваться основными методами, способами и средствами получения, хранения, переработки информации; пользоваться лингвистически ориентированными программными продуктами.
	ПК-11.3	Имеет практический опыт разработки электронных языковых ресурсов; опыт применения основных методов, способов и

		средств получения, хранения, переработки информации.
ПК-12. Способен использовать лингвистические технологии для проектирования систем автоматической обработки звучащей речи и письменного текста на естественном языке, лингвистических компонентов интеллектуальных и информационных электронных систем	ПК-12.1	Знает: основные системы автоматической обработки звучащей речи и текстов на естественном языке; базовые принципы автоматической обработки языковых данных; основные интеллектуальные и информационные электронные системы и принципы работы с ними.
	ПК-12.2	Умеет: пользоваться существующими системами автоматической обработки текста и звучащей речи, интеллектуальными и информационными электронными системами; проводить их сравнительный анализ; проектировать модули данных систем, составлять технические задания.
	ПК-12.3	Имеет практический опыт работы с системами автоматической обработки текста и звучащей речи; проектирования модулей таких систем.
ПК-13. Способен проводить квалифицированное тестирование лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем	ПК-13.1	Знает: типы, характеристики и особенности основных доступных в Интернете лингвистических ресурсов.
	ПК-13.2	Умеет: сравнивать данные, полученные с использованием различных электронных лингвистических ресурсов и систем; применять методы математического анализа и моделирования в профессиональной деятельности.
	ПК-13.3	Имеет практический опыт тестирования электронных лингвистических ресурсов, систем и компонентов.

1.3. Место дисциплины в структуре образовательной программы

Дисциплина «Компьютерная и корпусная лингвистика» относится к части, формируемой участниками образовательных отношений, блока дисциплин учебного плана. Дисциплина состоит из двух частей: вначале (семестр 4) студентам читается часть, посвященная компьютерной лингвистике, затем (семестр 6) — часть, посвященная корпусной лингвистике.

Для освоения дисциплины необходимы знания, умения и владения, сформированные в ходе изучения следующих дисциплин и прохождения практик: «Понятийный аппарат математики», «Вероятностные модели», «Введение в теорию языка», «Общая морфология», «Общий синтаксис».

В результате освоения дисциплины формируются знания, умения и владения, необходимые для изучения следующих дисциплин и прохождения практик: «Программирование в лингвистике», «Автоматический перевод».

2. Структура дисциплины

Общая трудоёмкость дисциплины составляет 5 з.е., 180 академических часов.

Структура дисциплины для очной формы обучения

Объем дисциплины в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении учебных занятий:

Семестр	Тип учебных занятий	Количество часов
	Лекции	22
	Семинары	48
	Всего:	70

Объем дисциплины (модуля) в форме самостоятельной работы обучающихся составляет 110 академических часов.

3. Содержание дисциплины

№	Наименование раздела дисциплины	Содержание
	I. КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА	
I.1	Компьютерная лингвистика и автоматическая обработка естественного языка	Компьютерная лингвистика и автоматическая обработка естественного языка. Лингвистический и инженерный подход к компьютерной лингвистике. Задачи компьютерной лингвистики. Сложности при обработке естественного языка: омонимия, синонимия, проблемы с пониманием прагматики и т. п.
I.2	Сегментация текста на токены и предложения. Проблемы токенизации	Сегментация текста на токены (≈ слова) и предложения. Проблемы токенизации и деления на предложения в языках с различными системами графики. Образец простейшего токенизатора с использованием регулярных выражений.

I.3	*n*-граммные языковые модели. Сглаживание	n-граммные языковые модели. Оценка вероятности для последовательности слов. Оценка n-граммных моделей. Перплексивность. Сглаживание: метод Лапласа, интерполяция и откат.
I.4	Стемминг, лемматизация и морфологическая разметка	Понятия стемминга, лемматизации, частеречная разметка и морфологическая разметка. Стандарты морфологической разметки для русского и английского языка. Омонимия и её разрешение. Скрытые марковские модели. Алгоритм Витерби. Таггер Брилла.
I.5	Формальное представление синтаксиса. Основные алгоритмы парсинга	Формальное представление синтаксиса: структура зависимостей и структура составляющих. Синтаксически аннотированные корпуса. Типология формальных грамматик. Основные алгоритмы парсинга. Stanford Parser, MaltParser.
I.6	Решение конкретных компьютерно-лингвистических задач	Оценка качества в компьютерной лингвистике. Автоматическая проверка орфографии. Машинный перевод. Классификация и кластеризация текстов. Чат-боты. Информационный поиск.
	II. КОРПУСНАЯ ЛИНГВИСТИКА	
II.1	Основные методы лингвистического исследования и место корпусной лингвистики среди них	Интроспекция, эксперимент и наблюдение над реальностью. Место корпусной лингвистики в этом противопоставлении. Критика корпусной лингвистики со стороны Н. Хомского.
II.2	Лингвистические корпуса: определение и примеры применения в лингвистических исследованиях	Корпус как совокупность текстов, разметки и поиска. Применение корпусных методов для исследования морфологии, синтаксиса и семантики. Исследование
II.3	Корпуса русского и английского языков (обзор)	Основные корпуса русского и английского языков: Национальный корпус русского языка, ГИКРЯ, ruTenTen, Araneum Russicum; Brown Corpus, BNC, english-corpora.org
II.4	Типы разметки в корпусах. Стандарты морфологической разметки для русского и английского языка. Омонимия и её разрешение	Морфологическая разметка. Синтаксическая разметка. Прочие виды лингвистической разметки. Метаразметка
II.5	Количественные исследования на корпусном материале. Базовые методы статистики в корпусных исследованиях.	Таблица сопряжённости. Критерий хи-квадрат. Критерий Уилкоксона — Манна — Уитни.
II.6	Нормирование частотности языковых единиц в корпусах различного объёма. Частотные словари. Закон Ципфа.	Частотность на миллион. Зависимость частотности от ранга слова в частотном списке.

II.7	Исследование сочетаемости слов при помощи корпусов. Коллокации и меры их оценки. Лексические функции и их корпусное исследование.	Фразеологические сращения, единства и сочетания. Меры связанности коллокаций. Ожидаемая и наблюдаемая частота в корпусе. MI, z-score, t-score и др.
II.8	Проблема отбора текстов в корпус, репрезентативности и сбалансированности корпуса	Понятие сбалансированности и репрезентативности корпуса. Масштабируемость корпусных исследований.
II.9	Создание пользовательских корпусов. Применение корпусных методов в различных областях лингвистики	Оффлайн- и онлайн-конкордансеры. Создание специальных корпусов для различных исследовательских задач.

4. Образовательные технологии

№ п/п	**Наименование раздела**	**Виды учебных занятий**	**Образовательные технологии**
I	**КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА**		
I.1	Компьютерная лингвистика и автоматическая обработка естественного языка	Лекция 1	Вводная лекция с использованием презентации.
I.2	Сегментация текста на токены и предложения. Проблемы токенизации	Лекция 2	Лекция с использованием презентации.
		Семинар 1	Развёрнутое обсуждение прочитанной научной литературы. Совместное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам.
I.3	n-граммные языковые модели. Сглаживание	Лекция 3	Лекция с использованием презентации.
		Семинар 2	Развёрнутое обсуждение прочитанной научной литературы. Совместное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам.
		Семинар 3	Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч.

			текстовым корпусам, и совместное обсуждение после.
I.4	Стемминг, лемматизация и морфологическая разметка	Лекция 4	Лекция с использованием презентации.
		Семинар 4	Развёрнутое обсуждение прочитанной научной литературы. Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
I.5	Формальное представление синтаксиса. Основные алгоритмы парсинга	Лекция 5	Лекция с использованием презентации.
		Семинар 5	Развёрнутое обсуждение прочитанной научной литературы. Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
I.6	Решение конкретных компьютерно-лингвистических задач	Лекция 6	Лекция с использованием презентации.
		Семинар 6	Обсуждение существующих компьютерно-лингвистических задач (оценка качества в компьютерной лингвистике, информационный поиск) и способов их решения. Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
		Семинар 7	Обсуждение существующих компьютерно-лингвистических задач (автоматическая проверка орфографии, машинный перевод) и способов их решения. Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
		Семинар 8	Обсуждение существующих компьютерно-лингвистических задач (классификация и кластеризация текстов, чат-боты) и способов их решения. Индивидуальное выполнение

			заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
I.7	Зачёт		Защита и обсуждение докладов на ранее заданные темы.
II	**КОРПУСНАЯ ЛИНВГИСТИКА**		
II.1	Основные методы лингвистического исследования и место корпусной лингвистики среди них	Лекция 1.	Вводная лекция с использованием презентации
II.2	Лингвистические корпуса: определение и примеры применения в лингвистических исследованиях	Лекция 2	Лекция с использованием презентации
		Семинар 1	Развёрнутое обсуждение прочитанной научной литературы. Совместное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам.
II.3	Корпуса русского и английского языков (обзор)	Лекция 3	Лекция с использованием презентации
		Лекция 4	Лекция с использованием презентации
		Семинар 2	Совместное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам.
		Семинар 3	Совместное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам.
		Семинар 4	Совместное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам.
		Семинар 5	Совместное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и

			электронным ресурсам, в т.ч. текстовым корпусам.
II.4	Типы разметки в корпусах. Стандарты морфологической разметки для русского и английского языка. Омонимия и её разрешение	Лекция 5	Лекция с использованием презентации
		Семинар 6	Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
		Семинар 7	Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
II.5	Количественные исследования на корпусном материале. Базовые методы статистики в корпусных исследованиях.	Семинар 8	Развёрнутое обсуждение прочитанной научной литературы. Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
II.6	Нормирование частотности языковых единиц в корпусах различного объёма. Частотные словари. Закон Ципфа.	Семинар 9	Развёрнутое обсуждение прочитанной научной литературы. Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
II.7	Исследование сочетаемости слов при помощи корпусов. Коллокации и меры их оценки. Лексические функции и их корпусное исследование.	Семинар 10	Развёрнутое обсуждение прочитанной научной литературы. Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
II.8	Проблема отбора текстов в корпус, репрезентативности и	Семинар 11	Развёрнутое обсуждение прочитанной научной литературы. Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным

	сбалансированности корпуса		ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
		Семинар 12	Развёрнутое обсуждение прочитанной научной литературы. Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
		Семинар 13	Развёрнутое обсуждение прочитанной научной литературы. Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
II.9	Создание пользовательских корпусов. Применение корпусных методов в различных областях лингвистики	Семинар 14	Развёрнутое обсуждение прочитанной научной литературы. Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
		Семинар 15	Развёрнутое обсуждение прочитанной научной литературы. Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
		Семинар 16	Развёрнутое обсуждение прочитанной научной литературы. Индивидуальное выполнение заданий с помощью компьютера с доступом к сети «Интернет» и электронным ресурсам, в т.ч. текстовым корпусам, и совместное обсуждение после.
II.10	Экзамен		Защита и обсуждение докладов на ранее заданные темы.

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.

5. Оценка планируемых результатов обучения

5.1 Система оценивания

Форма контроля	**Макс. количество баллов**	
	За одну работу	**Всего**
I. Семестр 4: блок «Компьютерная лингвистика»		
Текущий контроль:		
- участие в дискуссиях в ходе лекций (в т.ч. в обсуждении прочитанной литературы)	*3 балла*	*12 баллов*
- выполнение заданий в ходе семинаров (темы 2-6)	*8 баллов*	*48 баллов*
Промежуточная аттестация – *зачет*		40 баллов
Итого за семестр		100 баллов
II. Семестр 6: блок «Корпусная лингвистика»		
Текущий контроль:		
- практические упражнения и задания, выполняемые на занятиях и в качестве домашней работы	*5 баллов*	*30 баллов*
- доклады студентов по прочитанной литературе	*5 баллов*	*30 баллов*
Промежуточная аттестация – *экзамен*		40 баллов
Итого за семестр		100 баллов

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 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

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	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 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	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	удовлетво-рительно/ 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	неудовлет-ворительно/ 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 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

I. Блок «Компьютерная лингвистика»

Примеры заданий к семинарам

1. Тагсет Penn: найти ошибки

- I/PRP need/VBP a/DT flight/NN from/IN Atlanta/NN
- Does/VBZ this/DT flight/NN serve/VB dinner/NNS
- I/PRP have/VB a/DT friend/NN living/VBG in/IN Denver/NNP
- What/WDT flights/NNS do/VBP you/PRP have/VB from/IN Milwaukee/NNP to/IN Tampa/NNP
- Can/VBP you/PRP list/VB the/DT nonstop/JJ afternoon/NN flights/NNS

2. Тагсет Penn: разметить

- It is a nice night.

3. Построить регулярные выражения, которые будут находить в тексте:

- IPv4 адреса
- телефонные номера
- инициалы и фамилии (А.С. Пушкин)

4. Оценить перплексивность униграммной, биграммной и триграммной модели, обученной на Национальном корпусе русского языка.

Примеры статей для докладов

- Kuhn, Tobias. 2014. A survey and classification of controlled natural languages. *Computational Linguistics* 40.1: 121–170.
- Kernighan, Mark, Kenneth Church & William Gale. 1990. A spelling correction program based on a noisy channel model.
- Johannes Schaback, Fang Li. 2007. Multi-level feature extraction for spelling correction.
- Hobbs, Jerry R. & Ellen Riloff. 2010. Information extraction. In: Indurkhya & Damerau (2010) (eds.). 511–532.
- Liu, Bing. 2010. Sentiment analysis and subjectivity. In: Indurkhya & Damerau (2010) (eds.). 627–666.
- Knight, Kevin & Graehl Jonathan. 1999. Machine transliteration. *Computational Linguistics* 24.4: 599–612.

II. Блок «Корпусная лингвистика»

Примеры заданий к семинарам

1. Суммарная Average Reduced Frequency (ARF) для всех слов в корпусе …, чем суммарная обычная частотность всех слов в корпусе.

(А) меньше; (Б) не меньше; (В) больше; (Г) не больше

2. Какую из этих мер рекомендуется применять только с установлением нижнего порога частотности для коллокации?

(А) MI; (Б) t-мера; (В) z-мера; (Г) simple-ll

3. Мера Average Reduced Frequency (ARF) теоретически может быть равна обычной частоте для слова, которое встречается в корпусе с частотой …

(А) 1; (Б) 2; (В) 100; (Г) с любой частотой

4. Сколько степеней свободы имеет таблица 3×3 с фиксированными суммами по строкам и по столбцам?

(А) 0; (Б) 1; (В) 4; (Г) 9

5. Какое сочетание слов в русском языке будет предсказано униграммной моделью как наиболее частотное?

(А) и и; (Б) и не; (В) а не; (Г) не а

6. Какое из этих значений не является стандартным порогом значимости, применяемым в статистических исследованиях?

(А) 0,2; (Б) 0,05; (В) 0,01; (Г) 0,001

7. Корпуса с известной степенью сходства (Known-Similarity Corpora) используются для …

(А) выявления ключевых слов; (Б) оценки лексического разнообразия; (В) оценки мер сравнения корпусов; (Г) машинного перевода

8. Если сравнивать частоту двух слов в разных корпусах, какие результаты даёт критерий χ^2?

(А) для высокочастотных слов результат почти всегда значим, а для низкочастотных почти всегда незначим; (Б) для высокочастотных слов результат обычно менее значим, чем для низкочастотных; (В) для высокочастотных слов результат почти всегда значим, и только для низкочастотных слов результат действительно зависит от наличия различий между корпусами; (Г) результат по критерию χ^2 не зависит от частоты слова

9. Если увеличить все значения в выборке в два раза, что произойдёт со стандартным отклонением?

(А) не изменится; (Б) уменьшится; (В) увеличится; (Г) невозможно определить

10. Какое минимальное значение R в частотном словаре Ляшевской и Шарова может иметь слово, которое встретилось в корпусе, легшем в основу словаря, 50 раз?

(А) 0; (Б) 1; (В) 50; (Г) 100

11. Слово w встретилось в корпусе объёмом N = 100000 токенов на 17-й, 548-й, 12319-й и 83500-й позиции. Вычислите ARF этого слова.

12. Каково минимальное и максимальное значение R для слов, которые в частотном словаре Ляшевской и Шарова имеют D=54?

13. С помощью критерия χ^2 оцените различие по частотности для слова «волна» в газетном и в поэтическом подкорпусе НКРЯ.

14. Используя корпус Araneum Russicum Minus (http://unesco.uniba.sk/guest/run.cgi/first_form), перечислите пять наиболее сильных по t-мере коллокатов, встречающихся непосредственно после слова «продать» (на уровне лемм).

15. Слово w встретилось в корпусе, разбитом на 10 частей, в общей сложности 40 раз. Приведите любое возможное распределение частотности этого слова по 10 сегментам корпуса, при котором D Жуйяна будет лежать в интервале [50;60].

16. При каких значениях n в формуле для вычисления %DIFF Адама Килгаррифа слово «system» окажется более характерным для British Academic Written English Corpus в сравнении с British Academic Spoken English Corpus, чем слово «strength»? (оба этих корпуса есть в SketchEngine)

17. В файле HP1and3.xls содержится частотный словарь для первой и третьей книги о Гарри Поттере. Приблизительно оцените, при каких значениях n в формуле для вычисления %DIFF Адама Килгарриффа в топ-10 слов, наиболее характерных для первой книги на фоне третьей, попадут не меньше четырёх нарицательных существительных.

Примеры статей для докладов

Из сборников:

Киселёва, Ксения Л., Владимир А. Плунгян, Екатерина В. Рахилина, Сергей Г. Татевосов (ред.). 2009. Корпусные исследования по русской грамматике. М: Пробел–2000.

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

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

6. Учебно-методическое и информационное обеспечение дисциплины

6.1 Список источников и литературы

Основная литература

- Захаров, В. П. Корпусная лингвистика : учебник / В. П. Захаров, С. Ю. Богданова. - 3-е изд., перераб. - Санкт-Петербург : СПбГУ, 2020. - 234 с. - ISBN 978-5-288-05997-1. - Текст : электронный. - URL: https://znanium.com/catalog/product/1244746

- Шунейко, А. А. Квантитативная лингвистика и новые информационные технологии : учебник для вузов / А. А. Шунейко, И. А. Авдеенко. — Москва : Издательство Юрайт, 2024. — 347 с. — (Высшее образование). — ISBN 978-5-534-15446-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/543983

- Шунейко, А. А. Корпусная лингвистика : учебник для вузов / А. А. Шунейко. — Москва : Издательство Юрайт, 2024. — 222 с. — (Высшее образование). — ISBN 978-5-534-13603-6. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/543746

- Jurafsky, Dan & James H. Martin. 2017. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* 3nd edition draft. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Дополнительная 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

- Моделирование в корпусной лингвистике: специализированные корпусы русского языка : монография / В. П. Захаров, И. В. Азаров, О. А. Митрофанова [и др.]. - СПб : Изд-во С.-Петерб. ун-та, 2019. - 208 с. - ISBN 978-5-288-05902-5. - Текст : электронный. - URL: https://znanium.com/catalog/product/1080953

6.2 Перечень ресурсов информационно-телекоммуникационной сети «Интернет».

- Национальный корпус русского языка (НКРЯ): http://ruscorpora.ru/
- Гайдлайны по стандартизованному представлению текстов разных видов https://tei-c.org/Guidelines/P5/
- Regular Expression Cheat Sheet https://www.cheatography.com/davechild/cheat sheets/regular expressions
- Universal Dependencies http://universaldependencies.org/
- Syntactic treebanks https://en.wikipedia.org/wiki/Treebank#Syntactic_treebanks

7. Материально-техническое обеспечение дисциплины

Лекционные занятия по проводятся с использованием компьютерных презентаций, поэтому в аудитории необходимы компьютер и проектор, а также соответствующее освещение. В ходе семинарских занятий студенты должны быть обеспечены компьютерами с лицензионным программным обеспечением и выходом в Интернет, т.е. занятия должны проходить в компьютерных классах.

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. Методические материалы

9.1 Планы семинарских занятий (блок «Компьютерная лингвистика»)

Семинар 1. Сегментация текста на токены и предложения. Проблемы токенизации

Вопросы для обсуждения:

Сегментация текста на токены (≈ слова) и предложения. Проблемы токенизации и деления на предложения в языках с различными системами графики. Образец простейшего токенизатора с использованием регулярных выражений.

Семинары 2-3. n-граммные языковые модели. Сглаживание

Вопросы для обсуждения:

n-граммные языковые модели. Оценка вероятности для последовательности слов. Оценка n-граммных моделей. Перплексивность. Сглаживание: метод Лапласа, интерполяция и откат.

Семинар 4. Стемминг, лемматизация и морфологическая разметка

Вопросы для обсуждения:

Понятия стемминга, лемматизации, частеречная разметка и морфологическая разметка. Стандарты морфологической разметки для русского и английского языка. Омонимия и её разрешение. Скрытые марковские модели. Алгоритм Витерби. Таггер Брилла..

Семинар 5. Формальное представление синтаксиса. Основные алгоритмы парсинга

Вопросы для обсуждения:

Формальное представление синтаксиса: структура зависимостей и структура составляющих. Синтаксически аннотированные корпуса. Типология формальных грамматик. Основные алгоритмы парсинга. Stanford Parser, MaltParser.

Семинары 6-8. Решение конкретных компьютерно-лингвистических задач

Вопросы для обсуждения:

Оценка качества в компьютерной лингвистике. Автоматическая проверка орфографии. Машинный перевод. Классификация и кластеризация текстов. Чат-боты. Информационный поиск.

Планы семинарских занятий (блок «Корпусная лингвистика»)

Семинар 1. Лингвистические корпуса: определение и примеры применения в лингвистических исследованиях

Вопросы для обсуждения:

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

Лингвистические корпуса: определение и примеры применения в лингвистических исследованиях.

Семинары 2–5. Корпуса русского и английского языков (обзор)

Вопросы для обсуждения:

Особенности устройства различных корпусов русского и английского языков. Практическая работа с корпусами.

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)
8. British National Corpus (BNC)
9. Corpus of Contemporary American English (COCA)
10. Corpus of Global Web-Based English (GloWbe)
11. Brown Corpus
12. Google Books: Google Ngrams Viewer и поисковый интерфейс на сайте Brigham Young University

Семинары 6–7. Типы разметки в корпусах. Стандарты морфологической разметки для русского и английского языка. Омонимия и её разрешение (обзор)

Вопросы для обсуждения:

Стандарты морфологической разметки для русского и английского языка. Омонимия и её разрешение. Основные морфоанализаторы и синтаксические парсеры для русского и английского языка.

Семинар 8. Количественные исследования на корпусном материале. Базовые методы статистики в корпусных исследованиях

Вопросы для обсуждения:

Противопоставление качественных и количественных исследований. Анализ частотности с помощью корпусов и преимущества данного метода по сравнению с другими способами получения количественных данных о языковых единицах.

Семинар 9. Нормирование частотности языковых единиц в корпусах различного объёма. Частотные словари. Закон Ципфа

Вопросы для обсуждения:

Проценты, промилле и вхождения на миллион. Сопоставление корпусов разных объёмов. Основные частотные словари русского языка. Особенности различных мер частотности, применяемых в частотных словарях. Зависимость частотности языковых единиц от их места в частотном списка. Проблема «длинного хвоста» — многочисленных редко встречающихся слов.

Семинар 10. Исследование сочетаемости слов при помощи корпусов. Коллокации и меры их оценки. Лексические функции и их корпусное исследование

Вопросы для обсуждения:

Фразеологические сращения, единства и сочетания. Меры связанности коллокаций. Ожидаемая и наблюдаемая частота в корпусе. MI, z-score, t-score, logDice и другие меры. Содержательная интерпретация автоматически извлечённых коллокаций.

Семинары 11–13. Проблема отбора текстов в корпус, репрезентативности и сбалансированности корпуса

Вопросы для обсуждения:

Понятие сбалансированности и репрезентативности корпуса. Масшатбируемость корпусных исследований. Оценка основных корпусов русского и английского языков с точки зрения репрезентативности и сбалансированности.

Семинары 14–16. Создание пользовательских корпусов. Применение корпусных методов в различных областях лингвистики

Вопросы для обсуждения:

Оффлайн- и онлайн-конкордансеры. Создание специальных корпусов для различных исследовательских задач. Использование базовых навыков программирования при разработке собственных корпусов.

9.2 Иные материалы

Рекомендуемая литература для более глубокого освоения программы.

- Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). 2016. Прикладная и компьютерная лингвистика. М.: URSS.
- Indurkhya, Nitin & Fred J. Damerau (eds.). 2010. *Handbook of Natural Language Processing*. 2nd ed. Boca Raton: Chapman & Hall/CRC.

Приложение 1. Аннотация
рабочей программы дисциплины

АННОТАЦИЯ РАБОЧЕЙ ПРОГРАММЫ ДИСЦИПЛИНЫ

Дисциплина реализуется в УНЦ компьютерной лингвистики Института лингвистики.

Цель дисциплины — освоение студентами базовых понятий компьютерной и корпусной лингвистики, овладение базовыми принципами автоматической обработки естественного языка.

Задачи

- познакомить студентов с основными задачами, стоящими в современной компьютерной лингвистике, и с методами их решения;

- указать на связь между содержательными характеристиками языковых явлений и способами их автоматической обработки при решении практических задач компьютерной лингвистики;

- научить студентов пользоваться базовыми программными продуктами, разработанными в компьютерной лингвистике, знать их области применения;

- обучить студентов основным понятиям и методам современной корпусной лингвистики;

- познакомить студентов с ключевыми проблемами ручной и автоматической разметки корпусных данных;

- научить студентов пользоваться существующими корпусами, понимать различие в интерфейсах поисковых запросов и форматов выдачи, обоснованно выбирать корпусной ресурс под решение конкретной исследовательской задачи.

Дисциплина направлена на формирование следующих компетенций:

- ПК-10. Способен пользоваться лингвистически ориентированными программными продуктами
- ПК-11. Владеет принципами создания электронных языковых ресурсов (текстовых, речевых и мультимодальных корпусов; словарей, тезаурусов, онтологий; фонетических, лексических, грамматических и иных баз данных и баз знаний) и умеет пользоваться такими ресурсами
- ПК-12. Способен использовать лингвистические технологии для проектирования систем автоматической обработки звучащей речи и письменного текста на естественном языке, лингвистических компонентов интеллектуальных и информационных электронных систем
- ПК-13. Способен проводить квалифицированное тестирование лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем.

В результате освоения дисциплины обучающийся должен:

знать:

- основные понятия и методы современной компьютерной лингвистики
- базовые принципы лингвистической разметки

- основные понятия и методы корпусной лингвистики
- устройство корпусов

уметь:

- анализировать различные уровни языковой структуры
- решать конкретные компьютерно-лингвистические задачи
- анализировать различные уровни языковой структуры
- решать лингвистические задачи с помощью методов корпусной лингвистики

владеть:

- современной терминологией компьютерной лингвистики
- методами решения компьютерно-лингвистических задач
- современной терминологией корпусной лингвистики
- методами решения лингвистических задач

По дисциплине предусмотрена промежуточная аттестация в форме *зачета (семестр 4) и экзамена (семестр 6)*.

Общая трудоемкость освоения дисциплины составляет 5 зачетных единиц.