МИНОБРНАУКИ РОССИИ

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

ИНСТИТУТ ЛИНГВИСТИКИ
УНЦ компьютерной лингвистики

Введение в компьютерную лингвистику

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ
45.03.03 Фундаментальная и прикладная лингвистика
Фундаментальная и прикладная лингвистика
Уровень квалификации выпускника *(бакалавр)*

Форма обучения *(очная)*

РПД адаптирована для лиц
с ограниченными возможностями
здоровья и инвалидов

Москва *2019*

Введение в компьютерную лингвистику.
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 компьютерной лингвистики
№ 1 от 28.08.2019

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 оценки по дисциплине
5.3. Типовые задания, используемые для текущего контроля успеваемости и промежуточной аттестации обучающихся по дисциплине
5.3.1. Примеры заданий к семинарам
5.3.2. Примеры статьей для докладов
6. Учебно-методическое и информационное обеспечение дисциплины
6.1. Список источников и литературы
6.2. Перечень электронных ресурсов
7. Материально-техническое обеспечение дисциплины
8. Обеспечение образовательного процесса для лиц с ограниченными возможностями здоровья и инвалидов
9. Методические материалы
9.1. Планы семинарских занятий
Приложения
Приложение 1. Аннотация дисциплины
Приложение 2. Лист изменений

1. Пояснительная записка

1.1. Цель и задачи дисциплины

Цель дисциплины – освоение студентами базовых понятий компьютерной лингвистики и автоматической обработки естественного языка.

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

- способность пользоваться лингвистически ориентированными программными продуктами (ПК-10)
- владение принципами создания электронных языковых ресурсов (текстовых, речевых и мультимодальных корпусов; словарей, тезаурусов, онтологий; фонетических, лексических, грамматических и иных баз данных и баз знаний) и умение пользоваться такими ресурсами (ПК-11)
- способность 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)
- способность проводить квалифицированное тестирование лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем (ПК-13)

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

В результате освоения дисциплины обучающийся должен:
знать:

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

уметь:

- анализировать различные уровни языковой структуры
- решать конкретные компьютерно-лингвистические задачи

владеть:

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

Компетенция (код и наименование)	**Индикаторы компетенций** (код и наименование)	**Результаты обучения**
ПК-10. Способен пользоваться	ПК-10.1	Знает: основные типы систем,

лингвистически ориентированными программными продуктами		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	ПК-12.1	Знает: основные системы автоматической обработки звучащей речи и текстов на естественном языке; базовые принципы автоматической обработки языковых данных; основные интеллектуальные и

интеллектуальных и информационных электронных систем		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

Курс «Ведение в компьютерную лингвистику» относится к вариативной части блока дисциплин учебного плана по направлению подготовки 45.03.03 «Фундаментальная и прикладная лингвистика». Дисциплина адресована бакалаврам, обучающимся по направлению «Фундаментальная и прикладная лингвистика» (Б1.В.02) без направленности (профиля). Курс читается на 1-м курсе в 2-м семестре УНЦ компьютерной лингвистики ИЛ РГГУ, форма промежуточного контроля – зачёт.

Для успешного освоения материала студент должен опираться на знания, умения и навыки, полученные в рамках курсов «Введение в теорию языка», «Общая фонетика», «Понятийный аппарат математики», а также на навыки изучения научной литературы, сформированные при подготовке к другим теоретическим курсам.

Курс направлен на углубление знаний усовершенствование умений и навыков студентов в сфере общей лингвистики, созданию у студентов представления об основных методах и задачах компьютерной лингвистики и автоматической обработки текста.

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

2. Структура дисциплины

Общая трудоёмкость дисциплины составляет 2 з.е., 72 ч., в том числе контактная работа обучающихся с преподавателем 28 ч., самостоятельная работа обучающихся 44 ч.

Тематический календарный план курса

№ № раздела	Раздел курса	Семестр	Виды учебной деятельности и трудоёмкость (в часах)			Форма контроля успеваемости
			лекции	семинары	СРС	
1	Введение. Компьютерная лингвистика и автоматическая обработка естественного языка	1	2		2	Контроль посещаемости студентов.
2	Сегментация текста на токены и предложения. Проблемы токенизации	1	2	2	4	Контроль посещаемости студентов. Обсуждение прочитанной научной литературы. Коллективное построение регулярных выражений.
3	*n*-граммные языковые модели. Сглаживание	1	2	4	6	Контроль посещаемости студентов. Обсуждение прочитанной научной литературы. Коллективный анализ перплексивности n-граммных моделей.
4	Стемминг, лемматизация и морфологическая разметка	1	2	2	6	Контроль посещаемости студентов. Обсуждение прочитанной научной литературы. Обсуждение тагсетов и ошибок автоматической разметки.
5	Формальное представление синтаксиса. Основные алгоритмы парсинга	1	2	2	6	Контроль посещаемости студентов. Обсуждение прочитанной научной литературы. Обсуждение проблем синтаксической аннотации.
6	Решение	1	2	6	16	Контроль посещаемости

	конкретных компьютерно-лингвистических задач					студентов. Участие в коллективном обсуждении задач.
7	Зачёт	1			4	Защита докладов.
	Итого:		**12**	**16**	**44**	

***3.* Содержание дисциплины**

№	Наименование раздела дисциплины	Содержание
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.
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.
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.
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.
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.
6	Решение конкретных компьютерно-лингвистических задач	Оценка качества в компьютерной лингвистике. Автоматическая проверка орфографии. Машинный перевод. Классификация и кластеризация текстов. Чат-боты. Информационный поиск.

4. Образовательные технологии

№ п/п	Наименование раздела	Виды учебных занятий	Образовательные технологии
1	Компьютерная лингвистика и автоматическая обработка естественного языка	Лекция 1.	Вводная лекция с использованием презентации.
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.
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.
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.
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.
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.
	Зачёт		Защита и обсуждение докладов на ранее заданные темы.

5. Оценка планируемых результатов обучения

5.1. Система оценивания

Форма контроля	**Макс. количество баллов**	
	За одну работу	**Всего**
Текущий контроль: *- участие в дискуссиях в ходе лекций (в т.ч. в обсуждении прочитанной литературы)* *- выполнение заданий в ходе семинаров (темы 2-6)*	*3 балла* *9 баллов*	*18 баллов* *56 баллов*
Промежуточная аттестация *(зачёт)*	*26 баллов*	*26 баллов*
Итого за дисциплину		*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 (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 (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 (удовлетвори-тельно)»/	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

	«зачтено»	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. 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. Типовые задания, используемые для текущего контроля успеваемости и промежуточной аттестации обучающихся по дисциплине.

5.3.1. 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.

5.3.2 Примеры статьей 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.

6. Учебно-методическое и информационное обеспечение дисциплины

6.1. Список источников и литературы

Основной учебник

- 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

Дополнительные учебники

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

6.2. Перечень электронных ресурсов

- 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 курсу «Введение в компьютерную лингвистику» 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. Для эффективной работы также необходима доска, размер и расположение которой должны позволять всем слушателям видеть написанное.

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

Семинар 1. Сегментация текста на токены и предложения. Проблемы токенизации (2 ч.).

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.

Список литературы:

- Николаев И. С., Митренина О. В., Ландо Т.М. (ред.). 2016. *Прикладная и компьютерная лингвистика.* M.: URSS.
- Санников В. З. 2002. *Русский язык в зеркале языковой игры*. Москва: Языки славянской культуры.
- Hausser, Roland R. 2014. *Foundations of computational linguistics: human computer communication in natural language*. Third edition. Heidelberg: Springer.
- Indurkhya, Nitin & Frederick J. Damerau (eds.). 2010. *Handbook of natural language processing*. Boca Raton, FL: Chapman & Hall/CRC.
- Jurafsky, Dan & James H. Martin. 2008. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 2nd ed. Upper Saddle River, N.J: Pearson Prentice Hall.

Материально-техническое обеспечение занятия:

В ходе лабораторных работ студенты должны быть обеспечены компьютерами с лицензионным программным обеспечением и выходом в Интернет.

Семинары 2-3. n-граммные языковые модели. Сглаживание (4 ч.)

Вопросы для обсуждения:

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

Список литературы:

- Hausser, Roland R. 2014. *Foundations of computational linguistics: human computer communication in natural language*. Third edition. Heidelberg: Springer.
- Indurkhya, Nitin & Frederick J. Damerau (eds.). 2010. *Handbook of natural language processing*. Boca Raton, FL: Chapman & Hall/CRC.
- Jurafsky, Dan & James H. Martin. 2008. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 2nd ed. Upper Saddle River, N.J: Pearson Prentice Hall.

Материально-техническое обеспечение занятия:

В ходе лабораторных работ студенты должны быть обеспечены компьютерами с лицензионным программным обеспечением и выходом в Интернет.

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..

Список литературы:

- Jurafsky, Dan & James H. Martin. 2017. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* 3nd edition draft.
- Indurkhya, Nitin & Fred J. Damerau (eds.). 2010. *Handbook of Natural Language Processing*. 2nd ed. Boca Raton: Chapman & Hall/CRC.
- Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). 2016. Прикладная и компьютерная лингвистика. М.: URSS.

Материально-техническое обеспечение занятия:

В ходе лабораторных работ студенты должны быть обеспечены компьютерами с лицензионным программным обеспечением и выходом в Интернет.

Семинар 5. Формальное представление синтаксиса. Основные алгоритмы парсинга (2 ч.)

Вопросы для обсуждения:

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

Список литературы:

- Jurafsky, Dan & James H. Martin. 2017. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* 3nd edition draft.
- Indurkhya, Nitin & Fred J. Damerau (eds.). 2010. *Handbook of Natural Language Processing*. 2nd ed. Boca Raton: Chapman & Hall/CRC.
- Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). 2016. Прикладная и компьютерная лингвистика. М.: URSS.

Материально-техническое обеспечение занятия:

В ходе лабораторных работ студенты должны быть обеспечены компьютерами с лицензионным программным обеспечением и выходом в Интернет.

Семинары 6-8. Решение конкретных компьютерно-лингвистических задач (6 ч.)

Вопросы для обсуждения:

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

Список литературы:

- Jurafsky, Dan & James H. Martin. 2017. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* 3nd edition draft.
- Indurkhya, Nitin & Fred J. Damerau (eds.). 2010. *Handbook of Natural Language Processing*. 2nd ed. Boca Raton: Chapman & Hall/CRC.
- Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). 2016. Прикладная и компьютерная лингвистика. М.: URSS.

Материально-техническое обеспечение занятия:

В ходе лабораторных работ студенты должны быть обеспечены компьютерами с лицензионным программным обеспечением и выходом в Интернет.

Приложение 1

АННОТАЦИЯ ДИСЦИПЛИНЫ

Курс «Ведение в компьютерную лингвистику» относится к вариативной части блока дисциплин учебного плана по направлению подготовки 45.03.03 «Фундаментальная и прикладная лингвистика». Дисциплина адресована бакалаврам, обучающимся по направлению «Фундаментальная и прикладная лингвистика» (Б1.В.02) без направленности (профиля). Курс читается на 1-м курсе в 2-м семестре УНЦ компьютерной лингвистики ИЛ РГГУ, форма промежуточного контроля – зачёт.

Дисциплина реализуется в учебно-научном центре компьютерной лингвистики.

Цель дисциплины – освоение студентами базовых понятий компьютерной лингвистики и автоматической обработки естественного языка.

Задачи дисциплины:

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

- способность пользоваться лингвистически ориентированными программными продуктами (ПК-10)
- владение 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)
- способность 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)
- способность проводить квалифицированное тестирование лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем (ПК-13)

В результате освоения дисциплины обучающийся должен:
знать:

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

уметь:

- анализировать различные уровни языковой структуры
- решать конкретные компьютерно-лингвистические задачи

владеть:

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

По дисциплине предусмотрена промежуточная аттестация в форме зачёта.

Общая трудоемкость освоения дисциплины составляет 2 зачетных единицы.

Приложение 2

ЛИСТ ИЗМЕНЕНИЙ

№	Текст актуализации или прилагаемый к РПД документ, содержащий изменения	Дата	№ протокола
1	Приложение №1	25.06.2020	4

Приложение к листу изменений №1

1. 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.

2. Перечень БД и ИСС (к п. 6.2 на 2020 г.)

№п/п	Наименование
1	Международные реферативные наукометрические БД, доступные в рамках национальной подписки в 2020 г. Web of Science Scopus
2	Профессиональные полнотекстовые БД, доступные в рамках национальной подписки в 2020 г. 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 Электронная библиотека Grebennikon.ru
4	Компьютерные справочные правовые системы Консультант Плюс, Гарант

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

№п/п	Наименование ПО	Производитель	Способ распространения (*лицензионное или свободно распространяемое*)
1	Adobe Master Collection CS4	Adobe	лицензионное
2	Microsoft Office 2010	Microsoft	лицензионное
3	Windows 7 Pro	Microsoft	лицензионное
4	AutoCAD 2010 Student	Autodesk	свободно распространяемое
5	Archicad 21 Rus Student	Graphisoft	свободно распространяемое
6	SPSS Statisctics 22	IBM	лицензионное
7	Microsoft Share	Microsoft	лицензионное

	Point 2010		
8	SPSS Statisctics 25	IBM	лицензионное
9	Microsoft Office 2013	Microsoft	лицензионное
10	ОС «Альт Образование» 8	ООО «Базальт СПО	лицензионное
11	Microsoft Office 2013	Microsoft	лицензионное
12	Windows 10 Pro	Microsoft	лицензионное
13	Kaspersky Endpoint Security	Kaspersky	лицензионное
14	Microsoft Office 2016	Microsoft	лицензионное
15	Visual Studio 2019	Microsoft	лицензионное
16	Adobe Creative Cloud	Adobe	лицензионное
17	Zoom	Zoom	лицензионное