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

«Утверждаю»
Заведующий кафедрой ТО и ЗИ


А.А. Сирота

03.05.2023 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Б1.В.02 Автоматическая обработка естественного языка

1. Шифр и наименование направления подготовки/специальности:
45.03.03 Фундаментальная и прикладная лингвистика

2. Профиль подготовки/специализации: Экспертно-аналитическая деятельность

3. Квалификация (степень) выпускника: бакалавр

4. Форма обучения: очная

5. Кафедра, отвечающая за реализацию дисциплины: кафедра Технологий обработки и защиты информации

6. Составители программы: Гаршина Вероника Викторовна, канд.тех.наук, доцент кафедры Технологий обработки и защиты информации

7. **Рекомендована:** Научно-методическим советом ФКН, протокол № 7 от 03.05.2023 г.

8. Учебный год: 2026/2027 **Семестр(-ы):**8

9. Цели и задачи учебной дисциплины: Ознакомление с принципами построения систем обработки, автоматического анализа, распознавания и синтеза естественно-языковых текстов и звучащей речи.
Основные задачи дисциплины:

- Получение навыков по применению математических методов обработки текстовой информации
- Знакомство с принципами проектирования программных систем, ориентированных на обработку естественно-языковых текстов и звучащей речи.

10. Место учебной дисциплины в структуре ООП: дисциплинаБ1.Б.28 Технологии обработки текста и звучащей речи входит вбазовую часть ООП. Для изучения дисциплины необходимы знания, умения и компетенции, сформированные дисциплин: Б1.Б.15 Введение в теорию языка, Б1.Б.27 Общая и компьютерная лексикография,Б1.Б.26Технологии корпусной лингвистики, Б1.Б.14 Информатика и основы программирования, Б1.В.ОД.2 Проектирование баз данных, Б1.В.ОД.4 Семантический WEB, Б1.В.ДВ.3.1 Компьютерная лингвистика

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

Код	Название компетенции	Коды	Индикаторы	Планируемые результаты обучения
ПК–5	Способен пользоваться лингвистически ориентированными программными продуктами	ПК-5.1 ПК-5.2.	Осуществляет постановку задачи на технологические исследования Анализирует результаты технологических исследований	знать: принципы работы лингвистически ориентированных программных продуктов уметь: пользоваться лингвистически ориентированными программными продуктами владеть: навыками использования лингвистически ориентированных программных продуктов
ПК-11	Владеет основными методами инструментального анализа звучащей речи	ПК-11.1 ПК-11.2	Проводит различные типы инструментального анализа звучащей речи Проводит запись, сегментацию, аннотацию речевого сигнала	знать: цели и задачи теоретической и практической фонетики уметь: сопоставлять фонетические факты английского и родного языков; делать фонетический анализ, объяснять фонетические явления, использовать теоретические знания о фонетической системе английского языка на практике для решения конкретных лингвистических задач владеть (иметь навык(и)): навыками фонологического анализа; фонетической терминологией

ПК–7	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-7.1 ПК-7.2 ПК-7.3.	Разрабатывает и документирует программные интерфейсы Пользуется электронными языковыми ресурсами для решения прикладных лингвистических задач Анализирует требования к программному обеспечению	Знать: методы описания денотативной, концептуальной, коммуникативной и прагматической информации Уметь: использовать лингвистически-ориентированные программные системы. Владеть: основами дисциплин, необходимых для формализации лингвистических знаний и процедур анализа и синтеза лингвистических структур.

12. Объем дисциплины в зачетных единицах/часах в соответствии с учебным планом — 3 ЗЕТ / 108 час.

Форма промежуточной аттестации зачет.

13. Виды учебной работы:

Вид учебной работы	Трудоемкость (часы)			
	Всего	По семестрам		
		№ сем.8	№ сем.	…..
Аудиторные занятия	36	36		
в том числе: лекции	18	18		
практические	0	0		
лабораторные	18	18		
Самостоятельная работа	72	72		
Итого:	108	108		
Форма промежуточной аттестации (зачет)				

13.1Содержание дисциплины:

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины
1. Лекции		
1.1	Компьютерная обработка звучащей речи, обработка акустических данных.	**Лекция 1.** Речевой сигнал, его основные акустические характеристики: частота основного тона, спектральный частотный состав, амплитуда, длительность, фазовые характеристики. Классификация звуков речи, основанная на артикуляторных признаках. Понятие формант. Анализ акустических характеристик речевого сигнала на основе спектограмм, сонограмм. Компьютерная обработка акустических данных. Звуковые редакторы.Компьютерные базы фонетических данных. Речевые базы данных. Аллофоны, дифоны, сэмплы для разработки речевой БД. Озвученные словари. Принципы организации и этапы разработки.
1.2	Методы автоматического синтеза речи (системы texttospeech)	**Лекция 2.** Автоматический синтез речи: история, первые синтезаторы. Методы автоматического синтеза речи: артикуляторный синтез, формантный синтез по правилам, компилятивный синтез, синтез на основе коэффициентов линейного предсказания (КЛП-синтез), нейросетерые алгоритмы синтеза речи. Фонетическое обеспечение для создания программ синтеза речи. Обобщенная функциональная структура синтезатора. Основные блоки, их назначение и практическая реализация. Проблемы формирования просоидических характеристик речи в задачах синтеза: интонации, паузирование. Программные приложения и библиотеки синтеза речи по тексту.
1.3	Построение систем распознавания речи (системы speechtotext).	**Лекция 3.** Системы распознавания речи: классификация, функциональная структура. Примеры реализации. Вопросы обучения и настройки на голос. Голосовые интерфейсы, голосовые Web порталы. Голосовой поиск и управление. Использование нейросетей для автоматического распознавания речи. Программные приложения и библиотеки для распознавания голоса.
1.4	Архитектура и функционирование автоматизированных систем обработки текстов (АСОТ) (системы texttotext).	**Лекция 4.** Классификация систем автоматической обработки текстов (АСОТ), их архитектуры и сферы применения. Алгоритмы лингвистического разбора и анализа текста. Уровни текстового анализа: графематический, фонетический, морфологический, синтаксический, семантический. Парсеры ЕЯ-предложений. Стемминг и лематизация. Алгоритмы стемминга. Программные приложения и библиотеки для работы с морфологией естественного языка. **Лекция 5.** Синтаксические парсеры, форматы представления синтаксического разбора: синтаксическое дерево, деревья составляющих. деревья зависимостей, КС - грамматики. Примеры синтаксических парсеров и инструменты их разработки и интеграции. Формат разметки CONLLU.
1.5	Построение систем распознавания, обработки и классификаций текстовых документов.	**Лекция 6.** Математическая постановка задачи распознавания образов и классификации. Формальные методы определения сходства ЕЯ документов. Векторная модель. Методы определения сходства и классификации текстовых документов. Кластерный анализ текстов (рубрицирование, стилистика). Деревья принятия решений.

		Лекция 7. Алгоритмы машинного обучения для задач компьютерной лингвистики: SVM, байесовский классификатор, нейросети. Распознавания печатных и рукописных текстов. Задачи автоматического индексирования (рубрицирования), аннотирования. Классификация тональности высказываний. Задача определения авторства и стилистики текстов. **Лекция 8.** Проблема семантического анализа в системах обработки текстов. TextMining - извлечение фактов из текстов (именованных сущностей и ключевых слов), установление взаимосвязей. Типы именованных сущностей и способы извлечения из текстов. Проблемы разрешения омонимии, анафоры и кореференции. Автоматические системы извлечения знаний из разнородных текстовых источников. Задачи структурирования текстовых данных. Формирование онтологии предметной области по тексту. Графовые БД. Построение семантической модели текста. Онтограф текста. Семантический анализ текстов на основе онтологии предметной области. Форматы представления, стандарты разработки, инструменты.
1.6	Проблемы автоматизации синтеза текста. TextGeneration	**Лекция 9.** Проблемы автоматизации синтеза текста. Алгоритмы синтеза текста для вербализации заданного содержания. Семантические, морфологические, синтаксические проблемы синтеза.
2. Практические занятия		
Учебным планом не предусмотрены		
3. Лабораторные занятия		
3.1	Компьютерная обработка звучащей речи, обработка акустических данных.	**Лабораторная работа 1.** Исследование речевого сигнала, его основных акустических характеристик, параметров записи звука, форматов представления. Знакомство с методами обработки речевых сигналов (фильтрация и модуляция) для повышения качества разборчивости речи. На примере звукового редактора.. **Лабораторная работа 2.** Разработка фонетического обеспечение для создания программ синтеза речи. Подготовка речевого материала для разработки фонетической БД. Разработка фонетической БД.
3.2	Методы автоматического синтеза речи (системы texttospeech)	**Лабораторная работа 3.** Исследование систем распознавания речи. Обучение и настройка на голос. Голосовые интерфейсы. Разработка системы синтеза речи по тексту.
3.3	Построение систем распознавания речи (системы speechtotext).	**Лабораторная работа 4.** Разработка системы распознавания речевых команд, тестирование эффективности распознавания. (библиотека Vosk)
3.4	Архитектура и функционирование автоматизированных систем обработки текстов (АСОТ) (системы texttotext).	**Лабораторная работа 5.** Работа с библиотекой Natasha. **Лабораторная работа 6.** Работа с фреймворком DeepPavlov.
3.5	Построение систем распознавания, обработки и классификаций текстовых документов.	**Лабораторная работа 7.** Исследование систем извлечения данных из неструктурированных естественно-языковых текстов, систем понимания текстов (TextMining). Работа с Yargy - парсером, разработка грамматик для извлечения фактов из текстов. **Лабораторная работа 8.** Применение НС различных архитектур для задач NLP

		(LSTM, BERT-модели)
3.6	Проблемы автоматизации синтеза текста. TextGeneration	**Лабораторная работа 9.** Знакомство с генеративными нейросетями на основе архитектур трансформеров (ChatGPT).

13.2 Темы (разделы) дисциплины и виды занятий:

№ п/п	Наименование темы (раздела) дисциплины	Виды занятий (часов)				
		Лекции	Практические	Лабораторные	Самостоятельная работа	Всего
1	Компьютерная обработка звучащей речи, обработка акустических данных.	2		4	12	18
2	Методы автоматического синтеза речи (системы texttospeech)	2		2	9	13
3	Построение систем распознавания речи (системы speechtotext).	2		2	9	13
4	Архитектура и функционирование автоматизированных систем обработки текстов (АСОТ) (системы texttotext).	4		4	18	26
5	Построение систем распознавания, обработки и классификаций текстовых документов.	6		4	12	22
5	Проблемы автоматизации синтеза текста. TextGeneration	2		2	12	16
	Итого:	18		18	72	108

14. Методические указания для обучающихся по освоению дисциплины:

1) При изучении дисциплины рекомендуется использовать следующие средства**:**

- рекомендуемую основную и дополнительную литературу;
- методические указания и пособия;
- контрольные задания для закрепления теоретического материала;
- электронные версии учебников и методических указаний для выполнения лабораторно - практических работ (при необходимости материалы рассылаются по электронной почте).

2) Для максимального усвоения дисциплины рекомендуется проведение письменного опроса (тестирование, решение задач) студентов по материалам лекций и лабораторных работ. Подборка вопросов для тестирования осуществляется на основе изученного теоретического материала. Такой подход позволяет повысить мотивацию студентов при конспектировании лекционного материала.

3) При проведении лабораторных занятий обеспечивается максимальная степень соответствия с материалом лекционных занятий и осуществляется экспериментальная проверка методов, алгоритмов и технологий обработки информации, излагаемых в рамках лекций.

4) При переходе на дистанционный режим обучения для создания электронных курсов, чтения лекций он-лайн и проведения лабораторно-практических занятий используется информационные ресурсы Образовательного

портала "Электронный университет ВГУ (https:\\edu.vsu.ru), базирующегося на системе дистанционного обучения Moodle, развернутой в университете.

13. Перечень основной и дополнительной литературы, ресурсов Интернет, необходимых для освоения дисциплины:

а) основная литература:

№ п/п	Источник
1	Боярский, К. К. Введение в компьютерную лингвистику : учебное пособие / К. К. Боярский. — Санкт-Петербург : НИУ ИТМО, 2013. — 72 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/70822 (дата обращения: 01.05.2023). — Режим доступа: для авториз. пользователей.
2	Ганегедара, Т. Обработка естественного языка с TensorFlow : руководство / Т. Ганегедара ; перевод с английского В. С. Яценкова. — Москва : ДМК Пресс, 2020. — 382 с. — ISBN 978-5-97060-756-5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/140584 (дата обращения: 01.05.2023). — Режим доступа: для авториз. пользователей.
3	Онтологии и тезаурусы: модели, инструменты, приложения : учебное пособие / Б.В. Добров, В.В. Иванов, Н.В. Лукашевич, В.Д. Соловьев. - Москва : Интернет-Университет Информационных Технологий, 2009. - 173 с. : ил.,табл., схем. - (Основы информационных технологий). - ISBN 978-5-9963-0007-5 ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=233056
4	Леонтьева, Нина Николаевна. Автоматическое понимание текстов: системы, модели, ресурсы : учебное пособие для студентов лингвистических факультетов вузов / Н.Н. Леонтьева .— М. : Академия, 2006 .— 302, [1] с.

б) дополнительная литература:

№ п/п	Источник
1	Белоногов Г.Г.Компьютерная лингвистика и перспективные информационные технологии.- М.:Русский мир, 2004.
2	Зубов А.В., Зубова И.И. Информационные технологии в лингвистике. - М.: Академия, 2004.
3	Всеволодова А.В. Компьютерная обработка лингвистических данных. М.:Наука, 2007.
4	Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы. – М.: Академия, 2006.
5	Кипяткова И.С., Ронжин А.Л., Крапов А.А. Автоматическая обработка разговорной русской речи. - Санкт-Петербург: ГУАП, 2013.
6	Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.
7	Бенгфорт Бенджамин, Билбро Ребекка, Охеда Тони Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. — СПб.: Питер, 2019.
8	Моделирование распознавания рукописного текста на основе скрытых марковских моделей : монография / И. Я. Львович, Я. Е. Львович, А. П. Преображенский [и др.]. — Воронеж : ВИВТ, 2016. — 164 с. — ISBN 978-5-4446-0838-8. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/157486 (дата обращения: 01.05.2023). — Режим доступа: для авториз. пользователей.
9	Маккинни, У. Python и анализ данных / У. Маккинни ; перевод с английского А. А. Слинкина. — 2-ое изд., испр. и доп. — Москва : ДМК Пресс, 2020. — 540 с. — ISBN 978-5-97060-590-5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/131721 (дата обращения: 01.05.2023). — Режим доступа: для авториз. пользователей.
10	Бонцанини, М. Анализ социальных медиа на Python. Извлекайте и анализируйте

	данные из всех уголков социальной паутины на Python / М. Бонцанини ; перевод с английского А. В. Логунова. — Москва : ДМК Пресс, 2018. — 288 с. — ISBN 978-5-97060-574-5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/108129 (дата обращения: 01.05.2023). — Режим доступа: для авториз. пользователей.
11	Теофили, Т. Глубокое обучение для поисковых систем : руководство / Т. Теофили ; перевод с английского Д. А. Беликова. — Москва : ДМК Пресс, 2020. — 318 с. — ISBN 978-5-97060-776-3. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/140574 (дата обращения: 01.05.2023). — Режим доступа: для авториз. пользователей.
12	Маннинг, Кристофер Д. Введение в информационный поиск = Introduction to Information retrieval / Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце ; [пер. с англ. Д.А. Клюшина] .— М. ; СПб. ; Киев : Вильямс, 2011 .— 520 с.

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

№ п/п	Источник
1	Электронный каталог Научной библиотеки Воронежского государственного университета. – (http // www.lib.vsu.ru/).
2	Образовательный портал «Электронный университет ВГУ».– (https://edu.vsu.ru/)
3	ЭБС Лань – Лицензионный договор №3010-14/37-23 от 07.03.2023 (срок предоставления с 12.03.2023 по 11.03.2024)
4	ЭБС «Университетская библиотека online» – Контракт №3010-06/23-22 от 30.12.2022 (срок предоставления с 12.01.2023 по 11.01.2024)
5	ЭБС «Консультант студента» – Лицензионный договор №3010-06/22-22 от 30.12.2022 (с дополнительным соглашение №1 от 09.01.2023) (срок предоставления с 12.01.2023 по 11.01.2024)
6	Лаборатория компьютерной лингвистики Института проблем передачи информации РАН. http://proling.iitp.ru/
6	Лаборатория общей компьютерной лексикологии и лексикографии МГУ. http://www.philol.msu.ru/~lex/library.htm
7	Научно-практический журнал РЕЧЕВЫЕ ТЕХНОЛОГИИ http://speechtechnology.ru/

16. Перечень учебно-методического обеспечения для самостоятельной работы:

№ п/п	Источник
1.	Информационные ресурсы Образовательного портала "Электронный университет ВГУ (https:\\edu.vsu.ru)
2.	SonySoundForge 9. Учимся убирать шум [Электронный ресурс]. – Режим доступа: http://cjcity.ru/content/sound-forge-denoise.php
3	Боярский К. К. Введение в компьютерную лингвистику. Учебное пособие. – СПб: НИУ ИТМО, 2013. – 72 с.

17. Образовательные технологии, используемые при реализации учебной дисциплины, включая дистанционные образовательные технологии (ДОТ), электронное обучение (ЭО), смешанное обучение):

Для реализации учебного процесса используются:

1) ПО Microsoft в рамках подписки "Imagine/Azure Dev Tools for Teaching", договор №3010-16/96-18 от 29 декабря 2018г.
2) Звуковой редактор SoundForge (Свободно-распространяемое ПО)
3) Парсер русского языка Yargy (Свободно-распространяемое ПО)
4) Язык программирования Python, IDEPysharm.
5) Фреймворк DeepPavlov (Свободно-распространяемое ПО)
6) Библиотекой Python Natasha (Свободно-распространяемое ПО)
7) Редактор онтологий и фреймворк для построения баз знаний Protege. Свободно-распространяемое ПО.

При проведении занятий в дистанционном режиме обучения используются технические и информационные ресурсы Образовательного портала "Электронный университет ВГУ (https:\\edu.vsu.ru), базирующегося на системе дистанционного обучения Moodle, развернутой в университете, а также другие доступные ресурсы сети Интернет.

18. Материально-техническое обеспечение дисциплины:

/ауд. 12/ - компьютерный класс: Компьютер Arbyte Tempo/AOC (12 шт.), Проектор Benq MW523 (1 шт.), Сканер Canon Canoscan LiDE 120 (5 шт.) Экран проекционный (1 шт.) /ауд. 14/ Проектор Benq MW523 (1 шт.) Экран проекционный (1 шт.) Компьютер Asus H81m-Plus (11 шт.)	г.Воронеж, пл.Ленина 10, ауд. 12, 14

19. Оценочные средства для проведения текущей и промежуточной аттестаций

Порядок оценки освоения обучающимися учебного материала определяется содержанием следующих разделов дисциплины:

№ п/п	Наименование раздела дисциплины (модуля)	Компетенции	Индикаторы достижения компетенции	Оценочные средства
1	1.1 Компьютерная обработка звучащей речи, обработка акустических данных. 1.2 Компьютерные базы фонетических данных. 1.3 Методы автоматического синтеза речи (системы texttospeech) 1.4 Построение систем распознавания речи (системы speechtotext). 1.5 Архитектура и функционирование 1.6 Построение систем распознавания текста. 1.7 Проблемы автоматизации синтеза текста. TextGeneration	ПК–5 ПК-11	Осуществляет постановку задачи на технологические исследования (ПК-5.1) Анализирует результаты технологических исследований (ПК-5.2) Проводит различные типы инструментального анализа звучащей речи (ПК-11.1) Проводит запись, сегментацию, аннотацию речевого сигнала (ПК-11.2) Разрабатывает и документирует программные интерфейсы (ПК–	Устный опрос, выполнение индивидуальных практических работ, Тест № 1, 2

		ПК–7	7.1) Пользуется электронными языковыми ресурсами для решения прикладных лингвистических задач (ПК–7.2) Анализирует требования к программному обеспечению (ПК–7.3)	

20 Типовые оценочные средства и методические материалы, определяющие процедуры оценивания

20.1 Текущий контроль успеваемости

Для оценивания результатов обучения на зачете используются следующие содержательные показатели (формулируется с учетом конкретных требований дисциплины):

1) знание теоретических основ учебного материала, основных определений, понятий и используемой терминологии;

2) умение проводить обоснование и представление основных теоретических и практических результатов (теорем, алгоритмов, методик) с использованием математических выкладок, блок-схем, структурных схем и стандартных описаний к ним;

3) умение связывать теорию с практикой, иллюстрировать ответ примерами, в том числе, собственными, умение выявлять и анализировать основные закономерности, полученные, в том числе, в ходе выполнения лабораторно-практических заданий;

4) умение обосновывать свои суждения и профессиональную позицию по излагаемому вопросу;

5) владение навыками программирования и экспериментирования в рамках выполняемых лабораторных заданий;

Различные комбинации перечисленных показателей определяют критерии оценивания результатов обучения (сформированности компетенций) на зачете:

- высокий (углубленный) уровень сформированности компетенций;
- повышенный (продвинутый) уровень сформированности компетенций;
- пороговый (базовый) уровень сформированности компетенций.

Для оценивания результатов обучения на зачете используется – зачтено (выше порогового уровня), не зачтено (ниже порогового уровня) по результатам тестирования.

Соотношение показателей, критериев и шкалы оценивания результатов обучения на государственном экзамене представлено в следующей таблице.

Критерии оценивания компетенций и шкала оценок

Критерии оценивания компетенций	Уровень сформированности компетенций	Шкала оценок

Обучающийся демонстрирует полное соответствие знаний, умений, навыков по приведенным критериям свободно оперирует понятийным аппаратом и приобретенными знаниями, умениями, применяет их при решении практических задач.	Повышенный уровень	Отлично
Ответ на контрольно-измерительный материал не полностью соответствует одному из перечисленных выше показателей, но обучающийся дает правильные ответы на дополнительные вопросы. При этом обучающийся демонстрирует соответствие знаний, умений, навыков приведенным в таблицах показателям, но допускает незначительные ошибки, неточности, испытывает затруднения при решении практических задач.	Базовый уровень	Хорошо
Обучающийся демонстрирует неполное соответствие знаний, умений, навыков приведенным в таблицах показателям, допускает значительные ошибки при решении практических задач. При этом ответ на контрольно-измерительный материал не соответствует любым двум из перечисленных показателей, обучающийся дает неполные ответы на дополнительные вопросы.	Пороговый уровень	Удовлетворительно
Ответ на контрольно-измерительный материал не соответствует любым трем из перечисленных показателей. Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки	–	Неудовлетворительно

Пример задания для выполнения лабораторной работы

Лабораторная работа №2

Знакомство с методами обработки речевых сигналов (фильтрация и модуляция) для повышения качества разборчивости речи. На примере звукового редактора SoundForge.

Теоретический материал

SoundForge — набор инструментов для редактирования цифровых аудиофайлов, подходящий как для любителей, так и профессионалов. SoundForge позволит получить полный контроль над процессом монтажа и мастеринга, производить запись, отладку и восстановление звука, записывать звуковые компакт-диски, кодировать и декодировать любые форматы аудио.

Основное назначение SoundForge – редактирование цифрового звука. С помощью этой программы можно обрабатывать фонограммы или звуковые дорожки фильмов практически всеми существующими способами. К возможностям программы относятся:

- Первоначальная запись и оцифровка звука с различных источников – микрофона, магнитофона, проигрывателя виниловых дисков и т. п. с заданным качеством. В результате появляется исходная, необработанная фонограмма.
- Монтаж фонограмм: удаление, вырезание и вставка, «склеивание» фрагментов.
- Наложение одних фонограмм на другие, целиком или частями, микширование.
- Исправление дефектов фонограммы: удаление или существенное снижение шума, щелчков, посторонних или нежелательных звуков в полуавтоматическом режиме.
- Точная «ручная» подчистка отдельных участков фонограммы.
- Частотная коррекция: изменение тембра, маскировка или подчеркивание отдельных частотных составляющих.
- Нормализация уровня (громкости), изменение динамического диапазона записей.
- Восстановление «срезанных» пиков – искажений, возникающих при записи фонограмм с чрезмерно большим уровнем сигнала.
- Изменение продолжительности фонограмм или отдельных их фрагментов.
- Применение специальных эффектов: вибрато, реверберации, эха. Всего доступно более тридцати различных эффектов.

По умолчанию в главном окне программы SoundForge отображаются лишь некоторые панели инструментов.

- Строка меню – как и в большинстве приложений Windows, эта панель находится под заголовком окна программы.
- Стандартная панель инструментов – содержит кнопки для вызова наиболее общих и часто используемых действий: создания, открытия и сохранения файла, копирования, вырезки и вставки, отмены и повтора последних действий, а также выбора функций указателя мыши.
- Панель передачи – очень напоминает пульт управления магнитофона или проигрывателя. Ее кнопки позволяют запустить воспроизведение и запись, остановить их, включить паузу, а также служат для «ускоренной перемотки» фонограммы.
- Рабочая область окна – предназначена для размещения окон данных и других средств работы.
- Индикатор уровня – показывает текущий уровень воспроизводимого сигнала в каждом из каналов. По умолчанию этот инструмент прикреплен к правому краю окна.
- Строка состояния – здесь выводятся подсказки и комментарии к выполняемым действиям.

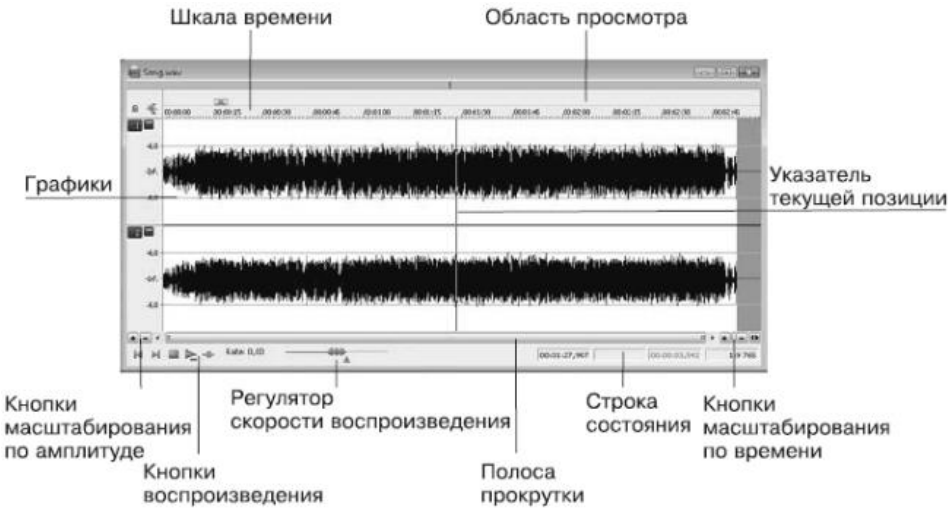

Рис. 4. Окно данных

В верхней части окна данных размещается Шкала времени, которая по умолчанию отградуирована в часах, минутах, секундах и тысячных долях секунды. Над шкалой времени находится полоса, на которой показываются значки установленных маркеров и областей, а также линия, отмечающая текущее положение, – область просмотра. В области просмотра всегда показывается весь файл, от начала и до конца, независимо от выбранного масштаба отображения графика. Большую часть окна занимают графики – изображения данных. Если открыта монофоническая (одноканальная) запись, то показывается один график; при стереофонической (двухканальной) записи графиков будет два. Верхний график изображает левый канал, нижний – правый. Расстояние от левого края графика до правого показывает длину файла.

Кроме прослушивания композиции, в окне данных можно устанавливать маркеры, выделять и создавать области, копировать их, вырезать и редактировать. Рассмотрим возможность создания маркеров.

Маркеры – это своеобразные флажки, метки, которые позволяют пометить на временной шкале точку редактирования. Например, маркерами можно помечать участки с плохим звучанием, чтобы впоследствии отредактировать их или удалить. Чтобы добавить маркер, нужно выполнить следующие действия.

1. Установить указатель текущей позиции в место, куда будет добавляться маркер.
2. Выполнить команду Специальные → Вставить маркер – откроется окно «Вставить маркер/Область».

3. В поле ввода со счетчиком **Начало** указано значение времени, в котором находится указатель текущего положения и куда будет устанавливаться маркер. При необходимости значение данного поля можно изменить.
4. В поле ввода **Имя** ввести обозначение маркера, например **Первый маркер**, и нажать кнопку **ОК** – данное окно закроется, а в окне данных на шкале времени появится маркер.

Удалить маркер можно, щелкнув на нем правой кнопкой мыши и выбрав в появившемся контекстном меню пункт **Удалить**.

Данные, отображаемые в области просмотра окна данных, можно масштабировать – уменьшать или увеличивать диаграмму, что позволит более детально рассматривать шумы, помехи и сделает работу более удобной.

В нижней части окна данных слева и справа от полосы прокрутки находятся две группы кнопок, на которых изображены знаки плюса и минуса. Правая группа кнопок позволяет увеличивать масштаб диаграммы горизонтально, корректируя значение времени. При нажатии кнопки **Увеличить масштабирование времени**, на которой изображен знак плюса, отображение данных увеличится и станет более детализированным.

Кнопка **Уменьшить масштабирование времени**, на которой изображен знак минуса, позволяет уменьшить отображение. При масштабировании времени значения линейки времени изменяются в сторону детализации. Индикатор разрешения **Коэффициент масштабирования** показывает текущую степень масштабирования в виде коэффициента. Степень масштабирования 1:1 дает наибольшую возможную детализацию диаграммы. По умолчанию степень масштабирования равна **1:4096**.

Рассмотрим такую команду редактирования, как подрезка. Это действие позволяет вырезать ненужные участки из фонограммы. Вырезанные участки можно удалять или перемещать в другие окна данных, создавая монтажи. Для редактирования нужно выполнить следующую последовательность действий.

1. Создать в окне данных выделенную область на участке, который нужно вырезать (рис. 2).

2. Увеличить выделенный участок, применив масштабирование, для чего следует нажать кнопку **Увеличить масштабирование времени.**

3. Выполнить команду меню **Правка** → **Вырезать** – выделенный фрагмент будет удален из окна данных и помещен в буфер обмена. Пока не выполнено новое сохранение в буфер обмена, можно вставить вырезанный фрагмент в новое окно данных или в новое положение в текущем окне данных. Если нужно удалить выделенный фрагмент без помещения его в буфер обмена, то следует выполнить команду **Правка** → **Удалить**. Если выполнить команду **Правка** → **Сократить/Обрезать**, то будет удалено все содержимое окна данных, кроме выделенного фрагмента.

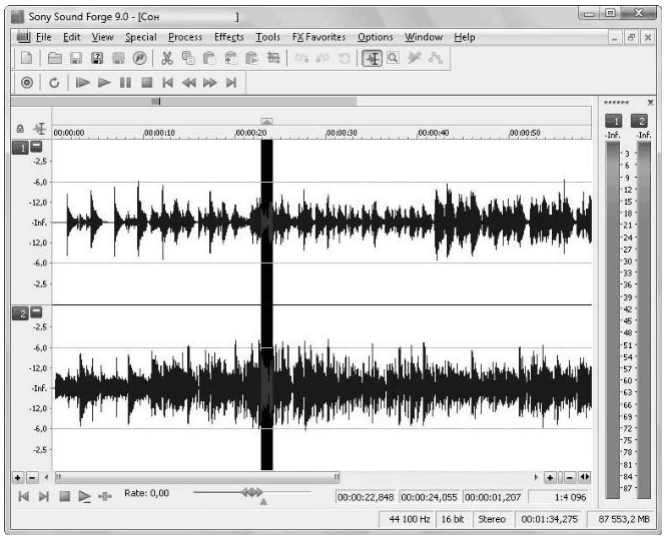

Рис. 5. Выделение фрагмента для удаления

Функция микширования позволяет сводить данные из буфера обмена и данные в открытом звуковом файле. Например, если требуется наложить одну композицию на другую, то можно использовать эту функцию. Для микширования данных необходимо выполнить следующие действия.

1. Выделить и скопировать в буфер данные, которые нужно микшировать.

2. В файле, в который предполагается вставить данные из буфера обмена, установить указатель текущей позиции в положение, куда требуется поместить данные из буфера обмена.

3. Выполнить команду **Правка** → **Специальная вставка** → **Микшировать** – откроется окно **«**Микшировать».

4. Установить требуемые значения громкостей с помощью ползунков **Громкость** и нажать кнопку **Просмотр**, чтобы проверить, не становится ли звук глухим. Данные ползунки регулируют соотношение громкостей данных из буфера обмена и из файла после микширования[11].

5. Если звук все же приглушился, то следует установить флажок **Инвертироватьданные**только для одного из сигналов, но не для обоих сразу. Это приводит к инвертированию сигнала и поможет разрешить проблему.

6. Нажать кнопку **ОК** – программа смикширует данные из буфера обмена и из файла в соответствии с выполненными установками.

Применение эффектов позволит придать звуковым композициям оригинальное звучание, сделать их объемными и расставить звуковые акценты. Для решения этой задачи в программе SoundForge 9.0 содержится обширный набор инструментов и эффектов, которые можно применить к звуковому фрагменту или всей композиции.

В программе доступно более 40 специальных эффектов и преобразователей звука.

Все эффекты можно разделить на группы.

- Эффекты эха – создают эффекты, добавляющие эхо.
- Эффекты высоты тона – создают эффекты звучания, основанные на изменении высоты тона.
- Эффекты громкости – создают эффекты на основе громкости, применяя амплитудную модуляцию, искажение, сжатие и др.
- Эффекты реверберации – эффекты, основанные на времени, дают ощущение пространства, позволяя имитировать звучание в концертном зале или закрытом помещении.

Общий алгоритм применения эффекта к звукам в программе SoundForge можно описать следующим образом.

1. Выделите звуковой файл или фрагмент файла.

2. Войдите в меню **Эффекты** и выберите в нем эффект, который нужно применить.

3. Настройте параметры эффекта по своему усмотрению или используйте заранее заданные установки из раскрывающегося списка **Предустановка**.

4. Прослушайте сделанные изменения, нажав кнопку **Просмотр**, и примените эффект.

Эффект **Простая задержка** позволяет применять базовые эффекты эха к звуковым фрагментам и композициям (Эффекты → Задержка/Эхо → Упрощенно).

Функция **Многоотводная задержка** позволяет создавать очень сложные эффекты эха за счет установки сразу нескольких задержек, подобно одновременной установке нескольких эффектов Simple (Эффекты → Задержка/Эхо → Мультисигнал).

К эффектам задержки, которые имеют разные настройки и дают различные варианты звучания, относятся также **Хор** и **Флэнджер/Вау-вау**. Программа содержит три различных эффекта, позволяющих изменять высоту тона аудиоданных разными способами: Отклонение высоты, Сдвиг высоты и **Вибрато**.

Эффект **Отклонение высоты** (Высота звука → Отклонение) позволяет изменять высоту тона аудиоданных на определенном периоде времени. Например, с помощью этого эффекта можно медленно повышать высоту тона от начала до конца записи.

Эффект **Сдвиг высоты** (Высота звука → Сдвиг) программы SoundForge может быть применен для изменения высоты тона без изменения длины аудиоданных, которое обычно сопровождает изменение тона.

SoundForge содержит также эффекты, которые основываются на изменении громкости. К ним относятся: Искажение, Пороговый шумоподавитель, Динамическое представление (Графическое, Многополосное), Огибающая, Интервал/Слияние и Амплитудная модуляция.

Амплитудная модуляция позволяет внести в звук такие эффекты, как тремоло на электрооргане. Тремоло (от итал. «дрожащий») обозначает специальный прием игры на музыкальных инструментах, заключающийся в многократном быстром повторении одного или нескольких звуков.

Искажение позволяет искажать звук, создавая неожиданные и яркие вкрапления в звучание, что может быть полезно, например, когда нужно добавить к звучанию голоса немного хрипоты.

Интервал/Слияние добавляет к записи фрагменты данных или обрезает их, позволяя генерировать эффекты тремоло, трели и заикания. Фрагменты данных, применяемые к записи или удаляемые из нее, настолько малы, что не разрушают запись, а приводят к появлению интересных звуков.

Пороговый шумоподавитель - тип эффекта, в котором используется цифровой шлюз сигнала для удаления части звуковых данных, например, если необходимо сделать промежутки между звуковыми паузами в композиции тихими и бесшумными.

Еще один эффект программы SoundForge – **реверберация**. Это особая форма эффекта затухания, добавляющую сложную последовательность очень коротких эффектов эха, имитирующих искусственную среду, то есть реверберация – это результат взаимодействия звука с помещением. Используя данный эффект, можно имитировать звучание аудиоданных в различных средах, например в концертном зале или пустой комнате. SoundForge содержит два эффекта реверберации: **Реверберация** и **Аккустическое зеркало**.

Задание

Смонтировать аудиозапись из двух медиафайлов с применением функции микширования, обработки звуковых дорожек и дополнительных звуковых эффектов.

Порядок выполнения

1. Загрузить медиафайл №1 в проект.
2. Используя маркеры/области выделения выделить ряд отрезков аудиозаписи для их последующего удаления.
3. Удалить выделенные отрезки.
4. Загрузить медиафайл №2 в проект.
5. Выделить в загруженном медиафайле отрезок, который впоследствии будет вставлен в медиафайл №1.
6. Скопировать выделенный отрезок аудиозаписи.

7. Вернуться к исходному файлу №1 и вставить отрезок аудиозаписи, скопированный из другого файла.
8. Повторить пункты 5-6 с целью последующего применения функции микширования.
9. Используя функцию микширования, наложите фрагмент одной композиции на другой.
10. Применить обработку аудиозаписи: команды постепенного изменения уровня сигнала (увеличить и уменьшить) для начала и конца записи.
11. Применить 2 дополнительных эффекта к получившейся аудиозаписи.
12. Сохранить получившийся аудиофайл и показать его преподавателю.
13. Подготовить отчет о выполнении лабораторной работы, используя «Документы Google». Открыть доступ к отчету преподавателю (с правами редактора).

Приведённые ниже задания рекомендуется использовать при проведении диагностических работ для оценки остаточных знаний по дисциплине

Вопросы с выбором 1-балл

1.Что относится к лингвистическим ресурсам для разработки программного обеспечения систем компьютерной лингвистики (множественный выбор):

1. Базы словосочетаний
2. Тезаурусы
3. Онтологии
4. Текстовые корпуса
5. Базы данных
6. Компьютерные словари

Ответ: 1,2,3,4,6.

2. Поставить соответствие:

1. Графематический анализ	a) *Выделение грамматической основы слова, определение частей речи, приведение слова к словарной форме.*
2.Фонетический анализ	b) *Выявление смысловых связей между словами и группами, извлечение семантических отношений.*
3. Морфологический анализ	c) *Выявление синтаксических связей между словами в предложении, построение синтаксической структуры предложения.*
4.Синтаксический анализ	*d) анализ звукового состава слова, позволяет вычленить в слове звуки и определить их характеристики.*
5.Семантический анализ	*e) Выделение из массива данных предложений и слов (токенов).*

Ответы: 1-e, 2-d, 3-a, 4-c, 5-b.

3. Поставить соответствие:

1. Токенизация	a) Процесс использует словарь и морфологический анализ, чтобы в итоге привести слово к его канонической

	форме (словарной форме).
2. Стемминг	b) Процесс разделения текста на предложения-компоненты или процесс разделения предложений на слова-компоненты.
3. Лемматизация	c) Процесс отсечения от слова окончаний и суффиксов, чтобы оставшаяся часть была одинаковой для всех грамматических форм слова.

Ответы: 1-b, 2-c, 3-a.

5.Какие формальные модели используются в компьютерной лингвистике? (множественный выбор)

a) Контекстно-свободные грамматики
b) Марковские модели
c) Мультиагентные модели
d) Графовые модели
e) Автоматные модели
f) Динамические модели

Ответы: a, b, d, e.

7. Какие утверждения верны для задания Контекстно-свободной грамматики (множественный выбор):

a) Конечное множество A - алфавит. Его элементы называются символами. Конечные последовательности символов образуют слова в данном алфавите.
b) Алфавит разделяется на терминальные (“окончательные”) и нетерминальные (“промежуточные”) символы.
c) Среди нетерминальных символов может быть выбран один – начальный.
d) Правила грамматики имеют вид K->X, где K-нетерминальный символ, а X- слово, в которое могут входить и терминальные, и нетерминальные символы.
e) Правила грамматики имеют вид aKb->aXb, где K нетерминальный символ, окруженный как нетерминальными, так и терминальными символами a,b. X- слово, в которое могут входить и терминальные, и нетерминальные символы, окруженное нетерминальными и терминальными символами.

Ответы: a, b, c, d.

Вопросы с коротким ответом 2-балла

Приведите название закона, отражающего эмпирическую закономерность распределения частоты слов естественного языка: “если все слова языка (или просто длинного текста) упорядочить по убыванию частоты их использования, то

частота *n*-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру *n.*”

Ответ: Закон Ципфа

Вопросы с развернутым ответом 3-балла

1. Опишите реализацию стеминга для русского языка на основе алгоритма Портера. Опишите шаги алгоритма.

Ответ:
Идея алгоритма: существует ограниченное количество словообразующих суффиксов, и стемминг слова происходит без использования каких-либо баз основ: только множество существующих суффиксов и вручную заданные правила.
Алгоритм состоит из пяти шагов.
На каждом шаге отсекается словообразующий суффикс и оставшаяся часть проверяется на соответствие правилам (например, для русских слов основа должна содержать не менее одной гласной). Если полученное слово удовлетворяет правилам, происходит переход на следующий шаг. Если нет — алгоритм выбирает другой суффикс для отсечения.
На первом шаге отсекается максимальный формообразующий суффикс,
на втором — буква «и»,
на третьем — словообразующий суффикс,
на четвертом — суффиксы превосходных форм, «ь» и одна из двух «н».

20.2 Промежуточная аттестация

Промежуточная аттестация может включать в себя проверку теоретических вопросов, а также, при необходимости (в случае не выполнения в течение семестра), проверку выполнения установленного перечня лабораторных заданий, позволяющих оценить уровень полученных знаний и/или практическое (ие) задание(я), позволяющее (ие) оценить степень сформированности умений и навыков.

Для оценки теоретических знаний используется перечень контрольно-измерительных материалов. Каждый контрольно-измерительный материал для проведения промежуточной аттестации включает два задания - вопросов для контроля знаний, умений и владений в рамках оценки уровня сформированности компетенции. При оценивании используется количественная шкала. Критерии оценивания представлены в приведенной ниже таблице

Для оценивания результатов обучения на экзамене используются следующие содержательные показатели (формулируется с учетом конкретных требований дисциплины)

1. знание теоретических основ учебного материала, основных определений, понятий и используемой терминологии;
2. владение навыками проведения компьютерного эксперимента, тестирования компьютерных алгоритмов обработки информации.

3. владение навыками программирования и экспериментирования с компьютерными моделями алгоритмов обработки информации в рамках выполняемых лабораторных заданий;
4. умение обосновывать свои суждения и профессиональную позицию по излагаемому вопросу;
5. умение связывать теорию с практикой, иллюстрировать ответ примерами, в том числе, собственными, умение выявлять и анализировать основные закономерности, полученные, в том числе, в ходе выполнения лабораторно-практических заданий;
6. умение проводить обоснование и представление основных теоретических и практических результатов (теорем, алгоритмов, методик) с использованием математических выкладок, блок-схем, структурных схем и стандартных описаний к ним;

Различные комбинации перечисленных показателей определяют критерии оценивания результатов обучения (сформированности компетенций) на зачете: высокий (углубленный) уровень сформированности компетенций; повышенный (продвинутый) уровень сформированности компетенций; пороговый (базовый) уровень сформированности компетенций.

Для оценивания результатов обучения на зачете с оценкой используется 4-балльная шкала: «отлично», «хорошо», «удовлетворительно», «неудовлетворительно». Для оценивания результатов обучения на зачете используется – зачтено, не зачтено по результатам тестирования.

Соотношение показателей, критериев и шкалы оценивания результатов обучения на зачете с оценкой представлено в следующей таблице.

Критерии оценивания компетенций и шкала оценок на зачете

Примерный перечень применяемых оценочных средств

№ п/п	Наименование оценочного средства	Представление оценочного средства в фонде	Критерии оценки
1	2	3	4
1	Устный опрос	Вопросы по темам/разделам дисциплины	Правильный ответ – зачтено, неправильный или принципиально неточный ответ - не зачтено
2	Контрольная работа по разделам дисциплины	Теоретические вопросы по темам/разделам дисциплины	Шкала оценивания соответствует приведенной в разделе 19.2
3	Лабораторная работа	Содержит 9 лабораторных заданий, предусматривающихосвоение программных системобработки текста и речи.	При успешно выполнении работы ставится оценка зачтено и осуществляется допуск к зачету, в противном случае ставится оценка не зачтено и обучающийся не допускается к зачету.
4	КИМ промежуточной аттестации	Каждый контрольно-измерительный материал для проведения промежуточной аттестации включает 2 заданий вопросов для контроля знаний, умений и владений в рамках оценки уровня сформированности	Шкалы оценивания приведены в разделе 19.2

		компетенции.	
5			

Пример контрольно-измерительного материала

УТВЕРЖДАЮ
Заведующий кафедрой технологий обработки и защиты информации

_____ А.А. Сирота
__.__.2023

Направление подготовки / специальность

45.03.03 Фундаментальная и прикладная лингвистика

Дисциплина Б1.В.02 Автоматическая обработка естественного языка

Форма обучения Очное

Вид контроля Зачет

Вид аттестации Промежуточная

Контрольно-измерительный материал № 1

1. Стемминг и лематизация. Стемминг: классификация алгоритмов стемминга. Алгоритмы стемминга для русского языка: Портера, Stemka, MyStem. Ошибки стемминга.
2. Автоматические системы извлечения знаний из разнородных текстовых источников. Задачи структурирования текстовых данных. Извлечение именованных сущностей и отношений между ними - подходы.

Преподаватель _____ В.В.Гаршина

Примерный перечень вопросов к зачету

1. Компьютерная лингвистика как междисциплинарная область.
2. Основные задачи, решаемые компьютерной лингвистикой. Направления исследований.
3. Проблемы моделирования естественного языка в компьютерной лингвистике: виды и особенности моделей.
4. Лингвистические ресурсы, используемые для обработки текста и речи.
5. Прикладные задачи компьютерной лингвистики.
6. Основные механизмы звукообразования речи. Речевой сигнал, его основные акустические характеристики: частота основного тона, спектральный частотный состав, амплитуда, длительность, фазовые характеристики.
7. Классификация звуков речи, основанная на артикуляторных признаках. Понятие формант. Анализ акустических характеристик речевого сигнала на основе спектограмм, сонограмм. Привести примеры компьютерных программ анализа звучащей речи.
8. Компьютерная обработка акустических данных. Звуковые редакторы: основные принципы работы. Примеры.
9. Компьютерные базы фонетических данных. Речевые базы данных. Озвученные словари. Принципы организации и этапы разработки.
10. Аллофонные и дифонные БД. Фонетическое обеспечение для создания программ синтеза речи.

11. Автоматический синтез речи: история, первые синтезаторы.
12. Методы автоматического синтеза речи: артикуляторный синтез, формантный синтез по правилам, компилятивный синтез, синтез на основе коэффициентов линейного предсказания (КЛП-синтез).
13. Обобщенная функциональная структура синтезатора. Основные блоки, их назначение и практическая реализация.
14. Проблемы формирования просоидических характеристик речи в задачах синтеза: интонации, паузирование.
15. Системы распознавания речи: классификация, функциональная структура. Примеры реализации.
16. Системы понимания речи, функциональная структура. Примеры реализации.
17. Компьютерное оборудование для обработки звучащей речи: АЦП, ЦАП, периферийное мультимедийное оборудование.
18. Автоматическая обработка текста как проблема компьютерной лингвистики. Классификация систем автоматической обработки текстов, сферы применения.
19. Архитектура автоматизированных систем обработки текстов (АСОТ). Лингвистический процессор: структура, функционирование.
20. Методы морфологического анализа, используемые в лингвистических процессорах. Морфологические словари.
21. Построение обобщенного синтаксического анализатора для АСОТ.
22. Варианты реализации семантического анализа для АСОТ.
23. Проблемы автоматизации синтеза текста. Алгоритмы синтеза текста для вербализации заданного содержания. Семантические, морфологические, синтаксические проблемы синтеза.
24. Автоматическое аннотирование и индексирование научно-технической документации. Автоматическое реферирование.
25. Проблемы автоматической обработки ошибок в печатных текстах. Автоматические корректоры.
26. Вероятностно-статистические характеристики текста, его элементов. Их применение в задачах лингвистики
27. Уровни текстового анализа: графематический, фонетический, морфологический, синтаксический, семантический. Основные задачи, их взаимосвязь.
28. Графематический анализ: задачи, методы реализации, примеры.
29. Графематический анализ: выделение структурных элементов в тексте: границы предложений, слов, словари сокращений.
30. Морфологический анализ. Задачи, методы реализации, примеры морфоанализаторов и инструментов разработки.
31. Стемминг и лематизация. Стемминг: классификация алгоритмов стемминга. Алгоритмы стемминга для русского языка: Портера, Stemka, MyStem. Ошибки стемминга.
32. Лематизация: используемые методы, примеры для русского языка, инструменты разработки.
33. Автоматическое выделение именованных сущностей и ключевых слов. Типы именованных сущностей и способы извлечения из текстов. Применение в системах обработки текстов/
34. Синтаксический анализ в компьютерной лингвистике. Способы представления синтаксического разбора: синтаксическое дерево, размеченное предложение. Примеры синтаксических парсеров и инструменты разработки.
35. Формальная модель представления синтаксиса: деревья составляющих. Грамматики составляющих.
36. Формальная модель представления синтаксиса: деревья зависимостей.
37. Формальная модель представления синтаксиса: КС- грамматики.

Прикладные системы обработки и хранения текстовых данных

38. Проблемы автоматизации синтеза (генерации) текста. Этапы генерации (схема). Методы генерации.
39. Шаблонные системы генерации. Генерация текстов на основе БД - простой отчет, связанный отчет. ЕЯ запросы к БД.
40. Алгоритмы синтеза текста для вербализации заданного содержания. Семантические, морфологические, синтаксические проблемы синтеза текстов.

41. Автоматическое аннотирование: архитектура построения систем, используемые методы, прикладное использование, примеры действующих систем.
42. Автоматическое реферирование: архитектура построения систем, используемые методы, прикладное использование, примеры действующих систем.
43. Информационный поиск: архитектура, модели представления документов, обработка поисковых запросов, извлечение документов.
44. Модели информационного поиска: инвертированная индексация, Булева и векторная модели. Метрики оценки близости документов. Оценка качества поиска: tf-idf, точность, полнота.
45. Вопросно-ответные системы: индексирование в информационно-поисковых системах, архитектура, способы обработки запросов, генерация различных типов ответов. Генерация диалогов в вопросно-ответных системах - чат-боты.
46. DataMining и TextMining. Извлечение фактов из текстов, установление взаимосвязей. Проблемы разрешения омонимии, анафоры и кореферентности.
47. Автоматические системы извлечения знаний из разнородных текстовых источников. Задачи структурирования текстовых данных. Извлечение именованных сущностей и отношений между ними. Подходы.
48. Извлечение фактов на основе контекстно-свободных грамматик, реализуемых в Томита и Yargy парсерах. Примеры грамматик.