Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Государственный институт русского языка им. А.С. Пушкина»

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Наименование дисциплины (модуля):

Технологии создания мультимедийного корпуса

Б1.В.ДВ.05.04

Уровень высшего образования:

магистратура

Направление подготовки:

Лингвистический анализ и экспертиза текста

(код и название направления подготовки)

Направленность (профиль) образования:

45.04.01 "Прикладная филология"

Форма обучения:

очная

Рабочая программа дисциплины (модуля) разработана в соответствии с федеральным государственным образовательным стандартом высшего образования по направлению подготовки, утвержденным приказом Минобрнауки России № 980 от 12.08.2020

1. Место дисциплины (модуля) в структуре ОПОП

В соответствии с учебным планом дисциплина Б1.В.ДВ.05.04 «Технологии создания мультимедийного корпуса» относится к части, формируемой участниками образовательных отношений, и изучается во 2 семестре.

2. Цель и задачи освоения дисциплины (модуля)

Цель курса: формирование системных компетенций в области проектирования, разработки и лингвистической разметки мультимедийных корпусов данных для проведения прикладных и фундаментальных исследований в области русского языка.

.

3. Планируемые результаты обучения по дисциплине (модулю), соотнесенные с требуемыми компетенциями выпускников

Компетенции выпускников (коды)	Индикаторы достижения компетенций	Планируемые результаты обучения по дисциплине (модулю), соотнесенные с индикаторами достижения компетенций
ПК-1 Способен проводить разноаспектное описание лексических единиц для решения	ПК-1.1 Выделяет и систематизирует значимые аспекты описания лексических единиц в соответствии с целями исследования и типом материала	Знать: основные понятия лексикографии и корпусной лингвистики: лексическая единица (лемма, вокабула, словоформа), коллокация, коллигация, валентность, семантические роли, прагматические маркеры, коннотации Уметь: формулировать цели

теоретических и прикладных задач		лексикографического описания применительно к конкретному мультимедийному корпусу и типу исследования (например, для задач машинного обучения, лексикографии или дискурс-анализа) Владеть: методами количественного и качественного анализа лексики в мультимедийной среде: подсчет совместной встречаемости лексемы и определенного жеста, анализ дистрибуции слова в разных регистрах речи (например, в студийном интервью и уличном опросе)
	ПК-1.2. Применяет цифровые и корпусные технологии для работы с лексическими данными	Знать: принципы разметки (аннотирования) лексического уровня в мультимедийных корпусах: стандарты *XML*, использование тегов для морфологических (*POS*-теггинг), семантических и просодических характеристик; типологию мультимедийных данных и их влияние на описание лексики Уметь: разрабатывать и адаптировать схемы разметки для лексического уровня, включая создание пользовательских тегов для

	ПК-1.3. Представляет результаты разноаспектного описания лексических единиц в формате, пригодном для разных пользователей	мультимодальных явлений. Владеть: применять специализированное программное обеспечение (*ELAN*, *EXMARaLDA*, *Audacity*, *Praat*) для сегментации мультимедийного потока и соотнесения лексических единиц с тайм-кодами, визуальными дорожками и звуковой волной Знать: методы извлечения и представления лексической информации: конкорданс, частотные словари, тезаурусы, *n-граммы*, *word embeddings* (векторные представления слов), дистрибутивная семантика. Уметь: проводить семантическую, прагматическую и мультимодальную разметку текстовых фрагментов, выделяя значимые аспекты (например, маркировать эмфазы, иронию, неуверенность, сарказм) с опорой на аудио- и видеоконтекст. Владеть: навыками проектирования и ведения глоссариев, тегсетов и руководств по разметке (*annotation guidelines*) для коллективной работы

		над мультимедийным корпусом.

4. Объем дисциплины (модуля) _____2_____з.е., в том числе _____32_____академических часов на контактную работу обучающихся с преподавателем; _____38_____академических часов на самостоятельную работу обучающихся.

(В конце данного пункта следует отметить, если дисциплина или часть ее реализуется с применением электронного обучения и дистанционных образовательных технологий)

5. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов

Наименование и краткое содержание разделов и тем дисциплины (модуля). Форма промежуточной аттестации по дисциплине (модулю)	Всего (часы)	В том числе							
		Контактная работа (работа во взаимодействии с преподавателем) *Виды контактной работы, часы*					Самостоятельная работа обучающегося *Виды самостоятельной работы, часы*		
		Занятия лекционного типа	Занятия семинарского типа (практические/ лабораторные)	Групповые консультации	Индивидуальная работа с обучающимися	Всего			Всего
Тема 1 **Теоретико-методологические основы корпусной**		2	2			4		8	12

лингвистики и мультимодальности. Понятие корпуса, аннотации, разметки. Специфика мультимодальных данных (текст, аудио, видео, жесты, пространственные данные) и их интеграция. Отличие мультимедийного корпуса от текстового и мультимодального.									
Тема 2 **Проектирование мультимедийного корпуса.** Формулирование исследовательских задач и гипотез. Определение целевой аудитории и типа корпуса (репрезентативный, диахронический, обучающий). Разработка архитектуры, схемы данных, выбор метаданных и принципов		2	2			4		6	10

балансировки выборки.									
Тема 3 **Сбор и предобработка мультимедийных данных.** Этика и право: авторское право, лицензии (Creative Commons), анонимизация персональных данных и получение информированного согласия. Технические аспекты: захват аудио и видео, оцифровка, извлечение текста из изображений (OCR) и видео (OCR/ASR), очистка и нормализация данных.		2	2			4		4	8
Тема 4 **Автоматическая и ручная аннотация мультимодальных данных.** Текстовые уровни (морфологическая, синтаксическая, семантическая, дискурсивная разметка, TEI, PAULA,		2	2			4		4	8

EXMARaLDA). Сегментация и синхронизация потоков (таймкоды, выравнивание аудио-текста). Разметка невербальных каналов: просодия, эмоции, жесты, направление взгляда, мимика. Использование инструментов (ELAN, Praat, ANVIL, Brat) и полуавтоматических пайплайнов.									
Тема 5 **Технологическая инфраструктура и разработка корпуса.** Выбор и настройка систем управления данными (CMS). Реляционные (PostgreSQL) и NoSQL (графовые, документные) базы данных для мультимедиа. Использование специализированных фреймворков (например, LaBB-CAT).		2	2			4		4	8

Принципы FAIR для языковых ресурсов. Разработка пользовательских интерфейсов (веб-вьюеры, такие как ANNIS) и API для доступа к корпусу.									
Тема 6 **Обработка сигналов и извлечение признаков.** Цифровая обработка звука: очистка от шума, нормализация громкости, выделение акустических признаков (форманты, F0, интенсивность). Компьютерное зрение для видео: распознавание лиц, трекинг объектов, анализ поз и жестов. Применение методов машинного обучения для автоматической классификации и кластеризации модальностей.		2	2			4		4	8
Тема 7 **Верификация, публикация и сопровождение**		2	2			4		4	8

корпуса. Оценка качества разметки (интерпретаторское согласие, метрики F1, Каппа). Версионирование данных и документации. Публикация в национальных и международных репозиториях (например, CLARIN, ELRA, LDC), получение постоянных идентификаторов (DOI). Обеспечение долгосрочной сохранности и масштабируемости ресурса.									
Тема 8 **Прикладные исследования на материале мультимедийных корпусов.** Использование корпуса для решения конкретных лингвистических, психолингвистических и		2	2			4		4	8

социолингвистических задач (анализ устной речи, межкультурная коммуникация, жестовые языки). Применение в образовательных технологиях, медиааналитике, разработке диалоговых систем и синтеза речи. Дизайн магистерских диссертаций на базе собранного корпуса.									
Промежуточная аттестация *(указывается форма проведения)*	*зачет*								2
Итого		16	16			32		38	72

6. Фонд оценочных материалов по дисциплине (модулю)

Типовые задания и иные материалы, необходимые для оценки результатов обучения

Модуль 1. Проектирование и аннотирование (теория и разметка)

1. Проектирование схемы корпуса. Разработать концептуальную модель мультимедийного корпуса (500 фрагментов) на материале региональных новостных репортажей. Описать цели, пользовательские сценарии, типы данных (аудио, видео, текст, метаданные) и структуру хранения.
2. Создание глоссария аннотации. Составить многоуровневый глоссарий тегов для разметки просодии (интонация, темп, паузы, громкость) и паралингвистики (смех, вздохи, покашливания) в устной речи, с примерами из русского языка.

3. Схема XML-разметки. Спроектировать иерархическую схему XML (или формат *.eaf для ELAN) для синхронной разметки видеофрагмента. Схема должна включать уровни: фоновый шум, жесты спикера, ключевые слова, интонационные контуры.
4. Разметка фрагмента в ELAN. Выполнить полную ручную разметку 3-минутного фрагмента ток-шоу в программе ELAN. Разметить не менее 5 уровней: реплики, перебивания, смену говорящего, жестикуляцию и смену плана в видео.
5. Автоматизация с помощью Praat. Написать скрипт на языке Praat для автоматического извлечения значений F0 (основной частоты) и интенсивности из 10 аудиофайлов, их нормализации и сохранения в таблицу CSV.
6. Выравнивание текста и аудио. Используя инструмент Forced Alignment (например, Montreal Forced Aligner или WebMAUS), провести автоматическое выравнивание расшифровки и аудиодорожки для 5-минутной лекции. Оценить точность и вручную исправить не менее 20 ошибок сегментации.

Модуль 2. Обработка данных и программирование (NLP и скрипты)

7. Пайплайн предобработки. Написать скрипт на Python, который принимает на вход папку с «сырыми» транскриптами и выполняет их очистку: удаление маркеров неуверенности (*эм*, *ну*), нормализацию пунктуации, токенизацию и лемматизацию с помощью библиотеки pymorphy2.
8. Извлечение именованных сущностей (NER). Применить модели Natasha или spaCy к транскриптам интервью для автоматического выделения персон, организаций и локаций. Написать скрипт для ручной валидации и исправления 100 случайных срабатываний модели.
9. Синхронизация субтитров. Написать Python-скрипт, который сопоставляет текстовые субтитры (.srt/.vtt) с таймкодами в медиафайле, автоматически корректирует задержки (drift) и объединяет короткие реплики (менее 1 секунды).
10. Визуализация зависимостей. Сгенерировать синтаксические деревья (Dependency Parse) для 20 сложных предложений из корпуса судебных заседаний. Визуализировать их с помощью displaCy и описать типичные синтаксические конструкции официально-делового стиля.
11. Кластеризация тем. Провести тематическое моделирование (LDA) на корпусе из 50 расшифровок подкастов. Визуализировать распределение тем во времени и подготовить отчет о доминирующих тематиках в разные годы.
12. Работа с API. Написать скрипт, который через API загружает видео с выбранной платформы, извлекает метаданные (длительность, описание, теги) и автоматически генерирует заглушки для текстового слоя корпуса.

Модуль 3. Мультимодальность и анализ (связь модальностей)

13. Корреляция жеста и интонации. Сформулировать исследовательскую гипотезу (например, «указательный жест вниз всегда совпадает с падением тона в конце фразы»). Проверить её на выборке из 50 фраз корпуса, используя статистический тест (хи-квадрат или точный тест Фишера).
14. Анализ эмоциональной валентности. Разметить 10 видеофрагментов по шкале валентности (отрицательная/положительная эмоция) на основе мимики спикера (используя OpenCV или библиотеки распознавания эмоций). Сопоставить результат с лексической тональностью текста (SentiWordNet или dostoevsky).
15. Сопоставление плана видео и дискурса. Проанализировать, как смена плана в видео (общий, средний, крупный) коррелирует с прагматическими функциями речи: введение новой темы, апелляция к зрителю, передача цифр.
16. Создание мультимодального конкорданса. Используя AntConc или Sketch Engine, создать конкорданс ключевого слова (например, «поэтому»). Для 30 вхождений вручную проверить и отметить, сопровождается ли это слово в видео характерным жестом или сменой интонации.

Модуль 4. Публикация и архитектура (базы данных и веб)

17. Проектирование базы данных. Спроектировать реляционную схему БД (SQL) или графовую схему (Neo4j) для хранения мультимедийного корпуса. БД должна поддерживать сложные запросы: например, «найти все фрагменты, где женщина старше 40 лет говорит о финансах с восходящей интонацией, стоя в полный рост».
18. Разработка прототипа интерфейса. Сверстать статический HTML/CSS/JS-прототип страницы конкорданса. Интерфейс должен отображать ключевое слово в контексте, слева показывать видеоплеер с привязкой к таймкоду, а справа — график частоты употребления.
19. Настройка поискового движка. Развернуть локальный экземпляр поискового движка (OpenSearch или Typesense). Загрузить в него 1000 документов корпуса, настроить ранжирование с учетом метаданных (дата, пол говорящего) и полнотекстового поиска.
20. Паспорт корпуса и лицензия. Написать финальный паспорт (metadata schema) для спроектированного в первом задании корпуса по стандартам OLAC или CMDI. Выбрать лицензию (например, CC BY-NC-SA), составить пользовательское соглашение и подготовить инструкцию для аннотаторов.

Методические рекомендации по использованию

- Баланс: задания сочетают ручную разметку (для понимания природы данных), элементы программирования (для масштабируемости) и лингвистический анализ (для исследовательской задачи).

- Материалы: для выполнения заданий рекомендуется использовать открытые русскоязычные данные (подкасты, записи заседаний Государственной Думы, региональные новости, интервью из НКРЯ, если их лицензия позволяет экспорт).

Шкала и критерии оценивания *(шкала и критерии оценивания могут быть едиными (типовыми) для всех дисциплин (модулей), входящих в основную образовательную программу)*

Критерии оценивания результатов обучения по дисциплине при проведении промежуточной аттестации:

<table>
<tr><td rowspan="2">Планируемые результаты обучения по дисциплине</td><td colspan="4">Критерии оценивания результатов обучения по дисциплине</td></tr>
<tr><td>«неудовлетворительно»</td><td>«удовлетворительно»</td><td>«хорошо»</td><td>«отлично»</td></tr>
<tr><td></td><td>«не зачтено»</td><td colspan="3">«зачтено»</td></tr>
</table>

Сформированность знаний (знать)	Уровень знаний ниже минимальных требований. Имели место грубые ошибки. Невозможно оценить полноту знаний вследствие отказа обучающегося от ответа.	Минимально допустимый уровень знаний. Допущено много негрубых ошибок.	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько несущественных ошибок.	Уровень знаний в объеме, соответствующем программе подготовки, или в объеме, превышающем программу подготовки, продемонстрирован без ошибок и погрешностей.

Сформированность умений (уметь)	Отсутствие минимальных умений. При решении стандартных задач не продемонстрированы основные умения. Имели место грубые ошибки. Невозможно оценить наличие умений вследствие отказа обучающегося от ответа.	Продемонстрированы основные умения. Решены типовые задачи с негрубыми ошибками. Выполнены все задания, но не в полном объеме.	Продемонстрированы все основные умения. Решены все основные задачи, но некоторые с негрубыми ошибками. Выполнены все задания, в полном объеме, но некоторые с недочетами.	Продемонстрированы все основные умения. Решены все основные задачи. Выполнены все задания, в полном объеме без недочетов.

Сформированность навыков (владеть)	При решении стандартных задач не продемонстрированы базовые навыки. Имели место грубые ошибки. Невозможно оценить наличие навыков вследствие отказа обучающегося от ответа.	Продемонстрирован минимальный набор навыков. Решены стандартные задачи с некоторыми недочетами.	Продемонстрированы базовые навыки. Решены стандартные задачи, но некоторые с недочетами.	Продемонстрировано всестороннее владение навыками. Задачи, в том числе нестандартные, решены без ошибок и недочетов.

7. Методические материалы

7.1. Перечень основной и дополнительной учебной литературы и иные материалы

Основная литературы

1. Грудева Е. В. Корпусная лингвистика: учебное пособие. — 4-е изд., стер. — Москва: ФЛИНТА, 2025. — 165 с. — ISBN 978-5-9765-1497-3.

Дополнительная литература:

1. Тайлакова Д. Н. Технология моделирования и создания электронного учебника // Молодой ученый. — 2013. — № 5. — С. 772–775.
2. Старостина С. А. Информационные технологии в филологии: учебное пособие / С. А. Старостина; Тамбовский государственный университет имени Г. Р. Державина. — Тамбов: Издательский дом "Державинский", 2022
3. Гуслякова А. В. Информационные технологии и лингвистика XXI века [Текст] : учебное пособие / А. В. Гуслякова ; М-во образования и науки Российской Федерации, Федеральное гос. бюджетное образовательное учреждение высш. образования Московский пед. гос. ун-т. - Москва : МГПУ, 2016. - 95 с. : ил.; 20 см.; ISBN 978-5-4263-0398-0

7.2. Перечень лицензионного программного обеспечения

8. Ресурсное обеспечение

8.1. Перечень профессиональных баз данных и информационных справочных систем:

Электронно-библиотечные системы		
ЭБС Юрайт	https://urait.ru	После регистрации с компьютера Института – доступ с любого компьютера, подключенного к сети Интернет.

8.2. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»:

Справочно-информационный портал ГРАМОТА.РУ – русский язык для	http://gramota.ru/	Свободный доступ

всех		
«ИНФРА–М»	http://znanium.com	Свободный доступ
BOOK.ru	https://www.book.ru	После регистрации с компьютера Института – доступ с любого компьютера, подключенного к сети Интернет
Научная электронная библиотека РИНЦ	https://elibrary.ru/	Свободный доступ
Научная электронная библиотека «КиберЛенинка»	https://cyberleninka.ru/	Свободный доступ

9. Язык преподавания: русский

10. Разработчики программыЛунькова Е. С., д.филол. наук, профессор кафедры ОиРЯ ФГБОУ ВО «Государственный институт русского языка им. А. С. Пушкина»