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

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Наименование дисциплины (модуля):

Корпусные исследования русского языка

Б1.В.08.06

Уровень высшего образования:

бакалавриат

Направление подготовки:

45.03.01 Филология

Направленность (профиль) образования:

Русский язык и литература

Форма обучения:

очная

Рабочая программа дисциплины (модуля) разработана в соответствии с федеральным государственным образовательным стандартом высшего образования по направлению подготовки 45.03.01 «Филология», утвержденным приказом Минобрнауки России № 986 от 12.08.2020.

1. Место дисциплины (модуля) в структуре ОПОП:

В соответствии с учебным планом бакалавриата 45.03.01 "Филология" по специальности "Русский язык и литература" дисциплина Б1.В.08.06 «Корпусные исследования русского языка» относится к части, формируемой участниками образовательных отношений, и изучается в 7 семестре.

2. Цель и задачи освоения дисциплины (модуля)

Целью изучения курса «Корпусные исследования русского языка» является развитие теоретических знаний студентов в области прикладной лингвистики и формировании навыков применения специализированных электронных корпусов текстов для решения научно-исследовательских задач. Эта цель конкретизируется в следующих задачах: формирование представлений о концепциях корпусной лингвистики, её месте в ряду лингвистических дисциплин и значении в рамках доказательной (эмпирической) лингвистики; изучение синтаксиса специализированных языков поисковых запросов, приемов алгоритмизации лингвистических задач и овладение навыками работы с корпусами русского языка; развитие навыков сбора, машинной обработки текстов, создания конкордансов и частотных словарей, а также применение корпусных данных в лексикографии и исследовании различных регистров 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 теоретических и прикладных задач	ПК-1.1 Выделяет и систематизирует значимые аспекты описания лексических единиц в соответствии с целями исследования и типом материала	Знать: основные понятия корпусной лингвистики: лексическая единица (лемма, вокабула, словоформа), коллокация, валентность, семантические роли, прагматические маркеры, коннотации. Уметь: формулировать цели описания единиц применительно к конкретному языковому корпусу и типу исследования. Владеть: методами количественного и качественного анализа лексики, извлеченной из корпуса
	ПК-1.2. Применяет цифровые и корпусные	Знать: принципы разметки

	технологии для работы с лексическими данными ПК-1.3. Представляет результаты разноаспектного описания лексических единиц в формате, пригодном для разных пользователей	(аннотирования) в корпусах, использование тегов для морфологических (*POS*-теггинг), семантических и просодических характеристик. Уметь: анализировать схемы разметки в различных подкорпусах. Владеть: применять специализированное программное обеспечение для извлечения информации из языкового корпуса Знать: Методы извлечения и представления языковой информации: конкорданс, частотные словари, тезаурусы, n-граммы. Уметь: проводить анализ семантической и прагматической разметки текстовых фрагментов, выделяя значимые аспекты. Владеть: навыками проектирования и ведения глоссариев.

4. Объем дисциплины (модуля) _3_з.е., в том числе __32__академических часа на контактную работу обучающихся с преподавателем; ___40__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. **Введение в корпусную лингвистику: объект, предмет и методология.** Определение корпуса, его отличие от простой текстовой коллекции, роль корпусов в		2	2			4		4	8

современной науке о языке, соотношение с компьютерной лингвистикой и обработкой естественного языка.									
Тема 2 **Типология и архитектура языковых корпусов.** Классификация по жанру (монологические, диалогические), модальности (письменные, устные), синхронии (синхронные, диахронные), сбалансированности (репрезентативные, мониторыные) и языку (одноязычные, многоязычные, параллельные).		2	2			4		4	8
Тема 3 **Национальный корпус и принципы его создания.**		2	2			4		4	8

Требования к репрезентативности и сбалансированности, отбор текстов, формирование подкорпусов, юридическая и этическая сторона сбора данных на примере Национального корпуса русского языка (НКРЯ).									
Тема 4. **Метатекстовая и лингвистическая разметка.** Структура текстовых метаданных (автор, дата, жанр, источник), виды аннотирования: морфологическая (POS-теггинг) разметка, синтаксическая (деревья зависимостей и непосредственных составляющих), семантическая и просодическая		2	2			4		4	8

разметка.									
Тема 5. **Технологии создания и обработки корпусов.** Процессы токенизации, лемматизации и стемминга, форматы представления данных, инструменты для разметки и автоматизация очистки данных.		2	2			4		4	8
Тема 6. **Корпусный инструментарий и языки запросов.** Поисковые системы корпусов: синтаксис запросов, использование фильтров, метаданных, регулярные выражения, работа с конкордансерами и генерация частотных списков.		2	2			4		6	10
Тема 7. **Корпусные методы**		2	2			4		8	10

лексикографии и анализа лексики. Изучение коллокаций, статистических мер ассоциации слов, построение тезаурусов, анализ частотности слов и выражений в СМИ, выявление неологизмов и архаизмов на основе исторических корпусов									
Тема 8. **Корпусный анализ дискурса и прикладные задачи.** Использование корпусов для решения прикладных задач: стилистический анализ, перевод, обучение языку, выявление языковых изменений, анализ речевых ошибок, а также междисциплинарные направления (стилометрия, когнитивные		2	2			4		6	10

исследования)**.**									
Промежуточная аттестация *(указывается форма проведения)*	*экзамен*								36
Итого		16	16			32		40	108

6. Фонд оценочных материалов по дисциплине (модулю)

Практические задания для текущего контроля

Программа курса «Корпусная лингвистика» для магистрантов предполагает баланс между теоретическим пониманием методов и практическими навыками работы с данными. Задания разделены на четыре тематических блока, от базовых операций до самостоятельного исследовательского проектирования.

Блок 1. Основы работы с корпусами и разметкой

1. **Сравнительный анализ интерфейсов.** Сравнить поисковые возможности и структуру метаданных двух корпусов (например, НКРЯ и Aranea для русского языка). Оценить, какие типы лингвистических и экстралингвистических запросов доступны в каждом из них.
2. **Оценка качества разметки.** Выбрать в корпусе фрагмент текста (100–150 токенов) со сложной синтаксической структурой. Сопоставить автоматическую разметку (морфологическую и синтаксическую) с ручным анализом, зафиксировать типы систематических ошибок парсера.
3. **Создание мини-корпуса.** Собрать из открытых источников (например, с сайта «Проза.ру» или из новостных архивов) 50 текстов по заданному критерию (жанр, год, автор). Очистить их от шума, привести к единому формату (TXT) и составить файл метаданных в формате TSV.

4. **Метаданные и репрезентативность.** Проанализировать состав подкорпуса устной речи НКРЯ. Определить, какие социальные и демографические группы в нем не представлены, и обосновать, как это искажает результаты исследования разговорных маркеров.
5. **Нормализация и лемматизация.** Написать скрипт (на Python) или использовать инструменты командной строки для лемматизации и очистки небольшого текстового файла от знаков препинания и стоп-слов. Сравнить результат работы двух разных лемматизаторов (например, PyMorphy2 и UDPipe) на омонимичных формах.

Блок 2. Частотный анализ и коллокации

6. **Ключевые слова (Keywords).** Используя Sketch Engine или AntConc, составить список ключевых слов (keywords) для корпуса текстов одного автора по сравнению с эталонным референтным корпусом. Интерпретировать топ-10 слов с точки зрения идиостиля.
7. **Динамика частотности.** Построить график частотности трех общественно-политических лексем (например, *суверенитет*, *цифровизация*, *токсичный*) в основном корпусе НКРЯ по десятилетиям с 1950 по 2020 год. Объяснить пики и спады частотности историческим контекстом.
8. **Анализ коллокаций.** Выявить коллокации для многозначного слова (например, *камень* или *ключ*) в основном корпусе. Разделить их на смысловые группы, соответствующие разным лексико-семантическим вариантам слова.
9. **Статистическая значимость.** Рассчитать для выбранной коллокации (например, *круглый дурак*) статистические показатели ассоциации (MI, LogDice, T-score). Объяснить, почему для разных типов лингвистических задач (поиск устойчивых выражений vs. анализ словосочетаний) предпочтительны разные меры.
10. **Семантические примитивы.** Составить частотный список глаголов движения в корпусе русской прозы XIX века. Сопоставить его со списком из корпуса текстов XXI века и описать произошедшие семантические сдвиги.

Блок 3. Синтаксис, семантика и прагматика

11. **Синтаксические конструкции.** В синтаксически размеченном корпусе (например, SynTagRus) найти все реализации дативного субъекта. Описать их частотность, лексическое наполнение и типичные предикаты.
12. **Диахронический сдвиг.** Отследить изменение управления глагола (например, *управлять* + творительный падеж vs. *управлять* + винительный падеж) на материале корпуса за последние 100 лет.
13. **Прагматические маркеры.** Составить список контекстных фильтров для поиска дискурсивных маркеров (*ну*, *типа*, *как бы*) в устном подкорпусе. Оценить их распределение в речи разных возрастных групп.
14. **Аспектуальная семантика.** Проанализировать конкуренцию видовременных форм глагола в контексте обстоятельств типа *вдруг*, *сразу*, *мгновенно*. Построить распределение частот совершенного и несовершенного вида в этих позициях.
15. **Корпусный анализ метафор.** Найти контексты со словом *яд* в прямом и метафорическом значениях. Описать семантические поля, в которые встраивается слово при метафоризации, используя инструмент Word Sketch.

Блок 4. Прикладные задачи и исследовательское проектирование

16. **Авторизация текста.** Разработать и протестировать на мини-корпусе (10 авторов, по 5 текстов на каждого) алгоритм атрибуции текста по 5–7 дистинктивным признакам (длина предложения, коэффициент лексического разнообразия TTR, частотность служебных слов).
17. **Сентимент-анализ.** Провести лексикографическую разметку полярности (позитив/негатив/нейтрально) для 200 слов, характерных для жанра кинорецензий. Оценить точность автоматической классификации тональности на тестовой выборке корпуса.
18. **Лингвистическая экспертиза.** Сформулировать корпусную стратегию для решения прикладной задачи (например, доказательство плагиата или установление авторства анонимного письма). Описать, какие подкорпуса и статистические критерии будут использованы.
19. **Проектирование исследовательского корпуса.** Написать развернутый паспорт исследовательского корпуса (2–3 тысячи словоупотреблений) для курсовой работы. Обосновать выбор текстов, принципы разметки (аннотации речевых актов, эмоций или синтаксиса), структуру XML/TEI и репрезентативность выборки.

20. **Визуализация данных.** На основе данных, полученных в задании №7 (динамика частотности), построить интерактивную визуализацию (например, в Python с использованием Plotly или библиотеки D3.js), добавив на временную шкалу маркеры ключевых исторических событий, повлиявших на употребление лексем.

Критерии оценки практических заданий

Общие принципы начисления баллов (для всех заданий)

- **Своевременность сдачи:** 0 баллов штрафа при сдаче в срок; -10 баллов при задержке до 3 дней; -20 баллов при задержке более 3 дней.
- **Академическая добросовестность:** обнаружение плагиата или автоматической генерации текста без авторской переработки — 0 баллов за задание.
- **Формат сдачи:** несоблюдение требований к оформлению (отсутствие комментариев в коде, неверный формат метаданных) — штраф от -5 до -15 баллов.

Блок 1. Основы работы с корпусами и разметкой (Задания 1–5) *Базовый уровень (0–60 баллов):* задание выполнено, файлы или скриншоты интерфейсов приложены, даны поверхностные описания. *Продвинутый уровень (61–85 баллов):* проведен корректный анализ, выявлены 3–5 ключевых различий или типов ошибок, приведены точные количественные примеры (5–10 контекстов). *Экспертный уровень (86–100 баллов):* глубокая аналитика с опорой на теорию корпусной лингвистики (репрезентативность, баланс), предложены способы устранения выявленных проблем разметки, для задания №3 составлен валидный файл метаданных в формате TSV или JSON, полностью готовый к загрузке в корпус-менеджер.

Блок 2. Частотный анализ и коллокации (Задания 6–10) *Базовый уровень (0–60 баллов):* списки слов или графиков извлечены, но в отчете отсутствует статистическая интерпретация и проверка значимости. *Продвинутый уровень (61–85 баллов):* приведены верные статистические показатели (MI, LogDice), выборка релевантна, дано лингвистически обоснованное объяснение результатов, графики наглядны и снабжены осями. *Экспертный уровень (86–100 баллов):* проведен сравнительный анализ разных статистических мер (например, почему для задания №9 выбрана именно мера LogDice), выявлены скрытые факторы (например, влияние

омонимии на коллокации), для задания №8 предложена собственная семантическая классификация, подтвержденная количественными данными.

Блок 3. Синтаксис, семантика и прагматика (Задания 11–15) *Базовый уровень (0–60 баллов):* найдены единичные примеры конструкций, фильтрация контекстов неполная, выводы носят характер общих рассуждений. *Продвинутый уровень (61–85 баллов):* выборка составляет не менее 50 релевантных примеров, фильтры для поиска (CQL или интерфейсные) описаны точно, распределение данных представлено в виде таблиц. *Экспертный уровень (86–100 баллов):* проведен количественный анализ распределения (проценты, абсолютные частоты), выявлена статистически значимая корреляция (например, между возрастом говорящих и частотой маркеров в задании №13), для задания №12 построена диахроническая модель изменения конструкции.

Блок 4. Прикладные задачи и исследовательское проектирование (Задания 16–20) *Базовый уровень (0–60 баллов):* алгоритм или стратегия описаны только на вербальном уровне, код не предоставлен или не запускается, признаки для анализа выбраны произвольно. *Продвинутый уровень (61–85 баллов):* код рабочий и прокомментированный, проведена оценка точности (accuracy/precision/recall), для задания №16 точность классификации превышает 70%, паспорт корпуса в задании №19 содержит все обязательные разделы. *Экспертный уровень (86–100 баллов):* для заданий №16 и №17 проведена кросс-валидация, признаки для классификации статистически обоснованы; в задании №18 предложенная корпусная стратегия полностью исключает ложноположительные результаты; в задании №19 паспорт корпуса соответствует стандарту TEI, выборка строго сбалансирована по жанрам и времени, а в задании №20 визуализация интерактивна и позволяет фильтровать данные по дополнительным параметрам.

Итоговая шкала перевода баллов в оценку за отдельное задание

- 90–100 баллов: Отлично
- 75–89 баллов: Хорошо
- 60–74 балла: Удовлетворительно
- 0–59 баллов: Неудовлетворительно (требуется пересдача)

Задания по курсу для самостоятельной работы студентов

Блок 1. Основы работы с корпусами и разметкой

1. **Сравнение типов корпусов.** Найти в открытом доступе (например, на сайтах Национального корпуса русского языка (НКРЯ) или Sketch Engine) два корпуса одного языка (например, синхронический и диахронический, или письменный и устный). Составить сравнительную таблицу их параметров: объем, хронологические рамки, жанровая разметка, наличие морфологической разметки.
2. **Чтение документации.** Изучить руководство пользователя (User Guide) для корпуса НКРЯ или British National Corpus (BNC). Написать краткую инструкцию (500–700 знаков) для одногруппника: как найти примеры употребления слова в текстах конкретного автора.
3. **Типы лингвистической разметки.** Взять один и тот же текст (например, басню И.А. Крылова) в трех видах: без разметки, с метатекстовой разметкой (автор, год, жанр) и с морфологической. Описать, какие новые типы лингвистических вопросов можно задать к каждому из этих текстов.
4. **Критерии оценки корпуса.** Проанализировать один из известных корпусов (например, Корпусы Славянских Языков или Corpus of Contemporary American English — COCA) по критериям репрезентативности и сбалансированности. Найти на сайте описание принципов его наполнения и выявить возможные лакуны (отсутствие редких диалектов, узких технических жанров).

Блок 2. Поисковые запросы и работа с конкордансами

5. **Построение конкорданса.** Составить конкорданс (список всех употреблений в контексте) для слова «зеленый» в основном корпусе НКРЯ. Выписать 10 примеров, где слово употреблено в переносном значении, и 5 примеров, где оно является частью имени собственного.
6. **Лексико-грамматические шаблоны.** Используя CQL (Corpus Query Language) или расширенный поиск, найти в корпусе все случаи употребления глаголов движения (*идти, ехать, бежать*) с предлогом «через». Проанализировать, с какими падежами существительных они сочетаются.
7. **Омонимичные формы.** Составить поисковый запрос, который находит слово «печь» только в значении глагола (инфинитив), исключая существительное. Описать логику построения этого запроса (использование тегов или отрицания).

8. **Коллокации.** Найти 10 наиболее частотных коллокатов (слов, встречающихся рядом) для прилагательного «острый» в радиусе двух слов. Определить, какие из них образуют устойчивые словосочетания (коллокации), а какие — свободные сочетания.
9. **Семантические примитивы.** Используя семантическую разметку НКРЯ (поиск по семантическим категориям), найти все глаголы перемещения, которые сочетаются с неодушевленными субъектами (например, *течение несет*, *поезд идет*).

Блок 3. Частотность и статистика

10. **Частотные словари.** Сравнить частотные списки первых 50 знаменательных слов в подкорпусе «Социальные сети» и в подкорпусе «Русская проза XIX века». Выявить 10 слов, которые входят в топ одного списка, но отсутствуют в другом, и предположить причины этой разницы.
11. **Ключевые слова (Keywords).** Взять текст современного новостного интервью и текст интервью 1980-х годов (из газетного корпуса). С помощью инструментов сравнения частот (например, в AntConc или онлайн-сервисах) выявить ключевые слова, характерные для современного текста.
12. **Закон Ципфа.** Построить график частотности слов для небольшого художественного текста (около 2000 слов) в Excel или Google Таблицах. Проверить, насколько распределение частот в тексте соответствует закону Ципфа, и отметить отклонения.
13. **Диахронический сдвиг.** Отследить изменение частотности слов «интернет», «гаджет» и «самоизоляция» по десятилетиям (с 1990 по 2020 год) в основном корпусе. Построить график изменения частот и описать точки резкого роста.

Блок 4. Корпусные менеджеры и самостоятельная разметка

14. **Работа с AntConc.** Скачать корпусный менеджер AntConc. Загрузить в него папку с 5–7 текстами (например, статьями одного автора). Найти с помощью инструмента *Clusters* (N-граммы) наиболее частотные четырехсловные сочетания во всем наборе текстов.
15. **Создание учебного корпуса.** Собрать мини-корпус объемом 10–15 тысяч словоупотреблений на узкую тему (например, «отзывы на бытовую технику» или «рецепты из кулинарных блогов»). Сохранить тексты в формате .txt, очистить от лишних символов и составить файл с метаданными (автор, дата, источник).

16. **Регулярные выражения.** Написать регулярное выражение (Regex) для поиска всех слов, начинающихся на «авиа-» или «авто-», и всех глаголов в форме прошедшего времени с возвратным суффиксом «-сь». Протестировать выражение на текстовом фрагменте.
17. **Простейшая ручная разметка.** Взять отрывок интервью (50 реплик). Разметить его вручную в формате XML или CSV: выделить реплики, атрибут говорящего, паузы и слова-паразиты (*ну, вот, значит*). Написать, какие трудности возникли при определении границ хезитационных конструкций.

Блок 5. Прикладные и исследовательские задачи

18. **Перевод и двуязычный корпус.** Используя английский корпус (например, OPUS или TenTen), найти параллельные контексты для многозначного русского слова «право» (в значении «закон» и в значении «моральное основание»). Сравнить, какими английскими эквивалентами (*law, right, act*) оно передается.
19. **Корпусный профиль слова.** Составить «портрет» глагола «достать» на основе данных корпуса. Описать его валентности (кто достает, что достает, откуда/от кого), типичные объекты (билет, документы, деньги) и стилистическую окраску (разговорное, просторечное).
20. **Проект мини-исследования.** Сформулировать исследовательский вопрос (например, «Как изменилось употребление этикетных формул приветствия в русском языке за последние 30 лет?»). Составить пошаговый план его проверки с помощью корпусов: какие подкорпуса выбрать, какие запросы составить, какую статистику собрать и как визуализировать результат.

Рекомендации по оценке:

- Задания 1–4, 15 проверяют понимание теории и документации.
- Задания 5–9, 13, 14, 16, 17 проверяют технические навыки работы с интерфейсами и инструментами.
- Задания 10–12, 18, 19, 20 требуют аналитического подхода и интерпретации данных.

Темы докладов

- **Параметризация метаданных в корпусе:** от базовых тегов (автор, дата, жанр) к многоуровневым аннотациям (социолингвистические переменные, прагматические пресуппозиции) и их влиянию на результаты поиска.
- **Автоматическое снятие омонимии (WSD) на материале национального корпуса:** сравнение алгоритмов, основанных на словарных определениях, и нейросетевых подходов.
- **Коллокации и коллигации в корпусном анализе:** статистические меры ассоциации (MI-score, T-score, LogDice) и их применимость для выявления устойчивых сочетаний в русском языке.
- **Диахронические корпуса и отслеживание языковых изменений:** корпусный анализ неологизмов, архаизмов и грамматикализации конструкций на материале исторического подкорпуса.
- **Корпусные методы в лексикографии:** от словарных помет к данным реального употребления (на примере создания или верификации статей толкового или двуязычного словаря).
- **Синтаксические парсеры и дерево зависимостей:** оценка качества автоматического синтаксического анализа на материале размеченного корпуса сложных предложений.
- **Корпусный анализ дискурсивных маркеров:** выявление и количественное описание прагматических функций частиц и коннекторов (например, «ведь», «же», «мол») в устной и письменной речи.
- **Создание узкоспециализированных корпусов:** методология сбора, очистки и аннотирования текстов профессиональной или субкультурной коммуникации (например, корпус медицинской выписки, геймерского сленга или судебной речи).
- **Корпусные исследования перевода:** анализ переводческих стратегий, явлений эксплицитации и имплицитации на материале параллельных корпусов (англо-русского, русско-китайского).
- **Субкорпусы устной речи:** проблемы транскрибирования, разметки паралингвистических явлений (паузы, вздохи, смех) и их учет в синтаксическом и просодическом анализе.
- **Применение дистрибутивной семантики и векторных моделей к корпусным данным:** выявление семантических полей и скрытых семантических сдвигов.
- **Корпусный анализ медиадискурса:** отслеживание динамики частотности идеологем, метафор и фреймов в новостных текстах за определенный период.
- **Методы оценки репрезентативности и баланса корпуса:** статистические критерии и практические проблемы при формировании сбалансированного подкорпуса.

- **Корпусные исследования усвоения второго языка:** анализ межъязыковой интерференции и типичных ошибок на материале learner corpora (например, RICLE — Russian Interlanguage Corpus of Learners of English).
- **Критическая корпусная лингвистика:** этические аспекты, проблема репрезентативности маргинальных социолектов и предвзятость (bias) в обучающих данных.

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

Оценка качества учебных достижений обучающихся по дисциплине осуществляется в виде текущего контроля успеваемости и промежуточной аттестации.

Требования к структуре и содержанию доклада:

1) регламент выступления - 10 минут; наличие текста доклада и компьютерной презентации (презентация и текст доклада не должны дублировать друг друга);

2) использование самостоятельно подобранного текстового материала и его анализа в аспекте, определяемом темой доклада;

3) наличие учебного задания для студентов группы по теме доклада.

Критерии оценки доклада

Критерии	Показатели
Наличие необходимых элементов	Есть письменный текст доклада, компьютерная презентация, не дублирующие друг друга, студент выступил с докладом
Соблюдение требований к	В докладе используется самостоятельно

структуре и содержанию доклада	подобранный текстовый материал, есть его анализ и учебное задание

Шкала оценивания доклада (в баллах):

- все требования соблюдены, необходимые элементы есть, выполнено качественно: 15 б.

- все требования соблюдены, необходимые элементы есть, выполнено качественно, но презентация и текст доклада дублируют друга друга – 10 б

- не все требования соблюдены, не все необходимые элементы в наличии (нет текста доклада/презентации/нет самостоятельно подобранного текстового материала/задания – 5-7 баллов).

- доклад не представлен – 0 баллов.

Вопросы к экзамену по дисциплине «Корпусная лингвистика»:

Раздел 1. Теоретические основы и история

1. Корпусная лингвистика как научная дисциплина: объект, предмет, цели и задачи.
2. Соотношение корпусной лингвистики с компьютерной, прикладной и математической лингвистикой.
3. История возникновения корпусной лингвистики: корпус Брауна, LOB, British National Corpus (BNC).
4. Роль отечественных лингвистов (Ю. Н. Караулов, В. А. Плунгян, Е. В. Рахилина) в развитии корпусных исследований.
5. Понятие репрезентативности и сбалансированности корпуса.
6. Парадигма «данные — теория» в корпусной лингвистике: индуктивный и дедуктивный подходы.
7. Корпус как эмпирическая база современной лингвистики.
8. Влияние корпусных данных на развитие лексикографии и грамматических описаний.

Раздел 2. Типология и архитектура корпусов

9. Классификация корпусов по форме представления (текстовые, мультимодальные, речевые).

10. Классификация по охвату языка (монолингвальные, параллельные, многоязычные, сопоставительные).
11. Динамические (мониторные) и статические корпуса.
12. Аннотированные и «сырые» (plain text) корпуса.
13. Специализированные корпуса: диалектные, исторические, поэтические.
14. Параллельные и многоязычные корпуса: принципы построения и выравнивания.
15. Национальные корпуса (на примере НКРЯ): структура, состав, принципы формирования.
16. Глобальные корпусы веб-текстов (WaCky, ukWaC): особенности сбора и очистки данных.
17. Диахронические корпуса и их роль в исторической лингвистике.
18. Корпус устной речи: специфика транскрипции и разметки.

Раздел 3. Аннотирование и разметка

19. Уровни лингвистической разметки: метатекстовая, морфологическая, синтаксическая, семантическая.
20. Принципы морфологической разметки (POS-tagging). Форматы разметки (XML, TEI).
21. Синтаксическая разметка: деревья зависимостей и деревьев составляющих.
22. Семантическая и дискурсная разметка: современные вызовы и стандарты.
23. Автоматические инструменты аннотирования (парсеры, тэггеры) и проблема точности.
24. Проблема разрешения омонимии при автоматической разметке.
25. Верификация и исправление ошибок автоматической разметки.
26. Стандарты кодировки текста (Unicode, UTF-8) и их значение для корпусов.

Раздел 4. Инструментарий и технические аспекты

27. Конкорданс: определение, структура конкордансной строки, виды сортировки.
28. Частотные списки (frequency lists) и их лингвистическая интерпретация.
29. Инструменты для работы с корпусами: AntConc, Sketch Engine, LancsBox, корпуса НКРЯ.
30. Регулярные выражения (Regular Expressions) как язык запросов к корпусу.
31. Языки запросов к корпусам (CQP, Poliqarp, синтаксис НКРЯ).
32. Архитектура информационно-поисковой системы лингвистического корпуса.
33. Методы сбора данных для корпусов: скрейпинг, парсинг, правовые аспекты.
34. Предобработка текста: токенизация, лемматизация, нормализация.
35. Хранение и индексация корпусных данных (реляционные базы данных, поисковые движки).

Раздел 5. Практическое применение корпусов

36. Изучение лексической семантики и многозначности на корпусном материале.
37. Коллокации и коллигации: методы выявления и статистические меры (MI-score, T-score).
38. Корпусный подход в грамматических исследованиях (грамматика конструкций).
39. Изучение вариативности и языковой нормы с помощью корпусов.
40. Диахронические исследования на материале динамических корпусов.
41. Корпусные методы в лексикографии (создание словарей на основе данных).
42. Использование корпусов в преподавании иностранных языков (Data-Driven Learning).
43. Корпусный анализ дискурса и медиалингвистики.
44. Применение корпусов в судебной лингвистике и автороведческой экспертизе.
45. Корпусные исследования перевода и контрастивная лингвистика.
46. Корпусная прагматика: изучение речевых актов и модальности.
47. Корпусный анализ фразеологии и идиоматики.

Раздел 6. Создание корпусов и статистика

48. Основные этапы проектирования и создания лингвистического корпуса.
49. Понятие статистической значимости в корпусных исследованиях. Ошибки интерпретации (например, парадокс Симпсона).
50. Этические и правовые вопросы создания корпусов: авторское право, анонимизация персональных данных, согласие информантов.

7. Методические материалы

7.1. Перечень основной и дополнительной учебной литературы и иные материалы

Основная литература:

1. Шунейко А. А. Корпусная лингвистика : учебник для вузов / А. А. Шунейко. — Москва : Издательство Юрайт, 2026. — 222 с. — (Высшее образование). — ISBN 978-5-534-13603-6. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/588309 (дата обращения: 10.10.2026).

Дополнительная литература:

1. Захаров В. И., Богданова С. Ю. Корпусная лингвистика: учебник. 3-е изд., перераб. СПб.: Изд-во С.-Петерб.ун-та, 2020. - 234с. ISBN: 978-5-288-05997-1

2. Моделирование в корпусной лингвистике: специализированные корпусы русского языка / В.П. Захаров, И.В. Азарова, О.А.Митрофанова, А.М.Попов, М.В. Хохлова; отв.ред. В.П.Захаров. – СПб.: Изд-во С.-Петерб.ун-та, 2019. - 208с. – ISBN: 978-5-288-05902-5

Периодические издания:

1. Вопросы языкознания: Научный журнал. - https://vja.ruslang.ru/ru/archive (Архив журнала с открытым доступом в Интернет).

7.2. Перечень лицензионного программного обеспечения

8. Ресурсное обеспечение

8.1. Перечень профессиональных баз данных и информационных справочных систем:

Наименование	Ссылка на ресурс	Доступность (свободный доступ / ограниченный доступ)
1	***2***	***3***
Информационно-справочные системы		
Библиографические базы данных по общественным наукам ИНИОН	http://inion.ru/	Свободный доступ
Методология: словарь системы основных	http://www.anovikov.ru/dict/met_sl.htm	Свободный доступ

понятий. Сайт академика РАО Новикова А.М.		

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

Наименование	Ссылка на ресурс	Доступность (свободный доступ / ограниченный доступ)
1	***2***	***3***
Электронно-библиотечные системы		
ЭБС Юрайт	https://urait.ru	После регистрации с компьютера Института – доступ с любого компьютера, подключенного к сети Интернет.
Профессиональные базы данных		
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 программы

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