МИНОБРНАУКИ РОССИИ

Федеральное государственное автономное образовательное учреждение высшего образования "Пермский государственный национальный исследовательский университет"

Авторы-составители: **Хорошева Наталья Владимировна**
Шустова Светлана Викторовна

Рабочая программа дисциплины

КОРПУСНЫЕ ТЕХНОЛОГИИ В ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

Код УМК 101016

Утверждено
Протокол №4
от «16» января 2024 г.

Пермь, 2024

1. Наименование дисциплины

Корпусные технологии в лингвистических исследованиях

2. Место дисциплины в структуре образовательной программы

Дисциплина входит в вариативную часть Блока « М.1 » образовательной программы по направлениям подготовки (специальностям):

Направление подготовки: **45.04.02** Лингвистика

направленность Цифровая лингвистика и перевод

3. Планируемые результаты обучения по дисциплине

В результате освоения дисциплины **Корпусные технологии в лингвистических исследованиях** у обучающегося должны быть сформированы следующие компетенции:

45.04.02 Лингвистика (направленность : Цифровая лингвистика и перевод)

ОПК.2 Способен учитывать в практической деятельности специфику иноязычной научной картины мира и научного дискурса в русском и изучаемом иностранном языках

Индикаторы

ОПК.2.2 Осуществляет анализ языковых изменений в условиях расширения глобализационных процессов

ОПК.7 Способен работать с основными информационно-поисковыми и экспертными системами, системами представления знаний и обработки вербальной информации

Индикаторы

ОПК.7.1 Использует информационно-поисковые системы открытого доступа для проведения экспертного анализа дискурса, средства автоматизированной обработки вербальной информации

4. Объем и содержание дисциплины

Направление подготовки	45.04.02 Лингвистика (направленность: Цифровая лингвистика и перевод)
форма обучения	очная
№№ триместров, выделенных для изучения дисциплины	1
Объем дисциплины (з.е.)	3
Объем дисциплины (ак.час.)	108
Контактная работа с преподавателем (ак.час.), в том числе:	36
Проведение лекционных занятий	24
Проведение практических занятий, семинаров	12
Самостоятельная работа (ак.час.)	72
Формы текущего контроля	Итоговое контрольное мероприятие (1) Письменное контрольное мероприятие (2)
Формы промежуточной аттестации	Зачет (1 триместр)

5. Аннотированное описание содержания разделов и тем дисциплины

Корпусные технологии в лингвистических исследованиях

Корпусная лингвистика - это раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования языков. Главная особенность корпуса состоит в возможности построить конкорданс - список всех употреблений исследуемого слова в контексте со ссылками на источник. С помощью корпуса можно получить статистические данные о речевых и языковых единицах, а также отследить их частотность, грамматические категории, лексемы, понаблюдать за сочетаемостью тех или иных лексических элементов. Для написания разнообразных словарей также могут быть использованы корпусы.

Тема 1. Корпус как современный лингвистический ресурс

Корпус как особый лингвистический ресурс. Определение корпуса. Проблема репрезентативности и полноты корпуса. Структура корпуса: корпус данных, корпус-менеджер. Корпусный (эмпирический) подход в сравнении с лингвистикой Н.А.Хомского. Единицы хранения корпуса данных. Требования к корпусу текстов. Корпус как информационно-справочная система.

Тема 2. Корпусная лингвистика в аспекте интердисциплинарности

Корпусная лингвистика в аспекте интердисциплинарности. Корпус и дискурс. Корпусная лингвистика как раздел языкознания. Место корпусной лингвистики в системе научных знаний о языке. История и перспективы корпусной лингвистики. Первые корпусы: The Brown Corpus, The Lancaster-Oslo-Bergen Corpus. Мегакорпусы, их особенности: British National Corpus; Чешский национальный корпус, Национальный корпус русского языка. Корпусная лингвистика и компьютерная лингвистика. Корпусная лексикография. Применение корпусов в области перевода и переводоведения. Основы правового регулирования сферы корпусной лингвистики. Лицензии Интернета. Законодательство РБ об авторском праве и о защите информации.

Тема 3. Типология корпусов

Типология корпусов. Типы лингвистических корпусов и основания их выделения. Устные и письменные корпусы. Смешанные корпусы. Одноязычные корпусы. Размеченные корпусы. Неразмеченные корпусы. Параллельный корпус: выравнивание, области применения. Диахрония и синхрония в корпусной лингвистике. Национальные корпусы. Лингвистические задачи, решаемые с применением корпусных технологий. Конкорданс. Машинный перевод. Корпусные технологии в лексикографии. Корпусные технологии в обучении иностранным языкам. Использование корпусов в информационном поиске. Международные проекты корпусной лингвистики. Стандарты TEI/CES/XCES, XML/SGML, EAGLES.

Тема 4. Корпусы современных славянских языков

Корпусы современных славянских языков.

- Корпус текстов украинского языка Лаборатории компьютерной лингвистики Киевского университета
- Открытые корпуса украинского языка
- Белорусский N-корпус
- Белорусский библейский корпус
- Экспериментальный корпус белорусского языка
- Corpus Albaruthenicum — корпус научных белорусских текстов
- Национальный корпус польского языка
- Польско-русский параллельный корпус
- Чешский национальный корпус
- Словацкий национальный корпус
- FIDA — словенский корпус

• GOS — устный словенский корпус
• GRALIS — параллельный корпус с участием сербских, хорватских и боснийских текстов Грацского университета
• Черногорско-английский параллельный корпус
• Болгарско-русский параллельный корпус

Тема 5. Корпусы германских и романских языков

Корпусы германских и романских языков
• Британский национальный корпус (BNC)
• Британский национальный корпус в версии Марка Дэвиса (BYU-BNC)
• Корпус современного американского английского (COCA)
• Исторический корпус американского английского (COHA)
• WaCKy — большие открытые веб-корпуса английского языка
• Немецкий справочный корпус (DeReKo)
• Банк данных разговорного немецкого (DGD)
• Корпуса немецкого языка на сайте CorpusEye
• KorpusDK: датский корпус
• Банк шведского языка
• Корпуса шведского языка на сайте CorpusEye
• Корпуса норвежского языка на сайте CorpusEye
• База французских текстов FranText
• Корпуса французского языка на сайте CorpusEye
• Лингвистическая база данных функционально эквивалентных фрагментов на материале поливариантного русско-французского корпуса
• Корпуса испанского языка на сайте CorpusEye
• Корпуса письменного итальянского языка CORIS и CODIS
• Корпуса итальянского языка на сайте Corpus Eye
• Корпуса португальского языка на сайте Corpus Eye

Тема 6. Корпусы искусственных языков и многоязычные корпусы

Корпусы искусственных языков и многоязычные корпусы
• Коллекции текстов на малых языках России (доступны для скачивания)
• Описание корпусов уральских языков на сайте Хельсинкского университета
• Обучающий корпус японского языка (доступен для онлайн-поиска)
• Японско-английский параллельный корпус (доступен для онлайн-поиска)
• Тайский корпус HSE
• Корпус эсперанто фонда «Esperantic Studies Foundation»
• Корпус слушаний Европарламента (доступен для скачивания)
• Корпус документов Евросоюза (более 20 языков) (доступен для скачивания)
• ParaSol: параллельный корпус славянских и других языков Бернского университета (бывший Регенсбургский)
• InterCorp: параллельные корпуса Пражского университета
• Universal Dependencies: размеченные в едином формате синтаксические корпуса 47 языков
• Многоязычные корпуса университета Осло
• TITUS — тезаурус материалов по индоевропейским языкам, Франкфурт
• Параллельный корпус русских и французских поэтических текстов первой трети XIX в.

Тема 7. Создание собственного корпуса

Лингвистические исследования на базе корпусов. Исследование частотности лексемы. Анализ распределения частотности употребления лексемы по жанрам. Изучение лексической сочетаемости. Исследование лексической синонимии. Исследование частеречного состава текста (корпусов текстов) при помощи морфологической разметки.

Тема 8. Аннотирование корпуса

Создание собственного корпуса. Планирование корпуса. Отбор источников корпуса. Определение жанрово-тематической структуры корпуса. Сбор и цифровка данных. Разметка текста.

6. Методические указания для обучающихся по освоению дисциплины

Освоение дисциплины требует систематического изучения всех тем в той последовательности, в какой они указаны в рабочей программе.

Основными видами учебной работы являются аудиторные занятия. Их цель - расширить базовые знания обучающихся по осваиваемой дисциплине и систему теоретических ориентиров для последующего более глубокого освоения программного материала в ходе самостоятельной работы. Обучающемуся важно помнить, что контактная работа с преподавателем эффективно помогает ему овладеть программным материалом благодаря расстановке необходимых акцентов и удержанию внимания интонационными модуляциями голоса, а также подключением аудио-визуального механизма восприятия информации.

Самостоятельная работа преследует следующие цели:

– закрепление и совершенствование теоретических знаний, полученных на лекционных занятиях;

- формирование навыков подготовки текстовой составляющей информации учебного и научного назначения для размещения в различных информационных системах;

- совершенствование навыков поиска научных публикаций и образовательных ресурсов, размещенных в сети Интернет;

- самоконтроль освоения программного материала.

Обучающемуся необходимо помнить, что результаты самостоятельной работы контролируются преподавателем во время проведения мероприятий текущего контроля и учитываются при промежуточной аттестации.

Обучающимся с ОВЗ и инвалидов предоставляется возможность выбора форм проведения мероприятий текущего контроля, альтернативных формам, предусмотренным рабочей программой дисциплины. Предусматривается возможность увеличения в пределах 1 академического часа времени, отводимого на выполнение контрольных мероприятий.

Процедура оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья по дисциплине предусматривает предоставление информации в формах, адаптированных к ограничениям их здоровья и восприятия информации.

При проведении текущего контроля применяются оценочные средства, обеспечивающие передачу информации, от обучающегося к преподавателю, с учетом психофизиологических особенностей здоровья обучающихся.

7. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

При самостоятельной работе обучающимся следует использовать:

– конспекты лекций;

– литературу из перечня основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля);

– текст лекций на электронных носителях;

– ресурсы информационно-телекоммуникационной сети "Интернет", необходимые для освоения дисциплины;

– лицензионное и свободно распространяемое программное обеспечение из перечня информационных технологий, используемых при осуществлении образовательного процесса по дисциплине;

– методические указания для обучающихся по освоению дисциплины.

8. Перечень основной и дополнительной учебной литературы

Основная:

1. Захаров, В. П. Корпусная лингвистика : учебник для студентов гуманитарных вузов / В. П. Захаров, С. Ю. Богданова. — Иркутск : Иркутский государственный лингвистический университет, 2011. — 161 с. — ISBN 978-5-88267-316-0. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. http://www.iprbookshop.ru/21088

2. Казарин, Ю. В. Лингвистический анализ текста : учебное пособие для академического бакалавриата / Ю. В. Казарин ; под научной редакцией Л. Г. Бабенко. — 2-е изд. — Москва : Издательство Юрайт, 2019 ; Екатеринбург : Изд-во Урал. ун-та. — 132 с. — (Университеты России). — ISBN 978-5-534-07556-4 (Издательство Юрайт). — ISBN 978-5-7996-1660-1 (Изд-во Урал. ун-та). — Текст : электронный // ЭБС Юрайт [сайт]. https://www.urait.ru/bcode/441460

Дополнительная:

1. Моисеева, И. Ю. История и методология науки. Часть 2 : учебное пособие / И. Ю. Моисеева. — Оренбург : Оренбургский государственный университет, ЭБС АСВ, 2017. — 160 с. — ISBN 978-5-7410-1712-8. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. http://www.iprbookshop.ru/71278.html

2. Гируцкий, А. А. Общее языкознание : учебник / А. А. Гируцкий. — Минск : Вышэйшая школа, 2017. — 240 с. — ISBN 978-985-06-2772-8. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. http://www.iprbookshop.ru/90799.html

9. Перечень ресурсов сети Интернет, необходимых для освоения дисциплины

http://charko.narod.ru/tekst/an4/1.html Национальный корпус русского языка
http://moluch.ru/conf/ped/archive/66/3305/ Информационный образовательный ресурс «Молодой ученый»
wortschatz-uni-leipzig.de Лаборатория корпусной лингвистики Лейпцигского университета
http://www.laurenceanthony.net/software/antconc Корпусный менеджер AntConc
elibrary.ru Научная электронная библиотека

10. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине

Образовательный процесс по дисциплине **Корпусные технологии в лингвистических исследованиях** предполагает использование следующего программного обеспечения и информационных справочных систем:
1) презентационные материалы (слайды по темам лекционных и практических занятий);
2) доступ в режиме on-line в Электронную библиотечную систему (ЭБС)
3) доступ в электронную информационно-образовательную среду университета;
4) интернет-сервисы и электронные ресурсы (поисковые системы, электронная почта, профессиональные тематические чаты и форумы).

Перечень необходимого лицензионного и (или) свободно распространяемого программного обеспечения:
1) офисный пакет приложений (текстовый процессор, программа для подготовки электронных презентаций);
2) программа демонстрации видеоматериалов (проигрыватель);
3) приложение, позволяющее просматривать и воспроизводить медиаконтент PDF-файлов.
Дисциплина не предусматривает использование специального программного обеспечения.

При освоении материала и выполнения заданий по дисциплине рекомендуется использование материалов, размещенных в Личных кабинетах обучающихся ЕТИС ПГНИУ (**student.psu.ru**).

При организации дистанционной работы и проведении занятий в режиме онлайн могут использоваться:
система видеоконференцсвязи на основе платформы BigBlueButton (https://bigbluebutton.org/).
система LMS Moodle (http://e-learn.psu.ru/), которая поддерживает возможность использования текстовых материалов и презентаций, аудио- и видеоконтент, а так же тесты, проверяемые задания, задания для совместной работы.
система тестирования Indigo (https://indigotech.ru/).

11. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Материально-техническая база обеспечивается наличием:
1) Для проведения занятий лекционного и семинарского типов (практических занятий) необходима учебная аудитория, оснащенная специализированной мебелью, демонстрационным оборудованием (проектор, экран, компьютер/ноутбук) с соответствующим программным обеспечением, меловой и (или) маркерной доской.

2) Для проведения мероприятий текущего контроля и промежуточной аттестации необходима учебная аудитория, оснащенная специализированной мебелью, демонстрационным оборудованием (проектор,

экран, компьютер/ноутбук) с соответствующим программным обеспечением, меловой и (или) маркерной доской.

3) Для самостоятельной работы используются помещения Научной библиотеки ПГНИУ, оснащенные компьютерной техникой и обеспечивающие доступ к информационно-телекоммуникационной сети «Интернет» и в электронную информационно-образовательную среду.

Помещения научной библиотеки ПГНИУ для обеспечения самостоятельной работы обучающихся:

1. Научно-библиографический отдел, корп.1, ауд. 142. Оборудован 3 персональными компьютера с доступом к локальной и глобальной компьютерным сетям.

2. Читальный зал гуманитарной литературы, корп. 2, ауд. 418. Оборудован 7 персональными компьютерами с доступом к локальной и глобальной компьютерным сетям.

3. Читальный зал естественной литературы, корп.6, ауд. 107а. Оборудован 5 персональными компьютерами с доступом к локальной и глобальной компьютерным сетям.

4. Отдел иностранной литературы, корп.2 ауд. 207. Оборудован 1 персональным компьютером с доступом к локальной и глобальной компьютерным сетям.

5. Библиотека юридического факультета, корп.9, ауд. 4. Оборудована 11 персональными компьютерами с доступом к локальной и глобальной компьютерным сетям.

6. Читальный зал географического факультета, корп.8, ауд. 419. Оборудован 6 персональными компьютерами с доступом к локальной и глобальной компьютерным сетям.

Все компьютеры, установленные в помещениях научной библиотеки, оснащены следующим программным обеспечением:

Операционная система ALT Linux;

Офисный пакет Libreoffice.

Справочно-правовая система «КонсультантПлюс»

Фонды оценочных средств для аттестации по дисциплине
Корпусные технологии в лингвистических исследованиях

Планируемые результаты обучения по дисциплине для формирования компетенции. Индикаторы и критерии их оценивания

ОПК.7

Способен работать с основными информационно-поисковыми и экспертными системами, системами представления знаний и обработки вербальной информации

Индикатор	Планируемые результаты обучения	Критерии оценивания результатов обучения
ОПК.7.1 Использует информационно-поисковые системы открытого доступа для проведения экспертного анализа дискурса, средства автоматизированной обработки вербальной информации	Знать возможности корпусных технологий, аннотирования лингвистического корпуса. Уметь планировать корпус, отбирать источники и определять жанрово-тематическую структуру корпуса. Владеть навыками сбора и оцифровки корпусных данных.	**Неудовлетворител** Не сформированы знания, умения и навыки, предусмотренные компетенцией. **Удовлетворительн** Знает возможности корпусных технологий, аннотирования лингвистического корпуса; умеет планировать корпус, но затрудняется отбирать источники и определять жанрово-тематическую структуру корпуса; не владеет навыками сбора и оцифровки корпусных данных. **Хорошо** Знает возможности корпусных технологий, аннотирования лингвистического корпуса; умеет планировать корпус, отбирать источники и определять жанрово-тематическую структуру корпуса; владеет навыками сбора и оцифровки корпусных данных, допуская отдельные ошибки методики. **Отлично** Знает возможности корпусных технологий, аннотирования лингвистического корпуса; умеет планировать корпус, отбирать источники и определять жанрово-тематическую структуру корпуса; владеет навыками сбора и оцифровки корпусных данных.

ОПК.2

Способен учитывать в практической деятельности специфику иноязычной научной картины мира и научного дискурса в русском и изучаемом иностранном языках

Индикатор	Планируемые результаты обучения	Критерии оценивания результатов обучения
ОПК.2.2 Осуществляет анализ	Знать возможности корпусных технологий для исследования	**Неудовлетворител** Не сформированы знания, умения и навыки,

Индикатор	Планируемые результаты обучения	Критерии оценивания результатов обучения
языковых изменений в условиях расширения глобализационных процессов	языковых изменений, частотности лексем, лексической сочетаемости, лексической синонимии. Уметь применять корпусные технологии для исследования частотности лексем, лексической сочетаемости, лексической синонимии. Владеть технологиями корпусной разметки текстов для анализа.	**Неудовлетворител** предусмотренные компетенцией. **Удовлетворительн** Знает возможности корпусных технологий для исследования частотности лексем, лексической сочетаемости, лексической синонимии; умеет применять корпусные технологии для исследования частотности лексем, лексической сочетаемости, лексической синонимии при консультативной поддержке; не владеет технологиями корпусной разметки текстов для анализа. **Хорошо** Знает возможности корпусных технологий для исследования частотности лексем, лексической сочетаемости, лексической синонимии; умеет применять корпусные технологии для исследования частотности лексем, лексической сочетаемости, лексической синонимии; не владеет технологиями корпусной разметки текстов для анализа. **Отлично** Знает возможности корпусных технологий для исследования частотности лексем, лексической сочетаемости, лексической синонимии; умеет применять корпусные технологии для исследования частотности лексем, лексической сочетаемости, лексической синонимии; владеет технологиями корпусной разметки текстов для анализа.

Оценочные средства текущего контроля и промежуточной аттестации

Схема доставки : Базовая

Вид мероприятия промежуточной аттестации : Зачет
Способ проведения мероприятия промежуточной аттестации : Оценка по дисциплине в рамках промежуточной аттестации определяется на основе баллов, набранных обучающимся на контрольных мероприятиях, проводимых в течение учебного периода.
Максимальное количество баллов : 100

Конвертация баллов в отметки

«отлично» - от 81 до 100
«хорошо» - от 61 до 80
«удовлетворительно» - от 43 до 60
«неудовлетворительно» / «незачтено» менее 43 балла

Компетенция (индикатор)	Мероприятие текущего контроля	Контролируемые элементы результатов обучения
ОПК.2.2 Осуществляет анализ языковых изменений в условиях расширения глобализационных процессов **ОПК.7.1** Использует информационно-поисковые системы открытого доступа для проведения экспертного анализа дискурса, средства автоматизированной обработки вербальной информации	Тема 1. Корпус как современный лингвистический ресурс **Письменное контрольное мероприятие**	Знание специфики прикладной лингвистики, корпусной лингвистики как ее раздела, места и роли корпусной лингвистики в отечественном и зарубежном языкознании; объекта, предмета и методов корпусной лингвистики. Умение определить круг задач корпусной лингвистики. Владение информацией о базовых категориях корпусной лингвистики.
ОПК.2.2 Осуществляет анализ языковых изменений в условиях расширения глобализационных процессов **ОПК.7.1** Использует информационно-поисковые системы открытого доступа для проведения экспертного анализа дискурса, средства автоматизированной обработки вербальной информации	Тема 7. Создание собственного корпуса **Письменное контрольное мероприятие**	Знание жанрово-тематической структуры корпусов. Умение собирать и производить оцифровку данных, выявлять базовые параметры корпуса при помощи программ анализа корпусов. Владение навыками описания и обобщения лингвистических особенностей корпуса, которые вскрылись в процессе выполнения проекта.

Компетенция (индикатор)	Мероприятие текущего контроля	Контролируемые элементы результатов обучения
ОПК.2.2 Осуществляет анализ языковых изменений в условиях расширения глобализационных процессов **ОПК.7.1** Использует информационно-поисковые системы открытого доступа для проведения экспертного анализа дискурса, средства автоматизированной обработки вербальной информации	Тема 8. Аннотирование корпуса **Итоговое контрольное мероприятие**	Знание понятийного аппарата корпусной лингвистики, корпусных технологий в лексикографии, грамматике; современных корпусов германских, романских и славянских языков. Умение применять корпусные технологии для исследования частотности лексем, лексической сочетаемости, лексической синонимии. Владение технологиями корпусной разметки текстов для анализа.

Спецификация мероприятий текущего контроля

Тема 1. Корпус как современный лингвистический ресурс

Продолжительность проведения мероприятия промежуточной аттестации: **1 часа**
Условия проведения мероприятия: **в часы самостоятельной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **30**
Проходной балл: **13**

Показатели оценивания	Баллы
Верное выполнение теста (по 0,5 балла за каждый правильный ответ)	13
Аргументация в пользу или в опровержение тезисов «Корпусная лингвистика – наука создания корпусов» и «Корпусная лингвистика – наука, базирующаяся на данных из корпусов»	10
Наличие выполненного задания 2: комментарий тезисов «Корпусная лингвистика – наука создания корпусов» и «Корпусная лингвистика – наука, базирующаяся на данных из корпусов»	7

Тема 7. Создание собственного корпуса

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**
Условия проведения мероприятия: **в часы самостоятельной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **30**
Проходной балл: **13**

Показатели оценивания	Баллы
Создание корпуса, согласно всем требованиям	30
Создание корпуса с некоторыми неточностями при оформлении и защите проекта.	25
Выполнение основных требований по созданию корпуса, 1-2 ошибки при защите проекта	13

Тема 8. Аннотирование корпуса

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**

Условия проведения мероприятия: **в часы самостоятельной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **40**
Проходной балл: **17**

Показатели оценивания	**Баллы**
Представлены результаты собственного корпусного исследования: создание текстового корпуса в рамках поставленной конкретной исследовательской задачи	17
Аргументированная защита собственного корпусного исследования с применением понятийного аппарата корпусной лингвистики	12
Верные ответы на вопросы теста (по 1 баллу за правильный вариант ответа)	11