

МИНОБРНАУКИ РОССИИ

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

Кафедра теоретического и прикладного языкознания

Авторы-составители: Худякова Екатерина Сергеевна

**Рабочая программа дисциплины
КОРПУСНАЯ ЛИНГВИСТИКА I
Код УМК 100397**

**Утверждено
Протокол №9
от «30» мая 2023 г.**

Пермь, 2023

1. Наименование дисциплины

Корпусная лингвистика I

2. Место дисциплины в структуре образовательной программы

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

Направление подготовки: **45.04.01** Филология

направленность Когнитивно-информационные технологии в гуманитарной сфере

3. Планируемые результаты обучения по дисциплине

В результате освоения дисциплины **Корпусная лингвистика I** у обучающегося должны быть сформированы следующие компетенции:

45.04.01 Филология (направленность : Когнитивно-информационные технологии в гуманитарной сфере)

ПК.5 Способен использовать системы искусственного интеллекта в гуманитарной сфере для сбора, обработки, анализа информации

Индикаторы

ПК.5.1 Использует системы искусственного интеллекта в гуманитарной сфере для сбора информации

ПК.5.2 Обрабатывает и анализирует информацию с использованием систем искусственного интеллекта

4. Объем и содержание дисциплины

Направление подготовки	45.04.01 Филология (направленность: Когнитивно-информационные технологии в гуманитарной сфере)
форма обучения	очная
№№ семестров, выделенных для изучения дисциплины	3
Объем дисциплины (з.е.)	3
Объем дисциплины (ак.час.)	108
Контактная работа с преподавателем (ак.час.), в том числе:	36
Проведение лекционных занятий	12
Проведение практических занятий, семинаров	24
Самостоятельная работа (ак.час.)	72
Формы текущего контроля	Защищаемое контрольное мероприятие (2) Итоговое контрольное мероприятие (1)
Формы промежуточной аттестации	Зачет (3 семестр)

5. Аннотированное описание содержания разделов и тем дисциплины

Корпусная лингвистика I

Курс является пропедевтикой в корпусную лингвистику. Основное внимание уделено базовым понятиям корпусной лингвистики и принципам компьютерной реализации корпуса. В результате освоения дисциплины студент должен быть способен реализовать проект создания лингвистического корпуса.

Корпусная лингвистика: проблематика, цели, применение.

Эмпиризм современной лингвистики. Пересечение с компьютерной лингвистикой: применение корпусов в приложениях NLP. Корпусы в лингвистических исследованиях. Краткая история корпусной лингвистики.

Понятие лингвистического корпуса.

Определение корпуса. Корпусный поворот в лингвистике. Корпусы и близкие понятия. Квантитативные единицы корпуса: токен и токенизация, n-граммы, коллокации. Статистические методы при использовании корпусов. Дистрибуция текста. Практикум по очистке текста и токенизации с применением Python. Библиотеки NLP для подготовки массивов текстов.

Типы и виды корпусов.

Типология корпусов: по количеству языков; по идиомам языка; по типу данных; по жанрам; по назначению, по синхронности / диахронности, по объему; по организации. Специализированные корпуса: прикладные без пользовательского интерфейса, мультимедийные, для специальных целей (корпуса диалектной речи, корпуса детской речи, корпуса билингвальной речи и др.).

Типы корпусной разметки.

Понятие разметки корпуса. Типы разметки: метаразметка и ее теги; структурная разметка (токенизация по словам, по предложениям, по структурным элементам текста); собственно лингвистическая разметка (морфологическая, синтаксическая, семантическая и др.). Стандарты разметки EAGLES (European Advisory Group on Language Engineering Standards), TEI (Text Encoding for Interchange), XCES (XML Corpus Encoding Standard). Представление аннотации в xml: атрибут -- значение.

Понятие корпусного менеджера.

Общее представление о корпусном менеджере. Элементы интерфейса. Инструмент разметки и инструмент поиска. Знакомство с UAM corpustool, Elan, Exmaralda. Практика по работе с корпусными менеджерами.

Аннотирование корпуса. Проект по созданию корпуса.

Кейс: создание авторского корпуса. Практическая работа предполагает: определение цели и типа корпуса; подбор текстов и их препроцессинг; обоснование типов разметки; осуществление аннотирования корпуса в корпусном менеджере.

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. Галич, Г. Г. Современные направления лингвистики : учебное пособие для студентов магистратуры, обучающихся по направлению 45.04.02 «Лингвистика», профиль подготовки «Практика и лингводидактика профессионально ориентированного перевода» / Г. Г. Галич. — Омск : Издательство Омского государственного университета, 2020. — 184 с. — ISBN 978-5-7779-2464-3. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. <https://www.iprbookshop.ru/108138>
2. Методы когнитивного анализа семантики слова: компьютерно-корпусный подход / В. И. Заботкина, Е. Е. Голубкова, М. А. Кронгауз [и др.] ; под редакцией В. И. Заботкиной. — 2-е изд. — Москва : Издательский Дом ЯСК, 2019. — 342 с. — ISBN 978-5-907117-73-0. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. <http://www.iprbookshop.ru/97629.html>
3. 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>

Дополнительная:

1. Рассказы о сновидениях. Корпусное исследование устного русского дискурса / О. А. Савельева-Трофимова, В. Л. Цуканова, А. О. Литвиненко [и др.] ; под редакцией А. А. Кибрик, В. И. Подлесская. — Москва : Языки славянских культур, 2009. — 736 с. — ISBN 978-5-9551-0303-7. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. <http://www.iprbookshop.ru/15133>
2. McCarthy M., O'Dell F. English collocations in use. How words work together for fluent and natural English: self-study and classroom use / M. McCarthy, F. O'Dell. — Cambridge: Cambridge University Press, 2006, ISBN 0-521-60378-1. — 190.
3. Fischer R. Lexical change in present-day English: a corpus-based study of the motivation, institutionalization, and productivity of creative neologisms / R. Fischer. — Tübingen: GNV, 1998, ISBN 3-8233-4940-6. — 209.
4. Лесников С. В. Метаязык лингвистики. в 2 т. Т. 1. Проблемы систематизации терминосистемы (гизаурис, классификация, синопсис) / С. В. Лесников. — Санкт-Петербург: Нестор-История, 2021, ISBN 978-5-4469-2007-5. — 512. — Библиогр.: с. 377-478
5. Корпусный анализ русского стиха: [сборник статей] / Рос. акад. наук, Ин-т рус. яз. им. В. В. Виноградова. — Москва: Азбуковник, 2013, ISBN 978-5-91172-073-5. — 2651. — Библиогр. в конце ст.

9. Перечень ресурсов сети Интернет, необходимых для освоения дисциплины

<https://ruscorpora.ru/> Национальный корпус русского языка

<https://www.korpus.cz/> Чешский национальный корпус

<https://www.english-corpora.org/bnc/> Британский национальный корпус

<https://transcriber.semograph.com/> Инструмент создания корпуса

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

Образовательный процесс по дисциплине **Корпусная лингвистика I** предполагает использование следующего программного обеспечения и информационных справочных систем:

Презентационные материалы (слайды по темам лекционных и практических занятий); доступ в режиме on-line в Электронную библиотечную систему (ЭБС); доступ в электронную информационно-образовательную среду университета.

Образовательный процесс по дисциплине предполагает использование следующего программного обеспечения и информационных систем:

1. Офисный пакет
 2. Приложения для создания корпусов (UAM corpustool, Elan, Exmaralda)
 3. Онлайн приложение для создания корпуса (<https://transcriber.semograph.com/>)
- Программы пп. 1–2 являются свободно распространяемыми (free software).

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

Для проведения лекционных занятий и текущего контроля необходима аудитория, оснащенная презентационной техникой (проектор, экран, компьютер/ноутбук) с соответствующим программным обеспечением, меловой (и) или маркерной доской.

Для проведения практических занятий необходим компьютерный класс, оснащенный персональными ЭВМ и соответствующим программным обеспечением. Состав оборудования определен в Паспорте компьютерного класса.

Аудитория для самостоятельной работы, оснащенный компьютерной техникой с возможностью подключения к сети «Интернет», обеспеченный доступом в электронную информационно-образовательную среду университета.

Помещения Научной библиотеки ПГНИУ.

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
Корпусная лингвистика I**

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

ПК.5

Способен использовать системы искусственного интеллекта в гуманитарной сфере для сбора, обработки, анализа информации

Индикатор	Планируемые результаты обучения	Критерии оценивания результатов обучения
ПК.5.1 Использует системы искусственного интеллекта в гуманитарной сфере для сбора информации	Использует лингвистический корпус для сбора информации для научного исследования: знает типологии корпусов и корпусной разметки, умеет осуществлять в нем поиск, владеет принципами интерпретации выдачи	Неудовлетворител Не знает принципов создания корпусов, типы и виды корпусов и корпусной разметки, не умеет осуществлять поиск в корпусе, не владеет интерфейсом корпуса для поиска нужной информации Удовлетворительн Частично знает принципы создания корпусов, типы и виды корпусов и корпусной разметки, частично умеет осуществлять поиск в корпусе, частично владеет интерфейсом корпуса для поиска нужной информации Хорошо Хорошо знает принципы создания корпусов, типы и виды корпусов и корпусной разметки, хорошо умеет осуществлять поиск в корпусе, хорошо владеет интерфейсом корпуса для поиска нужной информации Отлично Отлично знает принципы создания корпусов, типы и виды корпусов и корпусной разметки, умеет осуществлять поиск в корпусе, полностью владеет интерфейсом корпуса для поиска нужной информации
ПК.5.2 Обрабатывает и анализирует информацию с использованием систем искусственного интеллекта	Обрабатывает лингвистическую информацию с использованием корпуса: знает принципы создания корпуса, умеет применять корпусный менеджер, владеет технологиями создания корпуса	Неудовлетворител Не знает принципов создания корпуса, не умеет применять корпусный менеджер, не владеет технологиями создания корпуса Удовлетворительн Частично знает принципы создания корпуса, частично умеет применять корпусный менеджер, частично владеет технологиями создания корпуса Хорошо Хорошо знает принципы создания корпуса,

Индикатор	Планируемые результаты обучения	Критерии оценивания результатов обучения
		Хорошо хорошо умеет применять корпусный менеджер, хорошо владеет технологиями создания корпуса Отлично Отлично знает принципы создания корпуса, полностью умеет применять корпусный менеджер, отлично владеет технологиями создания корпуса

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
ПК.5.2 Обрабатывает и анализирует информацию с использованием систем искусственного интеллекта ПК.5.1 Использует системы искусственного интеллекта в гуманитарной сфере для сбора информации	Типы и виды корпусов. Защищаемое контрольное мероприятие	Знание типологии корпусов, умение определять тип корпуса для исследовательских задач, владение интерфейсом
ПК.5.2 Обрабатывает и анализирует информацию с использованием систем искусственного интеллекта ПК.5.1 Использует системы искусственного интеллекта в гуманитарной сфере для сбора информации	Типы корпусной разметки. Защищаемое контрольное мероприятие	Знание типологии корпусной аннотации; умение определять тип разметки и тип тега, владение интерфейсом корпуса

Компетенция (индикатор)	Мероприятие текущего контроля	Контролируемые элементы результатов обучения
ПК.5.2 Обрабатывает и анализирует информацию с использованием систем искусственного интеллекта ПК.5.1 Использует системы искусственного интеллекта в гуманитарной сфере для сбора информации	Аннотирование корпуса. Проект по созданию корпуса. Итоговое контрольное мероприятие	Знание технологий создания корпуса, владение инструментарием создания корпуса, умение создать учебный корпус

Спецификация мероприятий текущего контроля

Типы и виды корпусов.

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

Условия проведения мероприятия: **в часы аудиторной работы**

Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **30**

Проходной балл: **13**

Показатели оценивания	Баллы
Знает типологии корпусов, умеет определять тип корпуса для исследовательских задач, владеет интерфейсом. Представлено 10 корпусов, за полную и верную квалификацию каждого -макс. 3 балла. Ошибок нет.	30
Частично знает типологии корпусов, хорошо умеет определять тип корпуса для исследовательских задач, хорошо владеет интерфейсом. Представлено 10 корпусов, за полную и верную квалификацию каждого -макс. 3 балла. За ошибки баллы не начисляются	24
Частично знает типологии корпусов, частично умеет определять тип корпуса для исследовательских задач, частично владеет интерфейсом. Представлено 10 корпусов, за полную и верную квалификацию каждого -макс. 3 балла. За ошибки баллы не начисляются	18
Не знает типологии корпусов, не умеет определять тип корпуса для исследовательских задач, не владеет интерфейсом. Представлено 10 корпусов, за полную и верную квалификацию каждого -макс. 3 балла. За ошибки баллы не начисляются. Набрано менее 13 баллов.	12

Типы корпусной разметки.

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

Условия проведения мероприятия: **в часы аудиторной работы**

Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **30**

Проходной балл: **13**

Показатели оценивания	Баллы
Полностью знает типологии разметки, умеет определять тип аннотации для исследовательских задач, отлично владеет интерфейсом. Представлено 10 корпусов, за полную и верную квалификацию каждого -макс. 3 балла. Ошибок в определении типа разметки нет.	30

Частично знает типологии разметки, хорошо умеет определять тип аннотации для исследовательских задач, хорошо владеет интерфейсом. Представлено 10 корпусов, за полную и верную квалификацию каждого -макс. 3 балла. За ошибки баллы не начисляются	24
Частично знает типологии разметки, частично умеет определять тип аннотации для исследовательских задач, частично владеет интерфейсом. Представлено 10 корпусов, за полную и верную квалификацию каждого -макс. 3 балла. За ошибки баллы не начисляются.	18
Не знает типологии разметки, не умеет определять тип аннотации для исследовательских задач, не владеет интерфейсом. Представлено 10 корпусов, за полную и верную квалификацию каждого -макс. 3 балла. За ошибки баллы не начисляются. Набрано суммарно менее 13 баллов.	12

Аннотирование корпуса. Проект по созданию корпуса.

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

Условия проведения мероприятия: **в часы аудиторной работы**

Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **40**

Проходной балл: **17**

Показатели оценивания	Баллы
Отлично знает технологии создания корпуса, полностью владеет инструментарием создания корпуса, умеет создать учебный корпус. Корпус представителен (10 баллов), корпус предобработан и размечен (10 баллов), аннотации соответствуют цели проекта (5 баллов); поиск по корпусу осуществляется (5 баллов); единицы аннотированы верно (5 баллов), нет технических сбоев (5 баллов).	40
Хорошо знает технологии создания корпуса, хорошо владеет инструментарием создания корпуса, в целом умеет создать учебный корпус. Критерии оценки: корпус представителен (10 баллов), корпус предобработан и размечен (10 баллов), аннотации соответствуют цели проекта (5 баллов); поиск по корпусу осуществляется (5 баллов); единицы аннотированы верно (5 баллов), нет технических сбоев (5 баллов). Суммарно набрано 32 балла.	32
Частично знает технологии создания корпуса, частично владеет инструментарием создания корпуса, в целом умеет создать учебный корпус. Критерии оценки: корпус представителен (10 баллов), корпус предобработан и размечен (10 баллов), аннотации соответствуют цели проекта (5 баллов); поиск по корпусу осуществляется (5 баллов); единицы аннотированы верно (5 баллов), нет технических сбоев (5 баллов). Суммарно набрано 24 балла.	24
Не знает технологий создания корпуса, не владеет инструментарием создания корпуса, не умеет создавать учебный корпус. Критерии оценки: корпус представителен (10 баллов), корпус предобработан и размечен (10 баллов), аннотации соответствуют цели проекта (5 баллов); поиск по корпусу осуществляется (5 баллов); единицы аннотированы верно (5 баллов), нет технических сбоев (5 баллов). Суммарно набрано менее 16 балла.	15