

Минобрнауки России

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Оренбургский государственный университет имени В.А. Бондаренко»

Кафедра прикладной лингвистики и перевода

РАБОЧАЯ ПРОГРАММА

ДИСЦИПЛИНЫ

«Б1.Д.Б.28 Корпусная лингвистика»

Уровень высшего образования

БАКАЛАВРИАТ

Направление подготовки

45.03.04 Интеллектуальные системы в гуманитарной сфере
(код и наименование направления подготовки)

Языковые модели и искусственный интеллект
(наименование направленности (профиля) образовательной программы)

Квалификация

Бакалавр

Форма обучения

Очная

Год набора 2026

Рабочая программа дисциплины «Б1.Д.Б.28 Корпусная лингвистика» рассмотрена и утверждена на заседании кафедры

Кафедра прикладной лингвистики и перевода

наименование кафедры

протокол № 4 от "20" 03 2016 г.

Заведующий кафедрой

Кафедра прикладной лингвистики и перевода

наименование кафедры

подпись

Е.Д. Андреева

расшифровка подписи

Исполнители:

доцент

должность

подпись

Е.Д. Андреева

расшифровка подписи

должность

подпись

расшифровка подписи

СОГЛАСОВАНО:

Председатель методической комиссии по направлению подготовки

45.03.04 Интеллектуальные системы в гуманитарной сфере

код наименование

личная подпись

расшифровка подписи

Е.Д. Андреева

Заведующий отделом формирования фонда и научной обработки документов

личная подпись

С.А. Биктимирова

расшифровка подписи

Уполномоченный по качеству факультета

личная подпись

Т.В. Сапух

расшифровка подписи

№ регистрации _____

© Андреева Е.Д., 2026

© ОГУ, 2026

1 Цели и задачи освоения дисциплины

Цель (цели) освоения дисциплины:

формирование базовой системы компетенций, знаний, умений и навыков в сфере создания и применения современных корпусных технологий и корпусов в лингвистической теории и практике и ознакомление с существующими программными продуктами в этой сфере.

Задачи:

- освоить основы корпусной лингвистики и корпусных технологий в рамках эмпирического подхода к изучению языка;
- ознакомиться с историей, основными проблемами, понятиями и концепциями современной корпусной лингвистики;
- ознакомиться с основными корпусами и их возможностями, принципами из построения и разметки; научиться оценивать и сопоставлять различные корпуса;
- приобрести практические навыки работы с корпусами и связанными с ними компьютерными инструментами и программными продуктами (конкордансеры, корпусные менеджеры, программы для автоматического создания корпусов и др.);
- сформировать понимание межпредметных связей корпусной лингвистики с дисциплинами лингвистического, гуманитарного и математического циклов и видение возможностей использования корпусной лингвистики в других областях знания;
- сформировать понимание теоретического и практического значения корпусов для проведения научных лингвистических исследований и развития информационных технологий;
- сформировать умение использовать лингвистические корпуса для решения конкретных исследовательских задач;
- приобрести опыт работы с корпусными средствами для построения собственного исследовательского корпуса.

2 Место дисциплины в структуре образовательной программы

Дисциплина относится к базовой части блока Д «Дисциплины (модули)»

Пререквизиты дисциплины: *Б1.Д.Б.24 Компьютерная лингвистика, Б1.Д.В.11 Морфология и словообразование, Б1.Д.В.12 Семантика*

Постреквизиты дисциплины: *Б1.Д.Б.30 Компьютерная лексикография*

3 Требования к результатам обучения по дисциплине

Процесс изучения дисциплины направлен на формирование следующих результатов обучения

Код и наименование формируемых компетенций	Код и наименование индикатора достижения компетенции	Планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенций
ОПК-1 Способен применять в профессиональной деятельности методы математического анализа, логики и моделирования, теоретического и экспериментального исследования в информатике, лингвистике и гуманитарных науках	ОПК-1-В-1 Способен использовать основы математического анализа, логики и математического моделирования ОПК-1-В-2 Способен использовать математические методы для построения моделей в информатике, лингвистике и других гуманитарных дисциплинах ОПК-1-В-3 Владеет методами теоретического и экспериментального исследования	Знать: принципы функционирования автоматизированных лингвистических систем Уметь: применять методы корпусной лингвистики для анализа текстовой информации на естественном языке Владеть: методами корпусной лингвистики для анализа текстовой информации на естественном языке

Код и наименование формируемых компетенций	Код и наименование индикатора достижения компетенции	Планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенций
	в информатике, лингвистике и других гуманитарных дисциплинах	

4 Структура и содержание дисциплины

4.1 Структура дисциплины

Общая трудоемкость дисциплины составляет 4 зачетные единицы (144 академических часа).

Вид работы	Трудоемкость, академических часов	
	5 семестр	всего
Общая трудоёмкость	144	144
Контактная работа:	52,25	52,25
Лекции (Л)	18	18
Практические занятия (ПЗ)	34	34
Промежуточная аттестация (зачет, экзамен)	0,25	0,25
Самостоятельная работа: - выполнение индивидуального творческого задания (ИТЗ); - самоподготовка (проработка и повторение лекционного материала и материала учебников и учебных пособий; - изучение разделов курса в системе электронного обучения; - подготовка к практическим занятиям)	91,75	91,75
Вид итогового контроля (зачет, экзамен, дифференцированный зачет)	диф. зач.	

Разделы дисциплины, изучаемые в 5 семестре

№ раздела	Наименование разделов	Количество часов				
		всего	аудиторная работа			внеауд. работа
			Л	ПЗ	ЛР	
1	Теоретические основы корпусной лингвистики	20	4	4		12
2	Классификация корпусов	26	2	6		18
3	Создание и использование лингвистических корпусов	48	6	10		32
4	Сферы применения корпусной лингвистики	50	6	14		30
	Итого:	144	18	34		92
	Всего:	144	18	34		92

4.2 Содержание разделов дисциплины

1 Теоретические основы корпусной лингвистики

Возникновение количественных методов в языкознании. Основные понятия корпусной лингвистики. Понятие лингвистического корпуса. Лингвистические (языковые) и нелингвистические корпуса. Основные положения корпусной лингвистики. Корпусный подход и его специфика. Цели корпусной лингвистики. Принятие принципиальных теоретических решений составителями корпусов, зависимость валидности данных от предоставляемой информации. Основные свойства лингвистических корпусов, репрезентативность и сбалансированность. Корпусная лингвистика: современное состояние. Корпусная лингвистика в России.

2 Классификация корпусов

Корпусные и лексикографические исследования. Машинные фонды, базы данных, корпус. Критерии классификации корпусов. Типы корпусов. Классификация корпусов по хронологии, по жанрам, по аннотированности, по целям и задачам, по объему, по языкам, по формальным признакам. Корпусы языков мира. История появления корпусов языков мира: от картотеки к корпусу. Первые компьютерные корпусы. Современные корпусы языков мира. Национальный корпус русского языка. Другие корпуса русского языка (Упсальский корпус, Тюбингенские корпусы, компьютерный корпус текстов русских газет конца XX века, система Диктум-1, Большой корпус русского языка, Машинный фонд русского языка, корпус русского литературного языка, ХАНКО, OpenCorpora). Корпусы английского языка (Brown Corpus, British National Corpus, Lancaster Corpus, Uppsala Corpus, COCA, COHA, Learners' Corpus, Scottish Corpus, Penn Treebank, Оксфордский корпус, TenTen, Google Books Ngram, Groningen Meaning Bank). Специальные корпусы.

3 Создание и использование лингвистических корпусов

Технологии корпусной лингвистики. Принципы и способы составления корпусов. Конкурирующие технологии: rule-based, статистические, основанные на ограничениях (restriction-based), гибридные. Предварительные работы по созданию корпуса. Отбор источников. Внешние и внутренние критерии отбора. Создание формализованных описаний для разработки корпуса. Словарный и бессловарный способ. Предварительная обработка. Нормализация файлов. Графематический анализ. Токенизация, лемматизация, стемминг, парсинг. Разметка в лингвистическом корпусе. Виды лингвистической корпусной разметки, метатекстовая и лингвистическая разметка. Автоматический морфологический и синтаксический анализ. Средства создания и разметки корпусов. Качественные и количественные данные в корпусе. Стандартизация в корпусной лингвистике. Языковые средства представления размеченных текстов. Международные стандарты и проекты (TEI, EAGLES, CDIF, XCES). Параллельные корпусы. Проблема выравнивания. Создание своего корпуса. Планирование. Сбор и оцифровка данных. Кодировка текста. Техника и алгоритм создания корпуса письменных текстов. Техника и алгоритм создания корпуса устных текстов. Техники бутстрэппинга. Проблема авторских прав. Программы для оцифровки текстов. Расшифровка устной речи. Программы для расшифровки. Аннотирование корпуса. Хранение, публикация и обновление корпусов. Единица хранения корпуса данных. Формат представления информации. Онлайн корпуса. Оффлайн корпуса. Программы аннотирования корпусов. Проблема обновления корпусов. Закон Ципфа. Визуализация синтаксических деревьев с phpSyntaxTree. Регулярные выражения. NLTK – библиотека для Python.

4 Сферы применения корпусной лингвистики

Пользователи корпусов. Корпус и электронная библиотека. Корпус и поисковая система. Корпусные менеджеры. Корпусные менеджеры. Xaira. CQP. DDC. Manatee/Bonito. GATE. Sketch Engine. Game with a purpose. BootCaT. Корпус как информационно-поисковая система. Методы извлечения информации из корпуса. Языки запросов. Выходные интерфейсы. Web-интерфейсы к корпусам (COCA, НКРЯ). Конкорданс, коллокация, кластеры, чанки, коллигация. Пользовательский подкорпус. Конкордансер AntConc. Корпусные исследования. Лексикографические исследования, основанные на корпусах. Грамматические исследования, основанные на корпусах. Семантические исследования, основанные на корпусах. Исследования дискурса, основанные на корпусах. Корпусные методы в применении к различным дисциплинам. Международные организации корпусной лингвистики.

4.3 Практические занятия (семинары)

№ занятия	№ раздела	Тема	Кол-во часов
1	1	Основные понятия корпусной лингвистики. Основные положения корпусной лингвистики. Корпусный подход и его специфика.	2
2	1	Основные свойства лингвистических корпусов. Корпусная лингвистика в России.	2
3	2	История появления корпусов. Типы корпусов.	2
4	2	Корпусы русского языка.	2

№ занятия	№ раз-дела	Тема	Кол-во ча-сов
5	2	Корпусы английского языка.	2
6	3	Принципы и технологии составления корпусов Предварительные работы по созданию корпуса	2
7	3	Разметка в лингвистическом корпусе.	2
8	3	Стандартизация в корпусной лингвистике.	2
9	3	Параллельные корпуса	2
10	3	Создание своего корпуса	2
11-12	4	Корпусные менеджеры	4
13-14	4	Корпус как информационно-поисковая система	4
15-17	4	Корпусные исследования	6
		Итого:	34

5 Учебно-методическое обеспечение дисциплины

5.1 Основная литература

1. Грудева, Е.В. Корпусная лингвистика : учебное пособие / Е.В. Грудева ; науч. ред. Л.Н. Чурилина. – 3-е изд., стер. – Москва : ФЛИНТА, 2017. – 166 с. – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=364207>. – ISBN 978-5-9765-1497-3. – Текст : электронный.
2. Потапова, Р.К. Новые информационные технологии и лингвистика : учебное пособие / Р.К. Потапова. – Москва : ЛЕНАНД, 2016. – 364 с. – ISBN 978-5-9710-2394-4.

5.2 Дополнительная литература

1. Антопольский, А.Б. Лингвистические информационные ресурсы: монография/ А.Б. Антопольский ; науч. ред. Д. В. Ефременко. – Москва : Институт научной информации по общественным наукам (ИНИОН) РАН, 2022. – 466 с. – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=695077>. – Текст : электронный.
2. Ляшевская, О.Н. Корпусные инструменты в грамматических исследованиях русского языка / О.Н. Ляшевская. – Москва : Языки славянской культуры, 2016. – 520 с. – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=473302>. – Текст : электронный.
3. Мельчук, И.А. Опыт теории лингвистических моделей «Смысл <=> Текст» / И.А. Мельчук. – Москва : Школа «Языки русской культуры», 1999. – 370 с. – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=211373>. – Текст : электронный.
4. Прикладная и компьютерная лингвистика: коллективная монография / под ред.: И.С. Николаева, О.В. Митрениной, Т.М. Ландо. – Москва : ЛЕНАНД, 2017. – 315 с.

5.3 Периодические издания

Не предусмотрены.

5.4 Интернет-ресурсы

<http://aot.ru> – ПО в области автоматической обработки текста
<https://www.laurenceanthony.net/software.html> – конкордансер AntConc
<https://software.sil.org/lexiquepro/> – ПО для создания и публикации словарей
<http://billposer.org/Software/msort.html> – ПО для работы с файлами
<https://sourceforge.net/projects/tamsys/> – Text Analysis Markup System
<https://www.english-corpora.org/coca/> – Corpus of Contemporary American English (COCA)
<https://www.english-corpora.org/bnc/> – British National Corpus (BNC)
<http://www.natcorp.ox.ac.uk/> – British National Corpus (BNC)

<https://uniserv.iis.nsk.su/thes/> – Русско-английский тезаурус по компьютерной лингвистике
<https://uniserv.iis.nsk.su/cl/> – портал по компьютерной лингвистике
<https://nlpub.ru/> – каталог ресурсов для обработки естественного языка
<http://rusgram.ru/new/> – Русская корпусная грамматика
<https://lex.philol.msu.ru/proekty/> – Лаборатория общей и компьютерной лексикологии и лексикографии филологического факультета МГУ имени М.В. Ломоносова
<https://sysblok.ru/category/linguistics/> – статьи о корпусных исследованиях
<http://cfrl.ruslang.ru/> – Машинный фонд русского языка
<https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/0167> – Lancaster-Oslo-Bergen corpus of modern English (LOB)
<https://narusco.ru/> – Корпус русского литературного языка
http://www.philol.msu.ru/~lex/corpus/corp_descr.html – Компьютерный корпус текстов русских газет конца XX-ого века
<http://corpus.leeds.ac.uk/> – Collection of English corpora
<http://rykov-cl.narod.ru/> – сайт курса лекций и цикла статей по корпусной лингвистике Рыкова В.В.
<http://dict.ruslang.ru/freq.php> – О.Н. Ляшевская, С.А. Шаров «Частотный словарь современного русского языка (на материалах Национального корпуса русского языка)»
<https://www.english-corpora.org/> – online corpora
<https://www.dialog-21.ru/> – сайт международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог»
<https://www ldc.upenn.edu/> – Penn Treebank
<http://bokrcorpora.narod.ru/index.html> – БОКР (Большой Корпус русского языка)
<https://ruscorpora.ru/> – Национальный корпус русского языка (НКРЯ)
<http://www.bmanuel.org/index.html> – Languages, computers and humanities
<http://www.athel.com/corpus.html> – Text Corpora and Corpus Linguistics
<http://www.elsnet.org/resources.html> – Language and Speech Resources
<https://www.corpusdelespanol.org/> – исторический корпус испанского языка
«Введение в корпусную лингвистику» [Электронный ресурс]: онлайн-курс на платформе openedu.ru – «Открытое образование» / Разработчик курса: Национальный исследовательский университет «Высшая школа экономики», режим доступа: <https://openedu.ru/course/hse/CORPUS/>

5.5 Программное обеспечение, профессиональные базы данных и информационные справочные системы

1. Операционная система РЕД ОС.
2. Пакет офисных приложений «МойОфис Образование»
3. Для работы с ресурсами Интернет - веб-браузер Яндекс <https://yandex.ru/>.
4. ГАРАНТ Платформа F1 [Электронный ресурс]: справочно-правовая система. / Разработчик ООО НПП «ГАРАНТ-Сервис», 1999, Москва, Воробьевы горы, МГУ, [1990–2026]. – Режим доступа в сети ОГУ <http://garant.net.osu.ru>
5. Автоматизированная интерактивная система сетевого тестирования – АИССТ (зарегистрирована в РОСПАТЕНТ, Свидетельство о государственной регистрации программы для ЭВМ №2011610456, правообладатель – Оренбургский государственный университет), режим доступа – <http://aist.osu.ru>.

6 Материально-техническое обеспечение дисциплины

Учебные аудитории для проведения занятий семинарского типа, для проведения групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации.

Аудитории оснащены комплектами ученической мебели, техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой, подключенной к сети «Интернет», и обеспечены доступом в электронную информационно-образовательную среду ОГУ.