


Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования «Новосибирский национальный исследовательский
государственный университет» (Новосибирский государственный университет, НГУ)

Гуманитарный институт

СОГЛАСОВАНО

Директор ГИ


Зуев А.С.

«29» сентября 2020 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

КОРПУСНАЯ ЛИНГВИСТИКА

Направление подготовки: 45.04.01 Филология (магистратура)

Направленность (профиль): Русская филология, Филология, Русский язык, литература,
культура

Форма обучения: очная

Разработчик:

канд. филол. наук, доцент Лаврентьев А.М.



И.о. зав. кафедрой фундаментальной и прикладной
лингвистики, д-р филос. наук, профессор Савостьянов А.Н.



Руководитель программы:

д-р филол. наук, профессор Кошкарева Н. Б.



Новосибирск

Содержание

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы.....	3
2. Место дисциплины в структуре образовательной программы	3
3. Трудоемкость дисциплины в зачетных единицах с указанием количества академических часов, выделенных на контактную работу обучающегося с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающегося	3
4. Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий.....	4
5. Перечень учебной литературы	6
6. Перечень учебно-методических материалов по самостоятельной работе обучающихся..	6
7. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины	6
8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине	6
9. Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине	7
10. Оценочные средства для проведения текущего контроля и промежуточной аттестации по дисциплине.....	7
Приложение 1 Аннотация по дисциплине	
Приложение 2 Оценочные средства по дисциплине	

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

Результаты освоения образовательной программы (компетенции)	Результаты обучения по дисциплине		
	В результате изучения дисциплины обучающиеся должны:		
	знать	уметь	владеть
ОПК-4 способность демонстрировать углубленные знания в избранной конкретной области филологии	основные этапы развития и актуальные тенденции корпусной лингвистики; - основные термины и методы корпусных исследований, основные операторы языка запросов SQL.	- оценить основные характеристики и грамотно использовать доступные в Интернете лингвистические корпуса; - спроектировать и реализовать на платформе ТХМ корпус текстов или других языковых материалов, нацеленный на решение определенной исследовательской задачи.	Иметь навыки использования: - основных операторов языка SQL, регулярных выражений, языка разметки XML, инструментов импорта, количественного и качественного анализа корпуса на платформе ТХМ
ПК-1 владение навыками самостоятельного проведения научных исследований в области системы языка и основных закономерностей функционирования фольклора и литературы в синхроническом и диахроническом аспектах, в сфере устной, письменной и виртуальной коммуникации	теорию и практику создания компьютерных корпусов в лингвистических исследованиях	применять компьютерные корпуса в лингвистических исследованиях	принципами создания электронных языковых ресурсов (текстовых, речевых и мультимодальных корпусов; словарей, тезаурусов, онтологии; фонетических, лексических, грамматических и иных баз данных и баз знаний) и навыками использования таких ресурсов

2. Место дисциплины в структуре образовательной программы

Дисциплины (практики), изучение которых необходимо для освоения дисциплины
Корпусная лингвистика: Компьютерные технологии в филологии (курс бакалавриата).

Дисциплины (практики), для изучения которых необходимо освоение дисциплины
Корпусная лингвистика: Производственная практика, НИР (ОК-3, ПК-1,4).

3. Трудоемкость дисциплины в зачетных единицах с указанием количества академических часов, выделенных на контактную работу обучающегося с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающегося

Трудоемкость дисциплины: 2 з.е. (72 ч) -- Русская филология, Русский язык, литература, культура 2019,3 з.е. (108 ч) – Филология, Русский язык, литература, культура 2020.
Форма промежуточной аттестации 1/3 семестр – зачет.

№	Вид деятельности	набор	набор
		2019	2020

		Семестр 3	Семестр 1
1	Лекции, ч	32	32
2	Практические занятия, ч	16	16
3	Лабораторные занятия, ч		
4	Занятия в контактной форме, ч, из них	50	50
5	из них аудиторных занятий, ч	48	48
6	в электронной форме, ч	-	-
7	консультаций, час.		
8	промежуточная аттестация, ч	2	2
9	Самостоятельная работа, час.	58	22
10	Всего, ч	108	72

4. Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

Лекции (32 ч)

Наименование темы и их содержание	Объем, час
1. Возникновение и основные этапы развития корпусных исследований в лингвистике и других гуманитарных науках <ul style="list-style-type: none"> • Докомпьютерная корпусная лингвистика: стилометрия, механо-лингвистика, лексикография • Index Thomisticum • Брауновский корпус, SEU, LOB • Тенденции развития в 1980-е — 2000-е годы • Частотный словарь русского языка, проект машинного фонда русского языка 	2
2. Место корпусной лингвистики в ряду гуманитарных и компьютерных дисциплин <ul style="list-style-type: none"> • корпусная и традиционная лингвистика • компьютерная лингвистика, автоматическая обработка текста • цифровая гуманитаристика 	2
3. Основные задачи и понятия корпусной лингвистики <ul style="list-style-type: none"> • понятия корпуса, текстоформы, токена, текстовой единицы, конкорданса, частотности, вхождения, разметки, метаданных, репрезентативности 	2
4. Типология корпусов <ul style="list-style-type: none"> • по цели создания, по репрезентативности, по видам разметки, по характеру текстов, по размеру текстовых единиц, по количеству языков, по диахронической глубине, по условиям распространения и др. 	2
5. Национальные корпуса <ul style="list-style-type: none"> • Британский национальный корпус • Национальный корпус русского языка, Открытый корпус русского языка • Национальные корпуса других языков 	2
6. Всемирная паутина как корпус <ul style="list-style-type: none"> • Преимущества и методологические проблемы использования Интернета как корпуса • Инструменты для сбора корпусов из Интернета (кроулеры) • Специальные возможности поисковых машин 	2

7. Специализированные корпуса и текстовые базы <ul style="list-style-type: none"> • Диахронические корпуса • Параллельные корпуса • Корпуса устной речи, мультимедийные корпуса 	4
8. Проектирование и создание корпуса <ul style="list-style-type: none"> • Технологический процесс создания корпуса • Отбор и подготовка текстов • Подготовка метаданных • Разметка • Загрузка в корпус-менеджер • Публикация 	4
9. Лингвистическая разметка: виды, стандарты и инструменты <ul style="list-style-type: none"> • Токенизация, лемматизация, морфологическая, синтаксическая и др. • Встроенная и отдельно стоящая • Язык XML 	4
10. Филологическая разметка: XML-TEI <ul style="list-style-type: none"> • Причины создания и история TEI • Модель текста и основные разделы TEI • Инструменты для работы с TEI 	4
11. Инструменты для работы с корпусами, ТХМ <ul style="list-style-type: none"> • Основные функции корпус-менеджеров: загрузка корпуса, интерфейс для запросов и анализа, выгрузка результатов • Общее описание платформы ТХМ, порядок установки • Качественный анализ корпуса с ТХМ: чтение, конкордансы, индексы, запросы CQL • Количественный анализ: разбивка корпуса, лексическая таблица, специфичность, фактор-ный анализ • Выгрузка результатов • Подготовка и импортирование корпуса в ТХМ (модули Clipboard, ТХТ, ХТЗ), метаданные 	4

Практические занятия (16 ч)

Содержание практического занятия	Объем, час
Семинар по теме:	
1. Национальные корпуса	4
2. Проектирование и создание корпуса	4
3. Филологическая разметка: XML-TEI	2
4. Инструменты для работы с корпусами, ТХМ	6

Самостоятельная работа студентов

Перечень занятий на СРС	Объем, час	
	набор 2020	набор 2019
Работа в команде над проектом создания корпуса	17	6
Подготовка докладов о ресурсах в интернете	17	6
Самостоятельное изучение теоретического материала по разделам дисциплины	17	6
Подготовка к зачету	4	4
Итого	58	22

5. Перечень учебной литературы

5.1 Основная литература

1. Грудева, Е. В. Корпусная лингвистика : учебное пособие / Е. В. Грудева. — 3-е изд. — Москва : ФЛИНТА, 2017. — 165 с. — ISBN 978-5-9765-1497-3. — Текст : электронный // Лань : электронно-библиотечная система. — URL: <https://e.lanbook.com/book/106859>. — Режим доступа: для авториз. пользователей.

2. Захаров, В. П. Корпусная лингвистика : учебник / В. П. Захаров, С. Ю. Богданова. — 3-е изд., перераб. — Санкт-Петербург : СПбГУ, 2020. — 234 с. — ISBN 978-5-288-05997-1. — Текст : электронный // Лань : электронно-библиотечная система. — URL: <https://e.lanbook.com/book/139576>. — Режим доступа: для авториз. пользователей.

3. Копотев, Михаил. Введение в корпусную лингвистику [Текст: электронный ресурс] : учебное пособие для студентов филологических и лингвистических специальностей университетов / М. Копотев. Прага : Animedia Company, 2014. 194 с. : ил., табл. URL: <http://biblioclub.ru/index.php?page=book&id=375463>.

5.2 Дополнительная литература

4. Biber D. A Typology of English texts // Linguistics 27 (1989). P. 3–43 URL: <https://www.degruyter.com/view/j/ling.2013.51.issue-jubilee/ling-2013-0040ad.pdf>

5. Kilgarriff A. & Grefenstette G. Introduction to the Special Issue on Web as Corpus // Computational Linguistics 29 (3), 2003. P. 333-347. <doi:10.1162/089120103322711569> <<https://www.kilgarriff.co.uk/Publications/2003-KilgGrefenstette-WACIntro.pdf>>6. Перечень учебно-методических материалов по самостоятельной работе обучающихся

6. Перечень учебно-методических материалов по самостоятельной работе обучающихся

Материалы курса размещаются на портале eduportal.nsu.ru и на гугл-диске НГУ. Они включают презентации лекций, список литературы, список вопросов к зачету, ссылки на электронные ресурсы, инструкции для подготовки рефератов по онлайн корпусам и проектов создания корпуса

7. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины

Освоение дисциплины используются следующие ресурсы:

- электронная информационно-образовательная среда НГУ (ЭИОС);
- образовательные интернет-порталы;
- информационно-телекоммуникационная сеть Интернет.

Взаимодействие обучающегося с преподавателем (синхронное и (или) асинхронное) осуществляется через личный кабинет студента в ЭИОС.

7.1 Современные профессиональные базы данных:

Национальный корпус русского языка: www.ruscorpora.ru

7.2. Ресурсы сети Интернет

- Семинар «Корпусная лингвистика» СПбГУ <<http://corpora.iling.spb.ru>>
- Конференция Corpora Санкт-Петербург <<http://corpora.phil.spbu.ru>>
- Сообщество «Письменное наследие» <<http://textualheritage.org>>
- Школа El'Manuscript 2015 <http://gf.nsu.ru/www/?page_id=12388>
- Text Encoding Initiative (TEI): <<http://www.tei-c.org>>
- Corpus Encoding Standard for XML (XCES): <<http://www.xces.org>>

- Рассылка Corpora <<http://mailman.uib.no/listinfo/corpora>>
- Платформа ТХМ (корпус менеджер) <<http://textometrie.org>>

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

8.1 Перечень программного обеспечения

Для обеспечения реализации дисциплины используется стандартный комплект программного обеспечения (ПО), включающий регулярно обновляемое лицензионное ПО Windows и MS Office. Кроме того, в компьютерном классе устанавливаются следующее свободно распространяемое специализированное ПО:

Платформа ТХМ, с сайта <http://texometrie.org>

Текстовый редактор NotePad++

Редактор XML Copy Editor с сайта <https://sourceforge.net/projects/xml-copy-editor>

OpenOffice или LibreOffice

8.2 Информационные справочные системы

«Не используются».

9. Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине

Для реализации дисциплины используются специальные помещения:

1. Учебные аудитории для проведения занятий лекционного типа, занятий семинарского типа, курсового проектирования (выполнения курсовых работ), групповых и индивидуальных консультаций, текущего контроля, промежуточной и итоговой аттестации;

2. Помещения для самостоятельной работы обучающихся;

Учебные аудитории укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду НГУ.

Для проведения занятий лекционного типа предлагаются следующие наборы демонстрационного оборудования и учебно-наглядных пособий:

- комплект лекций-презентаций по темам дисциплины.

Материально-техническое обеспечение образовательного процесса по дисциплине для обучающихся из числа лиц с ограниченными возможностями здоровья осуществляется согласно «Порядку организации и осуществления образовательной деятельности по образовательным программам для инвалидов и лиц с ограниченными возможностями здоровья в Новосибирском государственном университете».

Реализация дисциплины может осуществляться с применением дистанционных образовательных технологий.

10. Оценочные средства для проведения текущего контроля и промежуточной аттестации по дисциплине

Перечень результатов обучения по дисциплине в виде индикаторов достижения компетенций, выраженных в знаниях, умениях и владениях представлен в разделе 1.

10.1 Порядок проведения текущего контроля и промежуточной аттестации по дисциплине

Текущий контроль успеваемости:

Студенты получают индивидуальные задания для самостоятельной работы и для проектной работы командами из 2-3 человек. Для индивидуальной работы предлагается

написание реферата и подготовка презентации по корпусу в интернете или по определенной методологической проблеме. В рамках проекта предлагается создание компьютерного корпуса, направленного на решение определенной учебной, научной или практической задачи. Более всего приветствуется создание корпуса, который может пригодиться студентам в их магистерских диссертациях.

Промежуточная аттестация:

Промежуточная аттестация по дисциплине проводится в форме зачёта. К сдаче зачёта допускаются студенты, подготовившие и представившие на семинаре (или загрузившие на платформу электронного обучения) реферат по корпусу в интернете или принявшие участие в проекте создания корпуса. Зачёт проводится в форме экзамена по билетам, содержащим один теоретический вопрос и практическое задание. Результаты прохождения аттестации оцениваются по шкале «незачет», «зачет».

10.2. Описание критериев и шкал оценивания результатов обучения по дисциплине

Таблица 10.1

Код компетенции	Результат обучения по дисциплине	Оценочное средство
ОПК-4	Знать основные этапы развития и актуальные тенденции корпусной лингвистики; основные термины и методы корпусных исследований, основные операторы языка запросов CQL.	Вопросы зачета
	Уметь оценить основные характеристики и грамотно использовать доступные в Интернете лингвистические корпуса; спроектировать и реализовать на платформе ТХМ корпус текстов или других языковых материалов, нацеленный на решение определенной исследовательской задачи.	Реферат
	Иметь навыки использования: основных операторов языка CQL, регулярных выражений, языка разметки XML, инструментов импорта, количественного и качественного анализа корпуса на платформе ТХМ	Проект
ПК-1	Знать теорию и практику создания компьютерных корпусов в лингвистических исследованиях	Вопросы зачета
	Уметь применять компьютерные корпуса в лингвистических исследованиях	Проект
	Владеть принципами создания электронных языковых ресурсов (текстовых, речевых и мультимодальных корпусов; словарей, тезаурусов, онтологий; фонетических, лексических, грамматических и иных баз данных и баз знаний) и навыками использования таких ресурсов	Проект

Таблица 10.2

Критерии оценивания результатов обучения	Шкала оценивания
<p><u>Зачет:</u> Студент демонстрирует базовые знания теорий и методов исследования корпусной лингвистики, в состоянии их использовать и делать собственные выводы, представлять полученные результаты.</p> <p><u>Реферат:</u> Обоснованность теоретическим и фактическим материалом, подкрепленным ссылками на научную литературу и источники, корректность и адекватность выбранных методов анализа источников, языковых фактов и их интерпретации, логичность и аргументированность изложения материала, точность и корректность применения терминов и понятий корпусной лингвистики.</p> <p><u>Проект:</u> Поставленные задачи выполнены полностью.</p>	<i>зачет</i>
<p><u>Зачет:</u> Студент допускает грубые ошибки в описании современных теорий и методов корпусной лингвистики, не может аргументировано отстаивать собственную точку зрения, отсутствуют ответы на дополнительные вопросы.</p> <p><u>Реферат:</u> Отсутствие теоретического и фактического материала, подкрепленного ссылками на научную литературу и источники, отсутствие анализа языковых фактов и их интерпретации, компилятивное, неосмысленное, нелогичное и неаргументированное изложение материала, грубые ошибки в применении терминов и понятий корпусной лингвистики.</p> <p><u>Проект:</u> Поставленные задачи не выполнены или выполнены неверно.</p>	<i>незачет</i>

10.3. Типовые контрольные задания и иные материалы, необходимые для оценки результатов обучения

Перечень вопросов для зачёта:

1. Корпусные исследования без применения компьютера.
2. Корпусная лингвистика в 1960-1980-е гг. Основные направления. Брауновский корпус. Корпус LOB.
3. Основные тенденции корпусных исследований после 2000 г.
4. Место корпусной лингвистики среди гуманитарных и компьютерных наук.
5. Использование корпусов в традиционной лингвистике.
6. Что такое корпус? Критерии определения корпуса.
7. Понятие текста применительно к корпусным исследованиям.
8. Основные понятия корпусной лингвистики: единица корпуса, метаданные, текст-форма, токен, вхождение, тэг, конкорданс (KWIC), частотность, специфичность, совместная встречаемость.
9. Основные виды разметки текстов в корпусе, их взаимозависимость.
10. Типология корпусов.
11. Основные корпуса русского языка.
12. Британский национальный корпус.

13. Интернет как корпус.
14. Язык XML: основные правила.
15. Инициатива по кодированию текстов TEI: цели создания, принципы организации
16. Технологический процесс создания корпуса: основные этапы
17. Методы анализа корпусов: конкордансы, частотные словари, язык запросов CQL
18. Правовые аспекты публикации корпуса. Открытые лицензии.

Задания для самостоятельной работы

Проект «Импортирование текстов доступного в интернете корпуса любого языка в корпус-менеджер TXM»

(группа 2-3 человека)

Задание: найти в интернете корпус с открытыми источниками, подготовить пакет для импорта этих источников с помощью модуля XTZ + CSV, с сохранением:

- метаданных
- всех видов разметки, присутствующей в корпусе.

Проект «Корпус новогодних обращений на русском языке»

(группа 2-3 человека)

Задание: собрать тексты новогодних обращений Президентов РФ, Президента СССР и генеральных секретарей ЦК КПСС за как можно более длительный период в формате TXT, подготовить таблицу с метаданными (год, фамилия говорящего) и импортировать в TXM с использованием модуля TXT+CSV.

Проект подготовки корпуса

(возможно группой 2-3 человека)

Презентация проекта должна включать:

1. Цель создания корпуса, например:
 - сравнить язык нескольких произведений одного автора или одного жанра
 - сравнить язык оригинала произведения и его перевода
 - ...
2. Принципы отбора и описания текстов (единиц корпуса), отвечающие целям корпуса
3. Методы оцифровки / верификации качества цифрового текста
4. Разметка корпуса
 - виды разметки, ее формат
5. Анализ корпуса с использованием корпус-менеджера
 - обоснование выбора корпус-менеджера, его основные функции
 - общее описание корпуса (размер в текстоформах и текстовых единицах, архитектура корпуса)
 - результаты качественного и количественного анализа
6. Правовой статус источников корпуса и условия распространения
7. Проблемы, возникшие при подготовке и анализе корпуса и их решение

Темы докладов и эссе

Студентам предлагается либо подготовить доклад по какому-либо национальному или специализированному корпусу, доступному в Интернете, либо подготовить реферат по одной из предложенных тем на основе научных статей.

ОБЗОР КОРПУСА

Выбрать какой-либо доступный в интернете корпус (кроме русского и британского) и подготовить обзор из следующих пунктов:

- история создания
- размеры корпуса
- типология текстов
- условия доступа
- виды разметки
- инструменты анализа / интерфейс, язык запросов
- использование в научной / учебной работе

Для выбора корпуса можно воспользоваться ссылками, данными в учебниках: [Захаров и Богданова 2013] или [Копотев 2014].

ТЕМЫ ДЛЯ РЕФЕРАТОВ

1) Сравнительный анализ принципов морфологической разметки в НКРЯ и КРЛЯ:

Ляшевская О.Н., Плунгян В.А., Сичинава Д.В. О морфологическом стандарте Национального корпуса русского языка // Национальный корпус русского языка: 2003 – 2005. Результаты и перспективы. М., 2005. С. 111 – 135.

<<http://ruscorpora.ru/sbornik2005/08lashevs.pdf>>

Венцов А.В., Грудева Е.В., Касевич В.Б. Морфологическая проблематика в Национальном корпусе русского литературного языка // Международная конференция «Корпусная лингвистика – 2004»: Тезисы докладов (12 – 14 октября 2004 г., С.-Петербург). СПб.: СПбГУ, 2004. С. 18 – 20. <<http://www.narusco.ru/PUBV006>>

2) Проблемы типологии текстов в Британском национальном корпусе

Lee, D. Y. W. 2001. 'Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle', Language Learning and Technology 5 (3): 37–72. <<http://lt.msu.edu/vol5num3/lee/default.html>>

3) Методика «многомерного анализа» текстов и ее критика

Biber, D. A typology of English texts // Linguistics, 27, 1989: 3-43

Критика:

[Lee 2001] (см. выше)

Учебник [McEnery & Hardie 2011: 104-115]

Altenberg, B. Review of D. Biber (1988), Variation across Speech and Writing. Studia Linguistica 43 (2), 1989: 167–74.¹

Doyle, P. 2005. Replicating corpus-based linguistics: investigating lexical networks in text // Proceedings from Corpus Linguistics 2005. University of Birmingham.

<<http://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2005-journal/Lexiconodf/COLING2005paper.pdf>>

Возможно также подготовить реферат по статьям на выбор студента из списка в учебниках по курсу (желательно представить несколько точек зрения на вопрос).

Оценочные материалы по дисциплине (приложение 2), предназначенные для проверки соответствия уровня подготовки по дисциплине требованиям ФГОС, хранятся на кафедре-разработчике РПД в печатном и/или электронном виде.

