МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное
учреждение высшего образования
«ТЮМЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

УТВЕРЖДАЮ
Первый проректор
А.В. Толстиков
29 марта 2022 г.

КОРПУСНАЯ (КВАНТИТАТИВНАЯ) ЛИНГВИСТИКА
И НОВЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

Рабочая программа
для обучающихся по научной специальности
5.9.8 Теоретическая, прикладная и сравнительно-сопоставительная лингвистика
форма обучения (очная)

Табанакова В.Д. Корпусная (квантитативная) лингвистика и новые информационные технологии. Рабочая программа для обучающихся по научной специальности 5.9.8. Теоретическая, прикладная и сравнительно-сопоставительная лингвистика, форма обучения (очная). Тюмень, 2022.

Рабочая программа составлена в соответствии с требованиями ФГТ, утверждёнными приказом МИНОБРНАУКИ России от 20.10.2021 г. № 951.

Рабочая программа дисциплины опубликована на сайте ТюмГУ: Корпусная (квантитативная) лингвистика и новые информационные технологии [электронный ресурс] / Режим доступа: https://www.utmn.ru/sveden/education/#.

1. Цели и задачи освоения дисциплины

Цель дисциплины – сформировать у аспирантов знания и навыки использования корпусных (квантитативных) методов и информационных технологий для осуществления грамотных лингвистических исследований. Данная цель ставит следующие **задачи**:

1. сформировать у аспирантов знания о границах и содержании таких направлений современной лингвистики, как квантитативная, компьютерная и корпусная лингвистика;
2. выработать у аспирантов умения и навыки использования терминологического аппарата и методов данных направлений.

2. Компетенции обучающегося, формируемые в результате освоения данной дисциплины:

УК-1 - способность к критическому анализу и оценке современных научных достижений, генерирование новых идей при решении исследовательских и практических задач, в том числе в междисциплинарных областях.

ОПК-1 - способность самостоятельно осуществлять научно-исследовательскую деятельность в соответствующей профессиональной области с использованием современных методов исследования и информационно-коммуникационных технологий.

ПК-8 - готовность проводить самостоятельные научные исследования в области общей теории языкознания, социолингвистики, психолингвистики, теории дискурса и теории отдельных языков, опираясь на систему основных понятий и категориальный аппарат современной теоретической лингвистики, педагогики, психологии, когнитивных и информационных наук для решения исследовательских задач.

3. Структура и объем дисциплины

Таблица 1

Вид учебной работы	**Всего часов** (академические часы)	**Часов в семестре** (академические часы)
		4
Общий объем зач. ед.	3	3
час	108	108
Из них:		
Часы аудиторной работы (всего):	22	22
Лекции	12	12
Практические занятия	10	10
Лабораторные / практические занятия по подгруппам	0	0
Часы внеаудиторной работы, включая самостоятельную работу обучающегося	50	50
Вид промежуточной аттестации (зачет, диф. зачет, экзамен)	36	Дифференцированный зачет 36

4. Система оценивания

Оценивание работы аспирантов в течение семестра осуществляется на основе балльно-рейтинговой системы. Баллы начисляются аспирантам за следующие активности:

1) посещение занятия – 1 балл;

2) выполнение заданий на практическом занятии: 0-5 балла;

3) подготовка лабораторных и рефератов: 0-5 балла.

Для получения зачета по дисциплине необходимо набрать не менее 61 балла. Аспиранты, набравшие по итогам работы в семестре менее 61 балла, сдают зачет по дисциплине в письменной форме. Такой зачет проводится в форме письменного изложения знаний, которые есть у аспиранта по одной из тем (список тем приводится далее), и имеет целью выявление уровня освоения дисциплины, характеризующего знания обучающегося в соответствии с определенными компетенциями.

Зачет проводится в форме письменной работы.

Темы для подготовки к зачету:

1. Количественные исследования в языкознании.
2. Соотношение корпусной и прикладной лингвистики.
3. Предмет, объект, цели и задачи корпусной лингвистики.
4. Статистические модели языка. Закон Ципфа.
5. Выборка. Виды выборки.
6. Переменная. Виды переменных.
7. Статистическая значимость. Нормальное распределение.
8. Корреляция. Коэффициент корреляций.
9. Показатели центральной тенденции: средняя арифметическая, мода, медиана.
10. Мера рассеяния признака. Показатели меры рассеяния признака.
11. Русскоязычные корпуса.
12. Интерпретация результатов, полученных методом статистического анализа.
13. Текстология и автороведение. Атрибуция текста.
14. Статистические характеристики гендера, возраста, социального статуса, происхождения автора текста.
15. Корпуса языков мира.
16. Инструменты анализа корпусов.
17. Новые информационные технологии в лингвистике.
18. Автоматизированные системы обработки устной и письменной речи. Парсинг.
19. Поисковые системы. Автоматическое индексирование, аннотирование и реферирование текстов.
20. Системы управления базами данных.
21. Системы машинного перевода.
22. Виды лингвистических корпусов.
23. Компьютерные модели языка.
24. Когнитивная лингвистика и модели представления знаний.

5. Содержание дисциплины

5.1. Тематический план дисциплины

Таблица 2

№ п/п	Наименование тем и/или разделов	Объем дисциплины (модуля), час.				
		Всего	Виды аудиторной работы (академические часы)			Иные виды контактной работы
			Лекции	Практические занятия	Лабораторные/ практические занятия по	

					подгруппам	
1	2	3	4	5	6	7
1	Параметризация языковых единиц	4	2	2	0	0
2	Метод статистического анализа текста	4	2	2	0	0
3	Меры рассеяния признака	4	2	2	0	0
4	Меры рассеяния признака - 2	4	2	2	0	0
5	Новые информационные технологии в лингвистике	4	2	2	0	0
6	Компьютерный анализ текста	4	2	0	0	0
7	Дифференцированный зачет	36	0	0	0	36
	Итого (часов)	58	12	10	0	36

5.2. Содержание дисциплины по темам

1. "Параметризация языковых единиц"

Статистические модели языка. Закон Ципфа. Определение необходимости проведения лингвостатистического анализа. Параметризация языковых единиц.

Выборка. Переменные. Виды переменных. Шкала переменных. Группирующие переменные. Ранжирование. Статистическая значимость. Нормальное распределение.

2. "Метод статистического анализа текста"

Корреляция. Коэффициент корреляций. Показатели центральной тенденции: средняя арифметическая, мода, медиана.

Рассчитать показатели центральной тенденции, составить график.

3. "Меры рассеяния признака"

Показатели меры рассеяния признака: лимиты, вариационный размах, среднее линейное отклонение, дисперсия, среднее квадратичное отклонение.

3. "Меры рассеяния признака"

Как создать выборку из определенных языковых единиц любого уровня на основе предложенного текста; обосновать принципы составления выборки; для решения каких задач можно считать данную выборку достаточной. Коэффициент вариации, квадратичная ошибка средней, t-критерий Стьюдента. Как рассчитать коэффициент вариации, квадратичную ошибку средней, t-критерий Стьюдента.

5. "Новые информационные технологии в лингвистике"

Текстология и автороведение. Атрибуция текста. Определение авторства текста. Статистические характеристики гендера, возраста, социального статуса, происхождения автора текста. Крупномасштабные проекты в рамках корпусной лингвистики: Национальный

Корпус Русского Языка (http://www.ruscorpora.ru/), WordNet (http://wordnet.princeton.edu/). Работа с системами анализа корпусов.

6. "Компьютерный анализ текста"

Автоматизированные системы обработки устной и письменной речи. Парсинг. Стемминг. Поисковые системы. Автоматическое индексирование, аннотирование и реферирование текстов. Системы управления базами данных. Системы машинного перевода. Системы анализа и синтеза устной речи. Знакомство с системами автоматической обработки текста, осуществляющими грамматический и лексический разбор.

6. Учебно-методическое обеспечение и планирование самостоятельной работы обучающихся

Таблица 3

№ Темы	Темы	Виды СРС
1	Параметризация языковых единиц	Чтение обязательной и дополнительной литературы Проработка лекций
2	Метод статистического анализа текста	Чтение обязательной и дополнительной литературы Проработка лекций
3	Меры рассеяния признака	Чтение обязательной и дополнительной литературы Проработка лекций
4	Меры рассеяния признака	Чтение обязательной и дополнительной литературы Проработка лекций
5	Новые информационные технологии в лингвистике	Чтение обязательной и дополнительной литературы Проработка лекций
6	Компьютерный анализ текста	Чтение обязательной и дополнительной литературы Проработка лекций

7. Промежуточная аттестация по дисциплине

7.1. Оценочные материалы для проведения промежуточной аттестации

Устный опрос проводится по теоретическому материалу на практических занятиях. Может проводиться в форме индивидуального собеседования или собеседования в малых группах по вопросам.

Письменная работа проводится по теоретическому и практическому материалу на практических занятиях в форме письменного ответа на вопросы.

Конспект (реферат) задается по теоретическому материалу на практических занятиях в форме письменного изложения в том числе в сокращенной форме выдержек из научной литературы.

Варианты заданий

УСТНЫЙ ОПРОС:

1. Каковы основные направления прикладной лингвистики? В чем ее отличие от теоретической лингвистики?
2. В чем различие между квантитативной и комбинаторной лингвистикой?
3. В каких лингвистических исследованиях применяются стохастические модели?
4. В каких исследованиях применяется метод лингвистических переменных?
5. Какие лингвистические методы основаны на количественном подсчете? Почему?
6. Какие лингвистические задачи решает параметризация языковых единиц?

ПИСЬМЕННАЯ РАБОТА (выполняется на персональных компьютерах):

1. Посмотреть в Интернете, как существуют на сегодня в России лаборатории, кафедры и институты прикладной, компьютерной, математической лингвистики. Описать сферу их деятельности, достижения.
2. Создать проект лаборатории квантитативной лингвистики. Распланировать размер помещения, необходимое оборудование и программное обеспечение, персонал. Составить круг направлений в работе лаборатории.

ТЕМЫ КОНСПЕКТОВ (РЕФЕРАТОВ):

1. Различия квантитативной и комбинаторной лингвистики.
2. Применение статистических методов в автороведении: возможности, ограничения.
3. Переход от количественных данных к качественному анализу в лингвостатистике.
4. Применение корпусной лингвистики в современной лексикографии.
5. Связь корпусной лингвистики с Web 2.0.

8. Учебно-методическое и информационное обеспечение дисциплины

8.1. Основная литература:

1. Кулаичев, А. П. Методы и средства комплексного анализа данных / А. П. Кулаичев. - 4-е изд., перераб. и доп. - Москва : НИЦ ИНФРА-М, 2016. - 511 с. - ISBN 978-5-16-104593-0 (online). - Текст : электронный. - URL: https://znanium.com/catalog/product/548836 (дата обращения: 22.03.2022). – Режим доступа: по подписке.

2. Ляшевская, О. Н. Корпусные инструменты в грамматических исследованиях русского языка = Corpus Instruments for Russian Grammar Studies/ О. Н. Ляшевская. - Москва: ЯСК ; Рукописные памятники Древней Руси, 2016. - 520 с.

3. Рик, Гаско Простой Python просто с нуля / Гаско Рик. — Москва : СОЛОН-Пресс, 2019. — 256 с. — ISBN 978-5-91359-334-4. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/94940.html (дата обращения: 23.03.2022). — Режим доступа: для авторизир. пользователей

8.2. Дополнительная литература:

1. Шапкин, А. С. Задачи с решениями по высшей математике, теории вероятностей, математической статистике, математическому программированию : учебное пособие для бакалавров / А. С. Шапкин, В. А. Шапкин. — 9-е изд., стер. — Москва : Издательско-торговая корпорация «Дашков и К°», 2020. — 432 с. - ISBN 978-5-394-03710-8. - Текст : электронный. - URL: https://znanium.com/catalog/product/1091871 (дата обращения: 22.03.2022). – Режим доступа: по подписке.

2. Хроленко, А. Т. Современные информационные технологии для гуманитария: практ. рук./ А. Т. Хроленко, А. В. Денисов. - Москва: Флинта: Наука, 2008. - 128 с.

3. Ржевский, С. В. Высшая математика I: линейная алгебра и аналитическая геометрия : учебное пособие / С.В. Ржевский. — Москва : ИНФРА-М, 2019. — 211 с. - ISBN 978-5-16-108269-0. - Текст : электронный. - URL: https://znanium.com/catalog/product/1065260 (дата обращения: 22.03.2022). – Режим доступа: по подписке.

4. Белозерова Н.Н. Шекспир и компания, или Использование электронных библиотек при лингвистическом исследовании: учеб. пособие/ Н.Н. Белозерова, Л.Е. Чуфистова. Тюмень: Изд-во ТюмГУ, 2011. 296 с.
5. Гвишиани Н.Б. Практикум по корпусной лингвистике: учеб. пособие по англ. яз. / Н.Б. Гвишиани. Москва: Высшая школа, 2008. 191 с.

8.3. Интернет-ресурсы:

1. https://colab.research.google.com/
2. drive.google.com
3. https://ruscorpora.ru/new/index.html
4. http://www.bl.uk/manuscripts/Browse.aspx
5. http://www.aot.ru/
6. https://gate.ac.uk/
7. http://opencorpora.org/
8. http://api.yandex.ru/speller/
9. http://books.google.ru/
10. http://scholar.google.com/
11. http://translate.google.com/
12. http://translate.yandex.ru/
13. http://wordnet.princeton.edu/
14. http://www.coli.uni-saarland.de/page.php?id=whatis
15. http://www.loa-cnr.it/DOLCE.html
16. http://www.ontologyportal.org/
17. http://www.opencyc.org/
18. http://www.ruscorpora.ru/
19. http://www.wikipedia.org/
20. Word-Tabulator – http://www.rvb.ru/soft/index.html
21. Морфологический анализатор – http://starling.rinet.ru/morph.htm
22. Программная среда DOE: http://www.eurecom.fr/~troncy/DOE/
23. Программная среда Protégé: http://protege.stanford.edu/download/protege/4.0/installanywhere/

8.4. Современные профессиональные базы данных и информационные справочные системы:

1. Знаниум - https://new.znanium.com/
2. Лань - https://e.lanbook.com/
3. IPR Books - http://www.iprbookshop.ru/
4. Elibrary - https://www.elibrary.ru/
5. Национальная электронная библиотека (НЭБ) - https://rusneb.ru/
6. Межвузовская электронная библиотека (МЭБ) - https://icdlib.nspu.ru/
7. "ИВИС" (БД периодических изданий) - https://dlib.eastview.com/browse

Электронная библиотека Тюмгу - https://library.utmn.ru/

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

- **Лицензионное ПО:**

1. платформа для электронного обучения Microsoft Teams
2. Microsoft Excel
3. Microsoft Speech API
4. Microsoft Word
5. Бесплатное программное обеспечение для статистического анализа R

6. Бесплатное программное обеспечение для статистического анализа Matplotlib для Python 3

10. Технические средства и материально-техническое обеспечение дисциплины

Для лекционных занятий мультимедийная учебная аудитория для проведения занятий лекционного типа, оснащенная проектором и компьютером. Обеспечено проводное подключение ПК к локальной сети и сети Интернет.

Для лабораторных занятий и самостоятельной работы компьютерный класс, оснащенный персональными компьютерами (ПК). На ПК установлено следующее программное обеспечение: операционная система MS Windows, офисный пакет MS Office, платформа MS Teams, антивирусное ПО. Обеспечено проводное подключение ПК к локальной сети и сети Интернет.

11. Средства адаптации преподавания дисциплины (модуля) к потребностям лиц с ограниченными возможностями

В случае необходимости, обучающимся из числа лиц с ограниченными возможностями здоровья (по заявлению обучающегося) могут предлагаться одни из следующих вариантов восприятия информации с учетом их индивидуальных психофизических особенностей, в том числе с применением электронного обучения и дистанционных технологий:

- для лиц с нарушениями зрения: в печатной форме увеличенным шрифтом; в форме электронного документа; в форме аудиофайла (перевод учебных материалов в аудиоформат); индивидуальные задания и консультации.
- для лиц с нарушениями слуха: в печатной форме; в форме электронного документа; индивидуальные задания и консультации.
- для лиц с нарушениями опорно-двигательного аппарата: в печатной форме; в форме электронного документа; в форме аудиофайла; индивидуальные задания и консультации.

12. Методические рекомендации обучающимся по выполнению самостоятельной работы

Проработка лекционного материала и конспектирование источников по теме требует систематизации концептуальных понятий и схематизации текстового изложения проблем. Для этого:

- избегайте копирования текстовых фрагментов,
- не переписывайте текстов определений терминов, а старайтесь их перефразировать и строить свои собственные краткие родовидовые определения,
- выстраивайте логико-понятийные схемы, отражающие отношения между специальными понятиями,
- формулируйте проблемы и определения понятий в нескольких вариантах.

При подготовке к практическим занятиям используйте свои собственные записи проработанных источников и моделируйте варианты выполнения практических заданий по теме.

При подготовке к дифференцированному зачету необходимо:

- грамотно и правильно использовать в ответах лингвистическую, статистическую и общенаучную терминологию;
- безошибочное владение категориальным аппаратом науки;
- умение обозначить основные проблемы сформулированных в билетах вопросов;
- безошибочное знание фактического материала;
- умение связать ответ на вопрос с темой диссертационного исследования;
- логичность, связность ответа.