Программа учебной дисциплины «Корпусные методы исследований языковых процессов»

Утверждена
Академическим советом ОП
Протокол № от __.__.20__

Разработчик	Попова Дарья Павловна, преподаватель, Школа лингвистики
Число кредитов	3
Контактная работа (час.)	32
Самостоятельная работа (час.)	82
Курс, Образовательная программа	1 курс, магистерская программа «Языковая политика в условиях этнокультурного разнообразия»
Формат изучения дисциплины	без использования онлайн курса

1. Цель, результаты освоения дисциплины и пререквизиты

Целями освоения дисциплины «Корпусные методы исследований языковых процессов» являются:

- знакомство с лингвистическими и социолингвистическими корпусами;
- знакомство с принципами аннотирования лингвистических и социолингвистических данных и с проблемами, возникающими при аннотировании данных;
- изучение основ квантитативного анализа в социолингвистике;
- ознакомление с возможностями квантитативных подходов в социолингвистике и с проблемами, с которыми они сталкиваются;
- умение формулировать исследовательские вопросы и представлять их в виде гипотез, которые можно протестировать квантитативными методами;
- умение критически оценивать качество статистического анализа;
- умение применять подходящие для целей исследования статистические методы к данным;
- умение программировать в R для самостоятельного решения исследовательских задач.

В результате освоения дисциплины студент должен:

знать:

- основные статистические методы анализа языковых данных;
- реализацию статистических методов в R;
- основные корпусные методы анализа данных.

уметь:

- производить поиск и анализ релевантной информации в лингвистических корпусах;
- форматировать лингвистические данные;
- оценивать адекватность проведенного статистического анализа;
- формулировать исследовательскую гипотезу;
- подбирать подходящий метод статистического анализа;

- проводить статистический анализ данных в R;
- оценивать степень достоверности результатов, полученных с помощью экспериментальных, корпусных или математических методов исследования;
- ориентироваться в потоке статистической информации.

владеть:

- навыками поиска по лингвистическим корпусам;
- навыками форматирования данных;
- методами статистического описания данных;
- навыками применения основных методов статистического анализа;
- навыками описания статистических исследований.

Изучение дисциплины «Корпусные методы исследований языковых процессов» не имеет пререквизитов: программа не предполагает, что студенты обладают навыками статистического анализа и знанием статистики.

Настоящая дисциплина относится к блоку обязательных дисциплин программы. В результате освоения данного курса студент должен уметь работать с корпусными данными, уметь критически оценивать использование статистических методов, владеть навыками программирования на языке R и уметь применять статистические методы для анализа данных. Эти навыки должны быть использованы студентом в дальнейшем при подготовке рефератов, написании курсовых, статей, проектов.

2. Содержание учебной дисциплины

Тема 1 **Корпусные исследования**

Основные понятия корпусных исследований: корпус, аннотация (разметка), поиск. Знакомство с существующими лингвистическими и социолингвистическими корпусами. Область применения корпусных исследований. Проблемы, возникающие при проведении корпусных исследований.
Количество часов аудиторной работы – 4 часа семинара. Общий объем самостоятельной работы: 8 часов. Самостоятельная работа подразумевает подготовку к тесту. Для освоения раздела предусмотрено проведение дискуссий, решение задач и рассмотрение опубликованных анализов данных.

Тема 2. **Представление данных и манипуляции с данными**

Знакомство с R. Понятие случайной величины.
Представление данных, сортировка данных в столбцах, строках. Форматы данных.
Простые графики.
Количество часов аудиторной работы – 4 часа семинара. Общий объем самостоятельной работы: 8 часов. Самостоятельная работа подразумевает подготовку домашнего задания.
Для освоения раздела предусмотрено проведение дискуссий, решение задач и рассмотрение опубликованных анализов данных на семинарах.

Тема 3. **Статистические распределения**

Понятие статистического распределения. Виды распределений. Нормальное распределение, распределения t, F, χ^2.

Количество часов аудиторной работы – 4 часа семинара. Общий объем самостоятельной работы: 8 часов. Самостоятельная работа подразумевает подготовку домашнего задания.
Для освоения раздела предусмотрено проведение дискуссий, решение задач и рассмотрение опубликованных анализов данных на семинарах.

Тема 4. **Базовые статистические методы**

Тесты для определения вида распределения. Зависимые и независимые переменные. Линейная регрессия. Ковариантность. Статистическая значимость.
Количество часов аудиторной работы – 4 часа семинара. Общий объем самостоятельной работы: 8 часов. Самостоятельная работа подразумевает подготовку домашнего задания.
Для освоения раздела предусмотрено проведение дискуссий, решение задач и рассмотрение опубликованных анализов данных на семинарах.

Тема 5. **Кластеризация и классификация**

Кластеризация – метод главных компонент, факторный анализ, иерархический кластерный анализ, correspondence analysis, multi-dimensional scaling. Классификация -- классификационные деревья.
Количество часов аудиторной работы – 4 часа семинара. Общий объем самостоятельной работы: 8 часов. Самостоятельная работа подразумевает подготовку домашнего задания.
Для освоения раздела предусмотрено проведение дискуссий, решение задач и рассмотрение опубликованных анализов данных на семинарах.

Тема 6. **Регрессионные методы**

Моделирование регрессии.
Количество часов аудиторной работы – 4 часа семинара. Общий объем самостоятельной работы: 8 часов. Самостоятельная работа подразумевает подготовку к тесту.
Для освоения раздела предусмотрено проведение дискуссий, решение задач и рассмотрение опубликованных анализов данных на семинарах.

Тема 7. **Модели со смешанным эффектом**

Использование моделей со смешанным эффектом.
Количество часов аудиторной работы – 4 часа семинара. Общий объем самостоятельной работы: 8 часов. Самостоятельная работа подразумевает подготовку к выполнению финального проекта.
Для освоения раздела предусмотрено проведение дискуссий, решение задач и рассмотрение опубликованных анализов данных на семинарах.

Тема 8. **Графическое представление данных**

Визуализация пройденных методов в R.
Количество часов аудиторной работы – 4 часа семинара. Общий объем самостоятельной работы: 8 часов. Самостоятельная работа подразумевает подготовку к выполнению финального проекта.

Для освоения раздела предусмотрено проведение дискуссий, решение задач и рассмотрение опубликованных анализов данных на семинарах.

18 часов самостоятельной работы отводится на работу над финальным проектом.

3. Оценивание

При выполнении домашних заданий, тестовых заданий и финального проекта студент должен продемонстрировать, что владеет соответствующим материалом, правильно употребляет пройденные статистические и корпусные методы, может обосновать свои решения, может критически оценить как собственный, так и чужой корпусный и статистический анализ данных.

Текущий контроль осуществляется с помощью тестов в рамках аудиторных занятий и домашних заданий.

Тесты включают в себя ряд вопросов, проверяющих усвоение материала соответствующих лекций и семинаров, – знание основных понятий, способность видеть явные ошибки в статистическом анализе, способность сформулировать тестируемую гипотезу, правильное понимание применения различных корпусных и статистических методов.

Домашние задания нацелены на развитие навыков решения статистических задач с помощью R.

Финальный проект представляет собой описание статистического анализа лингвистических данных, проведённого студентом. Студент должен отформатировать предоставленные данные, выдвинуть гипотезу, протестировать её с помощью подходящего статистического теста, проанализировать результат и описать проделанную работу.

Результирующая оценка рассчитывается по формуле:

$$O_{результ} = 0,4 *\cdot O_{проект} + 0,1 * O_{дз1} + 0,1 * O_{дз2} + 0,1 * O_{дз3} + 0,1 * O_{дз4} + 0,1 * O_{тест1} + 0,1 * O_{тест2}$$

Оценки по всем формам контроля выставляются по 10-ти балльной шкале (https://www.hse.ru/studyspravka/Scale/). Блокирующих элементов не предусмотрено. Пересдача проекта проводится в течение одной недели с первоначальной даты. Результирующая оценка округляется в пользу студента. В диплом выставляется результирующая оценка по учебной дисциплине.

4. Примеры оценочных средств

Блокирующих элементов не предусмотрено. Примеры тем финального проекта:

на основании файла с данными опроса в школах проекта «Языки Москвы», проведите статистический анализ взаимодействия одной или нескольких пар переменных:

1) язык, на котором ребёнку в детстве читали сказки и язык, который ребёнок считает родным;
2) родной язык мамы и родной язык ребёнка в случае, когда у родителей разные родные языки;
3) язык, на котором ребёнок разговаривает с родителями/с бабушками и дедушками, и язык, на котором ребёнок разговаривает с друзьями.

5. Ресурсы

5.1 Рекомендуемая основная литература

1. Материалы лекций.
2. Документация по языку/программе статистического анализа R: https://cran.r-project.org/manuals.html

5.2 Программное обеспечение

п/п	Наименование	Условия доступа
1.	MicrosoftWindows 7 Professional RUS MicrosoftWindows 10 MicrosoftWindows 8.1 Professional RUS	*Из внутренней сети университета (договор)*
2.	Программа статистического анализа R	*Свободно распространяемое ПО* https://www.r-project.org/

5.3 Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

п/п	Наименование	Условия доступа
	Профессиональные базы данных, информационно-справочные системы	
1.	Консультант Плюс	*Из внутренней сети университета (договор)*
2.	Электронно-библиотечная система Юрайт	URL: https://biblio-online.ru/
	Интернет-ресурсы (электронные образовательные ресурсы)	
1.	Национальный корпус русского языка	URL: www.ruscorpora.ru
2.	Открытый бесплатный курс DataCamp *Introduction to R*	URL: https://www.datacamp.com/courses/free-introduction-to-r
3.	Corpus of Contemporary American English	URL: https://corpus.byu.edu/coca/

5.4 Материально-техническое обеспечение дисциплины

Учебные аудитории для семинаров по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

– ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);

– мультимедийный проектор с дистанционным управлением.

Учебные аудитории для лабораторных и самостоятельных занятий по дисциплине оснащены ПЭВМ, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.

6. Особенности организации обучения для лиц с ограниченными возможностями здоровья и инвалидов

В случае необходимости, обучающимся из числа лиц с ограниченными возможностями здоровья (по заявлению обучающегося), а для инвалидов также в соответствии с индивидуальной программой реабилитации инвалида, могут предлагаться следующие варианты восприятия учебной информации с учетом их индивидуальных психофизических особенностей, в том числе с применением электронного обучения и дистанционных технологий:

6.1.1. *для лиц с нарушениями зрения:* в форме электронного документа; индивидуальные задания и консультации.

6.1.2. *для лиц с нарушениями слуха*: в форме электронного документа; индивидуальные задания и консультации.

6.1.3. *для лиц с нарушениями опорно-двигательного аппарата*: в форме электронного документа; индивидуальные задания и консультации.