Программа учебной дисциплины
«Корпусные методы исследований языковых процессов»

Утверждена
Академическим советом ООП
Протокол № 1 от «28» июня 2018 г.

Автор	Д.П. Попова
Число кредитов	3
Контактная работа (час.)	32
Самостоятельная работа (час.)	82
Курс	1
Формат изучения дисциплины	Full time (без использования онлайн-курса)

I. 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.

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Раздел 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 лекционных часа. Общий объем самостоятельной работы: 12 часов. Самостоятельная работа подразумевает подготовку к тесту.

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

Раздел 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 – 2 лекционных часа, 2 часа семинара. Общий объем самостоятельной работы: 10 часов. Самостоятельная работа подразумевает подготовку домашнего задания.

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

Раздел 3. **Статистические распределения**

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

Количество часов аудиторной работы – 2 лекционных часа, 2 часа семинара. Общий объем самостоятельной работы: 10 часов. Самостоятельная работа подразумевает подготовку домашнего задания.

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

Раздел 4. **Базовые статистические методы**

Тесты для определения вида распределения. Зависимые и независимые переменные. Линейная регрессия. Ковариантность. Статистическая значимость.

Количество часов аудиторной работы – 2 лекционных часа, 2 часа семинара. Общий объем самостоятельной работы: 10 часов. Самостоятельная работа подразумевает подготовку домашнего задания.

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

Раздел 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 – 2 лекционных часа, 2 часа семинара. Общий объем самостоятельной работы: 10 часов. Самостоятельная работа подразумевает подготовку домашнего задания.

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

Раздел 6. **Регрессионные методы**

Моделирование регрессии.

Количество часов аудиторной работы – 2 лекционных часа, 2 часа семинара. Общий объем самостоятельной работы: 10 часов. Самостоятельная работа подразумевает подготовку к тесту.

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

Раздел 7. **Модели со смешанным эффектом**

Использование моделей со смешанным эффектом.

Количество часов аудиторной работы – 2 лекционных часа, 2 часа семинара. Общий объем самостоятельной работы: 10 часов. Самостоятельная работа подразумевает подготовку к тесту.

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

Раздел 8. **Графическое представление данных**

Визуализация пройденных методов в R.

Количество часов аудиторной работы – 4 часа семинара. Общий объем самостоятельной работы: 10 часов. Самостоятельная работа подразумевает подготовку к выполнению финального проекта.

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

III. 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.

Оценки по всем формам текущего контроля и по итоговому контролю выставляются по 10-ти балльной шкале.

Тесты проводятся в рамках аудиторных занятий, домашние задания и комментарии по финальному проекту выкладываются on-line.

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. Финальный проект также предполагает устную презентацию проделанной каждым из студентов работы с последующей дискуссией между студентами.

Критерии оценки тестов и домашних заданий формулируются в тексте тестовых заданий и заданий на дом. Все тесты и домашние задания оцениваются по 10-ти балльной шкале. Оценка не округляется.

Критерии оценки финального проекта формулируются в отдельном файле за 2 недели до начала работы студентов над проектом. Проект оценивается по 10-ти балльной шкале. Оценка не округляется.

Результирующая оценка рассчитывается по формуле:

$$O_{результ} = 0,4 *\cdot O_{д/з} + 0,3 * O_{тест} + 0,3 * O_{проект},$$

где $O_{д/з}$ – оценка за домашние задания, $O_{тест}$ – оценка за тесты, $O_{проект}$ – оценка за финальный проект.

Результирующая оценка округляется в пользу студента. В диплом выставляется результирующая оценка по учебной дисциплине.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

1) Существуют две противоположные точки зрения на изучение иностранных языков детьми. Первая точка зрения: изучение иностранных языков детьми препятствует полноценному освоению родного языка. Такой точки зрения в своей языковой

политике традиционно придерживалась Япония. Вторая точка зрения: изучение иностранных языков детьми не препятствует полноценному освоению родного языка, возможно, даже способствует ему. Многие страны, например, Голландия, в своей языковой политике отражают вторую точку зрения.

Задание:

В файле hw3lr.csv находятся данные по 200 ученикам 10 лет школ одного и того же города. Переменная NofL (number of languages) характеризует детей по тому, сколькими языками они владеют и/или изучают в школе или в объёме не менее школьного. Например, если NofL=2, это может значить, что школьник билингв и владеет двумя языками с рождения, а может значить, что он владеет одним языком с рождения и изучает второй. Переменная Score отражает оценку за экзамен по родному языку. Переменная принимает значения от 0 до 100. Родной язык у всех школьников один и тот же.

- 1 балл: посчитайте в R среднее арифметическое и медиану для переменной NofL и для переменной Score. Приведите 4 числа и формулу, по которой Вы их находили в R.
- 1 балл: изобразите данные и линию регрессии между переменными NofL и Score на графике. Приведите формулы, которые Вы использовали при построении графика и сам график.
- 1 балл: приведите коэффициент корреляции между переменными NofL и Score и формулу, по которой Вы его посчитали.
- 1 балл: проинтерпретируйте результат. Существует ли линейная зависимость между двумя переменными? Если да, то какая: прямая (положительная) или обратная (отрицательная)?

1 балл: кратко (5-10 предложений) выскажите своё мнение по поводу языковой политики изучения детьми иностранных языков. Если Ваше мнение подтверждается проанализированными данными, используйте их в Вашей аргументации. Если Ваше мнение расходится с проанализированными в задании данными, объясните, почему их можно проигнорировать. Подсказка: наличие линейной зависимости между двумя переменными не гарантирует наличие причинно-следственной связи между ними.

2) На основании файла с данными опроса в школах проекта «Языки Москвы», ответьте на один из следующих вопросов.

- Есть ли корреляция между тем, сколько лет ребёнок живёт в России, и тем, насколько хорошо ребёнок говорит на русском/другом языке?
- Есть ли корреляция между тем, сколько лет ребёнок живёт в России, и тем, насколько хорошо ребёнок пишет на русском/другом языке?
- Есть ли корреляция между тем, сколько лет ребёнок живёт в России, и тем, насколько хорошо ребёнок читает на русском/другом языке?
- Есть ли корреляция между тем, сколько лет ребёнок живёт в России, и тем, насколько хорошо ребёнок понимает русский/другой язык?
- Есть ли корреляция между профессией матери/отца и тем, на каком языке ребёнок разговаривает с матерью/отцом?

- Есть ли корреляция между тем, разговаривает ли ребёнок на родном языке с бабушками/дедушками и разговаривает ли ребёнок на родном языке с братьями/сестрами?
- Есть ли корреляция между тем, разговаривает ли ребёнок на родном языке с бабушками/дедушками и разговаривает ли ребёнок на родном языке с друзьями?
- Есть ли корреляция между тем, различные ли национальности у родителей ребёнка, и тем, что ребёнок разговаривает только на русском языке?
- Есть ли корреляция между тем, одной ли национальности родители ребёнка, и тем, что ребёнок разговаривает на языке родителей?
- Есть ли корреляция между тем, на каком языке ему в детстве читали сказки, и тем, какой язык он считает родным?

V. РЕСУРСЫ

5.1 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.