Приложение 5

Федеральное государственное бюджетное образовательное учреждение высшего образования «Тамбовский государственный университет имени Г.Р. Державина»
Факультет истории, политологии и филологии
Кафедра зарубежной филологии и прикладной лингвистики

УТВЕРЖДАЮ:
Декан факультета



С. К. Лямин
«17» июня 2026 г.

РАБОЧАЯ ПРОГРАММА

по дисциплине ФТД.3 Статистические методы в лингвистических исследованиях

Направление подготовки/специальность: 45.03.01 - Филология

Профиль/направленность/специализация: Зарубежная филология

Уровень высшего образования: бакалавриат

Квалификация: Бакалавр

год набора: 2026

Тамбов, 2026

Автор программы:
Кандидат педагогических наук, доцент Кащеева Анна Владимировна

Рабочая программа составлена в соответствии с ФГОС ВО по направлению подготовки 45.03.01 - Филология (уровень бакалавриата) (приказ Министерства науки и высшего образования РФ от «12» августа 2020 г. № 986).

Рабочая программа принята на заседании Кафедры зарубежной филологии и прикладной лингвистики «16» июня 2026 г. Протокол № 11

Рассмотрена и одобрена на заседании Ученого совета Факультета истории, политологии и филологии, Протокол от «17» июня 2026 г. № 10.

СОДЕРЖАНИЕ

1. Цели и задачи дисциплины…… 4

2. Место дисциплины в структуре ОП бакалавриата……. 4

3. Объем и содержание дисциплины…… 5

4. Контроль знаний обучающихся и типовые оценочные средства…… 17

5. Методические указания для обучающихся по освоению дисциплины (модуля)…….. 23

6. Учебно-методическое и информационное обеспечение дисциплины…… 25

7. Материально-техническое обеспечение дисциплины, программное обеспечение, профессиональные базы данных и информационные справочные системы……. 25

1. Цели и задачи дисциплины

1.1 Цель дисциплины – формирование компетенций:

ОПК-7 Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности

1.2 Типы задач профессиональной деятельности, к которым готовятся обучающиеся в рамках освоения дисциплины:

- научно-исследовательский

1.3 Дисциплина ориентирована на подготовку обучающихся к профессиональной деятельности в сферах: 01 Образование и наука (в сферах: реализации основных образовательных программ основного общего, среднего общего образования, основных программ профессионального обучения, образовательных программ среднего профессионального образования, высшего образования, дополнительных профессиональных программ; научных исследований), Сфера перевода (устный и письменный (в том числе художественный) перевод), Сфера устной и письменной коммуникации

1.4 В результате освоения дисциплины у обучающихся должны быть сформированы:

Обобщенные трудовые функции / трудовые функции / трудовые или профессиональные действия (при наличии профстандарта)	Код и наименование компетенции ФГОС ВО, необходимой для формирования трудового или профессионального действия	Индикаторы достижения компетенций
	ОПК-7 Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности	Осуществляет поиск, систематизацию, критический анализ информации по использованию статистических методов в лингвистических исследованиях для решения поставленных задач

1.5 Согласование междисциплинарных связей дисциплин, обеспечивающих освоение компетенций:

ОПК-7 Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности

№ п/п	Наименование дисциплин, определяющих междисциплинарные связи	Форма обучения		
		Очная (семестр)		
		5	6	7
1	Информационные технологии в профессиональной деятельности	+		
2	Педагогическая практика			+
3	Переводческая практика (получение первичных навыков перевода)		+	

2. Место дисциплины в структуре ОП бакалавриата:

Дисциплина «Статистические методы в лингвистических исследованиях» изучается в 5 семестре.

3.Объем и содержание дисциплины

3.1.Объем дисциплины: 2 з.е.

Очная: 2 з.е.

Вид учебной работы	Очная (всего часов)
Общая трудоёмкость дисциплины	**72**
Контактная работа	32
Лекции (Лекции)	16
Практические (Практ. раб.)	16
Самостоятельная работа (СР)	40
Зачет	-

3.2.Содержание курса:

№ темы	Название раздела/темы	Вид учебной работы, час.			Формы текущего контроля
		Лекции	Практ. раб.	СР	
		О	О	О	
5 семестр					
1	Квантитативная лингвистика. Определение количества языковых явлений в тексте, среднего арифметического и частотности.	2	2	5	Опрос
2	Дисперсионный анализ в лингвистике. Репрезентация массива и диапазона данных. Нормальное распределение. Среднее квадратическое отклонение	2	2	5	Опрос
3	Регрессионный анализ в лингвистике.	2	2	5	Опрос; Контрольная работа
4	Кросскорреляционный анализ в лингвистике	2	2	6	Опрос
5	Факторный анализ в лингвистике	2	2	7	Опрос

6	Кластерный анализ в лингвистике	4	4	7	Опрос; Контрольная работа
7	Вероятностный анализ языковых явлений в тексте	2	2	5	Опрос

Тема 1. Квантитативная лингвистика. Определение количества языковых явлений в тексте, среднего арифметического и частотности. (ОПК-7)

Лекция.

Самым простым методом статистического анализа являются подсчет общего количества языковых явлений в выборке; среднего арифметического; частотности. Общее количество можно определить путем простого подсчета случаев употребления требуемого языкового явления в выборке. Среднее арифметическое – это показатель центральной тенденции, средняя частота использования языкового явления в выборке. Это понятие относится также к регрессионному анализу и распределению вероятностей, а поэтому является универсальным для разных квантитативных метод. Частотность – это процент использования искомого языкового явления по отношению ко всем языковым явлениям в выборке. Например, количество словоупотреблений на странице по отношению к общему количеству слов на странице. Частотность применима ко всем уровням языка: фонетическому, лексическому, грамматическому, орфографическому. Частотность в языковой выборке или языковой совокупности свидетельствует об употребительности языкового явления, его характерности для определенного типа текста или текста автора. Понятие частотности также может быть использовано в сравнительном анализе разных текстов. Например, в художественном отрывке (Приложение 1) можно подсчитать некоторые явления, характеризующие повествовательный текст. Например, если определить общее количество употреблений ключевых слов заголовка incident (0 случаев) и customs (7 случаев), то можно сделать вывод о том, что главная идея текста выражена иными лексическими способами, автор избегает прямого называния, или лексика заголовка использована в переносном смысле. Количество употреблений местоимения I (41) свидетельствует о повествовании от первого лица, а количество употреблений времени Past Simple (65) по сравнению с Past Progressive (6) – о том, что в повествовании факты преобладают над описаниями и разворачивающимися событиями. Если сравнить количество употреблений названных прошедших времен в тексте по сравнению с другими временами, то очевидно преимущество использования Past Simple в повествовании (рис. 1). В нижеследующих примерах представлены оригинальные тексты из газеты The Guardian (Приложение 2). Тексты относятся к деловому официальному стилю, для которого характерно использование официальной лексики, предложений в косвенной речи и страдательного залога глагола. Мы установили, к примеру, размер выборки каждые 200 слов, объединив подсчет в двух текстах, относящихся к единому стилю. По нашим подсчетам, среднее арифметическое употреблений слова “government” на каждые двести слов (3, 0, 2, 0, 1, 1, 1) – 1,14, слова budget – 0,45, слова bureaucracy – 0,14 на каждые 200 слов. Таким образом, для каждого из текстов официальная лексика общего значения менее характерна, чем тематическая лексика того же регистра (adopt и производные, assess и производные, rating, debt). Среднее арифметическое предложений в прямой речи составило 2 случая на каждые 200 слов, что довольно часто для текста официального стиля, вопреки ожиданиям. Среднее арифметическое предложений в страдательном залоге – 1,7 случаев на 200 слов, из чего можно сделать вывод о преобладании активного залога. На примере художественного текста Skin (Приложение 3), в котором мы проследили частотность некоторых явлений в выборке каждой страницы, одна из тем – талант художника – лексически выражена в тексте следующим образом. Частотность использования лексики по теме «Изобразительное искусство» на каждой из 7 страниц текста следующая: стр. 1 – 9 (3,8% от 237 слов), стр. 2 – 7 (1,4% от 484 слов), стр. 3 – 14 (3% от 457 слов), стр. 4 – 24 (10,4% от 230 слов), стр. 5 – 13 (2,7% от 485), стр. 6 – 6 (1,4% от 419 слов), стр. 7 – 5 (1,1% от 457 слов). Таким образом, можно говорить о достаточной лексической репрезентативности темы «изобразительное искусство» в тексте. Частотность служебных слов (частиц, предлогов, союзов, артиклей) на каждой странице текста стр.1 – 46 (27%), стр. 2 – 116 (23% от 502 слов), стр. 3 – 89 (20,3% от 437слов), стр. 4 – 103 (22%) от 462 слов. По сравнению с вышеупомянутыми текстами официального стиля частотность составляет на каждую страницу текста 24,2% (77 на 317 слов), 28,3% (88 на 311 слов), 25,6% (87 на 340 слов). Сравнение показало, что нет принципиальной разницы между частотностью служебных слов в художественном и публицистическом тексте.

Практическое занятие.

Знакомство с основными понятиями квантитативной лингвистики, определение понятий. Овладение статистическими приемами определения количества языковых явлений в тексте, среднего арифметического и частотности

Задания для самостоятельной работы.

Выполнение статистических расчетов на основе текстов разных жанров на английском языке.

Тема 2. Дисперсионный анализ в лингвистике. Репрезентация массива и диапазона данных. Нормальное распределение. Среднее квадратическое отклонение (ОПК-7)

Лекция.

Дисперсионный анализ (Analysis of Variance: ANOVA) – это метод в математической статистике, используемый для выявления влияния разных факторов на исследуемую переменную. Иначе говоря – это изучение вариации (рассеивания, дисперсии) какого-либо признака или переменной. Дисперсией называют меру отклонения от среднего значения. В современных негуманитарных исследованиях термин используется для сопоставления измерений одного типа, выполненных при разных условиях. Дисперсионный анализ проверяет значимость различия между средними значениями с помощью сравнения дисперсий. Дисперсию одного измеряемого признака разлагают на независимые слагаемые, каждое из которых характеризует влияние того или иного фактора или их взаимодействия. Сравнение представляет собой анализ отклонения каждого из значений от усредненной оценки, которое и является вариацией признака. Следовательно, данные с низкой дисперсией мало отличаются друг от друга, а данные с высокой дисперсией имеют значительные отличия. Дисперсия σ2 – мера вариации, определяемая как средняя из отклонений признака от его средней величины, возведенная в квадрат. Возведение в квадрат необходимо потому, что отклонение от средней величины может быть отрицательным. Для того чтобы мера отклонения соответствовала значению признака, в математике используют формулу среднего квадратического отклонения. Это наиболее совершенная характеристика вариации, называемая иначе стандартом или стандартным отклонением. В квантитативной лингвистике отклонение может быть абсолютным (отдельна величина признака в отдельной выборке) или средним (среднее арифметическое всех отклонений во всех выборках). В математике среднее квадратическое отклонение σ равно квадратному корню из среднего квадрата отклонений отдельных значений признака от средней арифметической. Оно используется для определения значений У (ординат) кривой нормального распределения при оценке границ вариации признака (в конечном ряду). Среднее квадратическое отклонение вычисляется по формуле: Для лингвистических исследований использование величины среднего квадратического отклонения означает измерение меры дисперсии (рассеивания) языкового явления или среднего отклонения изменения количества какого-либо языкового явления от среднего арифметического в разных текстах или постранично. Дисперсионный анализ лингвистических явлений может затрагивать структуру предложений и употребление частей речи в зависимости от жанра текста, а также зависимости одних лексических явлений от других. Дисперсионный анализ лингвистических данных можно проводить с помощью программы Excel, для чего задаются массивы и диапазоны (способы организации данных). Массивом называют упорядоченную совокупность элементов одного типа (переменной). Каждый элемент массива имеет индексы, определяющие порядок элементов. Число индексов характеризует размерность массива. Каждый индекс изменяется в некотором диапазоне [a,b]. Иначе говоря, диапазон - это две и более ячеек, а массив представляет собой строки и столбцы, в которые вносятся средние показатели какого-либо признака. Например, чтобы вычислить дисперсию какого-либо лингвистического явления с помощью статистических формул программы Excel, надо в каждую ячейку одной строки внести количество лингвистических явлений в каждой новой выборке (тексте), затем в новой ячейке ряда нажать знак «=» и выбрать из формул ДИСП (в выборке). Пример 1. Вычисление дисперсии. Для вычисления дисперсии мы сравнили влияние вида печатного издания, в котором публикуется текст, на количество употреблений лингвистических форм того или иного вида. Мы отобрали по 4 неадаптированных текста из английских газет: серьезной газеты “The Guardian” и таблоида “The Daily Mirror” (Приложения 5,6). Тексты статей были взяты из одинаковых рубрик: политика, бизнес, отдых, окружающая среда. В каждой из выборок было посчитано количество лингвистических явлений, характерных или нехарактерных для официального стиля изложения в газетном тексте. Мы поставили задачу проследить, насколько статус газеты влияет на использование этих лингвистических явлений в текстах. Таблица словоупотреблений выглядит следующим образом: Количество словоупотреблений The Guardian The Daily Mirror Дисперсия Сокращенные формы 3 32 242 Прямая речь и цитаты 14 21 24,5 Составные слова 21 21 0 Цифры 60 60 0 Учитывая полученные данные, мы пришли к выводу о том, что статус газеты практически не влияет на количество использований цифр и составных слов типа pre-condition, jungle-covered, year-long, wholly-owned, cash-strapped, ill-treatment, которые в некоторых случаях заменяют объясняющие обороты и экономят количество слов. Различия в количестве предложений в прямой речи также невелики, что свидетельствует о типичности таких предложений для газетного стиля изложения вообще. Самый большой показатель дисперсии относится к использованию сокращенных форм, нетипичных для официального стиля и приемлемого для текстов таблоидов.

Практическое занятие.

Дисперсионный анализ в лингвистике. Репрезентация массива и диапазона данных. Нормальное распределение. Среднее квадратическое отклонение.

Знакомство с сущностью дисперсионного анализа в лингвистике. Овладение способами репрезентации массива и диапазона языковых данных в диаграмме рассеяния (scattergram), определение нормального распределения данных и среднего квадратического отклонения

Задания для самостоятельной работы.

Выполнеие статистических расчетов на основе текстов разных жанров на английском языке.

Тема 3. Регрессионный анализ в лингвистике. (ОПК-7)

Лекция.

Исследования рассеяния какого-либо языкового признака тесно связаны с математической функцией регрессии. Регрессионный анализ заключается в установлении зависимости между постоянной (критериальной) переменной и одной или несколькими независимыми. Регрессия показывает, на какую величину может измениться значение одного признака, если другой меняется на определенную единицу измерения. Для статистики регрессионный анализ важен, потому что он позволяет сделать выводы о предполагаемой величине одного признака по значениям другого. Для измерений регрессии в математике обычно используют средние величины двух переменных. Если следовать точному определению процедуры регрессионного анализа в статистике, то оно включает: - анализ диаграммы разброса, - расчет регрессионных коэффициентов, - расчет коэффициента корреляции, - проверку остатков на случайность. Величину регрессии можно вычислить с помощью Excel и STATISTICA с использованием формул в меню «Функции», подменю «Статистические». Тем не менее, для лингвистического анализа цифровой показатель регрессии не столь информативен, как анализ диаграммы рассеяния и центральной тенденции. Например, есть диаграммы рассеяния признаков, в которых невозможно определить центральную тенденцию. Графически одну из них их можно представить в следующем виде (рис.10): Рисунок 10. Для анализа лингвистических явлений также важен вид используемой регрессии. В математической статистике различают линейную и нелинейную регрессию. Нелинейная регрессия анализирует влияние нескольких факторов на исследуемый признак, линейная – одного. В лингвистических исследованиях наиболее применима линейная регрессия, которая позволяет выявить взаимосвязь характеристики лингвистического объекта (У) и величины, влияющей на её изменения (Х). Такая регрессия всегда графически выражена прямой, которая может быть направлена вверх по отношению к оси Х (положительная корреляция) или вниз (отрицательная корреляция). Взаимосвязь функции регрессии и свойств рассеяния в диаграмме выражена прямой, которая называется центральной тенденцией и характеризуется отклонением от оси Х в зависимости от коэффициента. Чем больше коэффициент отклонения одной переменной (Х) от другой (У) больше, то есть, величина, на которую изменяется Х при изменении У на установленную единицу измерения больше, тем больше угол прямой. Это означает, что тем больше связь между этими двумя признаками. В математической статистике переменную Х иначе называют откликом, а переменную У – фактором или регрессором. На рисунке 11 показана центральная тенденция с отрицательной корреляцией признаков, анализ которых был проведен по данным языкового корпуса. График построен в программе STATISTICA. График корреляции двух признаков имеет отрицательную тенденцию. С увеличением числа компонентов фразового глагола (Х) число словоупотреблений этого глагола (У) сокращается (рис. 11). Для корреляционно-регрессионного анализа языковых явлений мы использовали их числовые и качественные проявления. Предположим, что можно линейно выразить взаимосвязь следующих лингвистических явлений. Например, как изменится количество употреблений имен существительных в выборке, если число определенного и неопределенного артиклей возрастет на какую-то величину. Таким образом, мы можем проследить зависимость использования артиклей от категории используемых имен существительных (существительные в единственном числе, выражающие общие и вещественные понятия и существительные во множественном числе).

Практическое занятие.

Регрессионный анализ в лингвистике. Центральная тенденция.

Ознакомление с сущностью и процедурой регрессионного анализа в лингвистике. Репрезентация данных на осях ординат. Определение центральной тенденции

Задания для самостоятельной работы.

Выполнеие статистических расчетов на основе текстов разных жанров на английском языке.

Тема 4. Кросскорреляционный анализ в лингвистике (ОПК-7)

Лекция.

Используя термин «корреляционно-регрессионный анализ», уточним особенности понятий «регрессия» и «корреляция». Регрессия показывает, имеется ли связь между двумя явлениями, а корреляция – насколько она сильна. Кроме того, можно проследить корреляцию нескольких признаков в лингвистических исследованиях, что будет составлять кросс-корреляционный анализ. Для удобства представления данных с целью подсчета коэффициента корреляции используют матрицу, в столбцы которой вносят изменения значений исследуемых признаков. Далее для подсчета коэффициента корреляции используют статистические формулы Excel и анализируют данные в диапазоне от -1 до 1. Если значение коэффициента корреляции ближе к единице, то два признака имеют сильную степень взаимосвязи. В лингвистических исследованиях этот показатель свидетельствует о тенденции взаимного влияния лингвистических признаков. Как правило, мы проводим исследования выборок, а не всего массива или генеральной совокупности (например, всех текстов автора), что вызывает неточность подсчетов. Очевидно, что при отрицательном коэффициенте корреляции r увеличение словоупотреблений одной переменной вызывает уменьшение словоупотреблений другой, а положительный коэффициент свидетельствует об обратном. Следовательно, правильнее говорить о тенденциях функционального отталкивания или функционального притяжения явлений.

Практическое занятие.

Коэффициент корреляции и кросскорреляционный анализ в лингвистике.

Выявление возможностей корреляционного и кросскорреляционного анализа в изучении явлений языка и текста.

Задания для самостоятельной работы.

Выполнеие статистических расчетов на основе текстов разных жанров на английском языке.

Тема 5. Факторный анализ в лингвистике (ОПК-7)

Лекция.

Следующим статистическим методом, который может быть использован в лингвистике, является метод факторного анализа. С точки зрения статистики он является многомерным исследовательским методом, поскольку процедура факторного анализа достаточно сложна и проходит несколько этапов. Факторный анализ использует уже известные нам понятия распределения, рассеивания, матричной организации данных и корреляции. В статистике под факторным анализом понимают совокупность методов, которые позволяют обобщать характеристики изучаемых явлений и процессов в достаточно больших выборках. Они помогают выявить скрытые (латентные) причины, или скрытые факторы, по которым разные признаки коррелируют между собой. Группы обобщенных таким образом признаков и называют факторами. Обобщение признаков объектов называют R-техниками или R-анализом, а обобщение самих объектов – Q-анализом. В результате R-анализа получают несколько групп (комбинаций) признаков, а в результате Q-анализа – комбинации объектов. Выделение групп признаков связано с их дисперсией, потому что факторы обычно выделяются последовательно: первый фактор объясняет наибольшую долю дисперсии признаков (то есть, включает больший процент от общего количества признаков), второй – меньшую долю дисперсии и т.д. Статистические исследования имеют четкую процедуру, которая устанавливает, сколько надо выделить факторов. В лингвистических исследованиях эта процедура может принимать упрощенный вид, поскольку классификация явлений по общему признаку более важна, чем ограничение количества признаков. Известно, что факторный анализ преследует две цели: группирование или классификация переменных в большие группы по степени близости и сокращение переменных (анализ главных компонент) из множества значений. Однако обе цели позволяют четче представить более простую факторную структуру изучаемого объекта или явления из всего многообразия признаков. В факторном анализе корреляция отдельного признака и фактора представлена двумя коэффициентами. Факторная нагрузка – это мера влияния фактора на признак, изменяющаяся в диапазоне от -1 до 1. Чем ближе коэффициент к единице, тем больше фактор влияет на признак, а чем ближе к 0, тем меньше это влияние. В лингвистическом анализе факторную нагрузку можно интерпретировать как степень проявления лингвистических явлений в том или ином факторе (например, принадлежность окончаний (суффиксов) к факторам видовременных форм глагола; к фактору окончаний имен существительных; к притяжательному падежу существительных; к окончаниям прилагательных). Другой коэффициент называется факторным весом. Если сравнивать проявление одного и того же фактора у разных объектов (например, типов текстов), то мера (коэффициент) проявления и есть факторный вес. Чем ближе факторный вес к 0 или -1, тем меньше степень проявления фактора. Чм ближе факторный вес к 1, тем больше степень проявления фактора. Если составить матрицу факторных весов, то строки будут соответствовать количеству объектов, а столбцы – количеству факторов. Такая матрица позволяет судить о распределении объектов по каждому фактору (анализ строк по горизонтали) и распределении факторов по мере их проявления в разных объектах или группах объектов (по вертикали в столбцах). Анализ матрицы результатов с целью выделения факторов проходит несколько этапов. На первом этапе выделяются исходные (неповернутые) факторы, то есть, формируется матрица первичных результатов, которая показывает взаимосвязь факторов и отдельных переменных. Тем не менее, она редко позволяет однозначно интерпретировать результаты факторного анализа, потому что одна и та же переменная (явление) можно отнести к нескольким факторам, их позиция погранична. Для того чтобы получить итоговое количество факторов, с помощью которых можно объяснить связи между различными данными, используют прием вращения факторов, иначе – перегруппировки, при которой их количество меняется. Говоря языком статистики, сложную матрицу упрощают, в ней соотношение факторных нагрузок становится более понятным и однозначным. При вращении получают несколько решений факторного анализа из одного набора данных. Известно несколько методов вращения, но наиболее распространены методы «варимакс» и «кватримакс». Оба метода вращения выбирают факторные нагрузки с большим диапазоном значений, при этом увеличивая большие значения и уменьшая маленькие. Метод «варимакс» используется чаще, потому что он позволяет увидеть разброс нагрузок для каждого фактора в отдельности, а метод «кватримакс» - для всех факторов вместе. Вращение можно использовать неоднократно, задавая в программе необходимое количество факторов, которое наилучшим образом позволит объяснить изучаемое явление.

Практическое занятие.

Изучение исследовательских возможностей факторного анализа для изучения явлений морфологии, синтаксиса и семантики языка.

Задания для самостоятельной работы.

Выполнеие статистических расчетов на основе текстов разных жанров на английском языке.

Тема 6. Кластерный анализ в лингвистике (ОПК-7)

Лекция.

Кластерный анализ является видом многомерного статистического анализа, который заключается в разбиении объектов на кластеры (укрупненные и упрощенные группы), каждый из которых состоит из схожих объектов, а сами кластеры различаются между собой. Кластерный анализ используют в том случае, если существует много объектов с разными признаками, которые надо сгруппировать по признакам. Допустим, существует некоторое количество предложений, которые характеризуются подлежащим, выраженным существительным или инфинитивом, наличием определения, наличием сложных причастных или инфинитивных конструкций, наличием атрибутивных придаточных предложений, общим количеством слов, количеством знаменательных слов. В результате кластерного анализа мы получим группы предложений, которые ближе всего по всем указанным признакам. Мы можем ограничиться двумя или тремя самыми значимыми признаками и группировать кластеры слов, текстов, предложений по ним. Очевидно, что кластерный анализ как метод классификации может затрагивать все языковые уровни. Есть несколько способов кластерного анализа в статистике. Первый иерархический или древовидный (joining or tree clustering) способ состоит в том, чтобы последовательно выделять кластеры признаков по степени близости, начиная с самых близких. Так пошагово группируются каждые две ближайшие группы объектов по ближайшим признакам до получения желаемого результата. Другой метод К-средних (K-means clustering) заключается в построении путем случайной группировки по признакам заранее известного количества кластеров, которые максимально отличаются друг от друга. С помощью этого метода на первом этапе определяются К кластеров-эталонов, далее каждый объект присоединяется к ближайшему эталону. С образованием нового кластера эталон пересчитывается, и объекты снова присоединяются к ближайшим кластерам до стабилизации процесса кластеризации. Эти процедуры доступны в программе STATISTICA, дополнительной опцией является метод двухвходового объединения (two-way joining). Он позволяет группировать и признаки, и кейсы одновременно. Тем не менее, самым известным является древовидный метод. Программа дает графическое изображение кластеров в виде кубиков и показывает меру их близости в системе координат. В случае, если кластеры группируются по двум или трем признакам, возможно также получить диаграмму рассеяния кластеров. Для выполнения кластеризации объектов в программе нам потребуется создать матрицу из переменных, признаков (столбцы) и наблюдений (строки). С помощью функции Cluster analysis следует выбрать расстояние между объектами кластеров/amalgamation или метрику (обычно – евклидова метрика/squared Euclidian distances), а затем проанализировать появившуюся дендрограмму. Приведем примеры древовидного кластерного анализа, для которого мы взяли корпус текстов, использовавшихся в разделе «Дисперсионный анализ». Условно разделив тексты на группы художественных и нехудожественных произведений, попытаемся проследить, насколько близки корпусы художественных текстов друг к другу по выражению концепта «любовь» в словах love, affection, desire, devotion, которые и будут являться параметрами измерения. Другим примером кластеризации будет наша попытка сгруппировать нехудожественные тексты по видам модальности, представленной глаголами should, might, could, can’t, must, may, dare, ought to. Выбрав 7 текстов и соответствующие понятию «любовь» слова love, affection, desire, devotion, составляем матрицу количества словоупотреблений в каждом из текстов (рис. 24). Love affection desire Devotion Pride and Prejudice 60 58 9 2 Beowolf 68 14 9 2 Jane Eyre 219 41 34 8 Great Expectations 219 41 34 8 Ulysses 390 46 53 9 The Sonnets (W. Shakespeare) 585 46 64 9 The Lord of the Rings. 611 47 96 9 Выбрав последовательно функции «многомерный анализ», «анализ кластера», отмечаем переменные для анализа и получаем следующие данные кластерного анализа в таблице (рис. 25). Number of variables: 4 Number of cases: 7 Joining of variables Missing data were casewise deleted Amalgamation (joining) rule: Single Linkage Distance metric is: Euclidean distances (non-standardized) Далее нажмем кнопку для получения графика (Vertical Icicle Plot) и получим дендрограмму кластеров (рис. 26). Дендрограмма кластеров показывает, что переменные 2 и 3 (desire, affection) образуют единый кластер, и расстояние между ними по вертикали невелико, а также расстояние между этим кластером и переменными devotion и love почти одинаково. Далее, на втором по степени близости шаге этот кластер объединен с кластером devotion (расстояние по вертикали также мало). На третьем шаге первые два кластера объединяются с достаточно удаленным от них понятием love, который относит все понятия к единому кластеру «любовь». Таким образом, чтобы «расшифровать» кластерную диаграмму, мы начинаем анализ снизу, анализируя явления, которые соединены в единый «кубик» - кластер. Далее, идя вверх, смотрим, с какими явлениями или кластерами объединяется этот кластер. Иными словами, осуществляется анализ того, в какой очередности языковые явления объединяются по степени близости. Данные анализа нашего примера говорят о том, что семантически слова desire, affection наиболее близки семантически. Близко по

Практическое занятие.

Изучение исследовательских возможностей кластерного анализа для изучения явлений морфологии, синтаксиса и семантики языка.

Задания для самостоятельной работы.

Выполнеие статистических расчетов на основе текстов разных жанров на английском языке.

Тема 7. Вероятностный анализ языковых явлений в тексте (ОПК-7)

Лекция.

Вероятность в широком смысле – это мера появления события (А), если осуществляется комплекс условий. Если в ходе испытания нельзя предсказать результат при повторяющихся условиях, то такой результат называют случайным событием. Вероятность есть объективная характеристика того, что случайное событие А может произойти. Если таких событий много в серии экспериментов, то частота его появления и составляет некоторое постоянное число вероятности Р. Как отмечает Б.Н. Головин [4], вероятность есть объективное свойство развивающегося языка, которая неразрывно связана с частотностью явлений. Статистическая обработка частотности языковых явлений позволяет вычислять вероятность их появления в речи. В свою очередь, знание вероятностей позволяет установить частотность речевых структур. Вероятность показывает возможность возникновения явления в пределах от 0 до 1. Близость показателя к единице свидетельствует о большой вероятности или достоверности события. Анализ вероятности (Probabilistic Analysis) в лингвистике подчиняется не только статистическим законам распределения. По мнению Р.Г. Пиотровского[9], кроме классического и статистического способа определения вероятности в языкознании возможно ещё и субъективное определение, которое подразумевает попытку субъекта оценить достоверность вхождения лингвистических объектов в множества (событие может принадлежать или не принадлежать множеству). Субъективное определение вероятности позволяет измерять семантическую информацию. Под классическим определением вероятности понимается формула P=F:N, где F – количество случаев «успеха» (количество проявлений искомого лингвистического явления), а N – общее количество лингвистических явлений. Примером может быть отношение двусложных слов (F) к общему числу слов на странице или в заданном диапазоне текста (N). Статистическое определение вероятности применимо в тех случаях, когда исследуется не вся совокупность, а выборка, в результате чего исследователь получает относительные частоты случайных событий. Если сложить частоты в нескольких выборках и поделить полученное число на количество выборок, то мы получим относительную частоту явления. Поскольку язык относится к сложным гуманитарным системам, то определение вероятности приобретает особую значимость при проверке гипотез. Экспериментальные данные, полученные при проверке гипотезы, дают возможность построить новые вероятности, которые Р.Г. Пиотровский назвал апостериорными в отличие от начальных, априорных [8]. Самым простым, классическим способом определения вероятности лингвистического события является вычисление по формуле: частотность явления в выборке деленная на количество выборок. Например, частотность явления на каждой из страниц текста, деленная на число страниц. Таким образом, получаем среднюю величину в диапазоне от 0 до 1. В свою очередь, частотность определяют как количество словоупотреблений по отношению к общему количеству слов на странице. Более сложные вычисления вероятности можно осуществить с помощью программ Excel и STATISTICA, которые позволяют рассчитать два вида распределения вероятностей: дискретные и непрерывные. Дискретные распределения моделируют наступление отдельных событий. Приведем примеры распределения вероятностей дискретных величин. 1. Классическая вероятность есть собственно отношение числа успехов случайного события к общему числу исходов, иначе её называют статистической вероятностью. Взяв, к примеру, вероятность употребления артиклей в английском языке, предположим, что из 1000 артиклей в тексте артикль the был употреблен 347 раз, следовательно, вероятность его появления в этом тексте составила 0,347. 2. Биномиальное статистическое распределение или распределение Бернулли (распределение двух исходных признаков). Косвенно этот тип вероятности связан с нормальным распределением, потому что условием биномиального распределения является постоянная вероятность наступления события в каждом из испытаний. Предположим, надо посчитать число вероятных употреблений артикля the (по сравнению с артиклем а) из общего числа артиклей в выборке. 3. Полиномиальное или мультиномиальное – это распределение взаимоисключающих величин или совместное распределение групповых частот. Примером может служить появление определенного числа the и a из всего количества артиклей в разных выборках. В отличие от биномиального, этот вид распределения представляет сразу несколько величин, а не одну. 4. Геометрическое и гипергеометрическое распределение являются частным случаем биноминального распределения. Гипергеометрическое распределение справедливо для больших совокупностей, где надо вычислить вероятность, исходя из вероятности в отдельно взятой выборке. С помощью этой функции мы можем вычистить, например, вероятность того, что артикль the встретится в совокупности из 100 случаев употребления артиклей, если в первой малой выборке из 6 случаев он встретился 3 раза. Геометрическое распределение помогает установить количество испытаний, которое необходимо провести до получения первичного «успеха», или первого употребления исследуемого явления. Непрерывные распределения моделируют вероятность непрерывных событий, таких как

Практическое занятие.

Использование средств вероятностного анализа наступление лингвистических событий в текстовом массиве (фонетика, морфология, синтаксис, семантика).

Задания для самостоятельной работы.

Выполнеие статистических расчетов на основе текстов разных жанров на английском языке.

4. Контроль знаний обучающихся и типовые оценочные средства

4.1. Распределение баллов:

5 семестр

- посещаемость – 10 баллов
- текущий контроль – 70 баллов
- контрольные срезы – 2 среза по 10 баллов каждый
- премиальные баллы – 20 баллов

Распределение баллов по заданиям:

№ темы	Название темы / вид учебной работы	Формы текущего контроля / срезы	Max. кол-во баллов	Методика проведения занятия и оценки
1.	Квантитативная лингвистика. Определение количества языковых явлений в тексте, среднего арифметического и частотности.	Опрос	10	10 баллов - студент владеет темой и методикой статистического лингвистического анализа. 7 баллов - студент владеет темой и методикой статистического лингвистического анализа в достаточной степени, допускает незначительные ошибки в ответе. 5 баллов - студент владеет основами темы, затрудняется в статистическом анализе. 2 балла - студент фрагментарно отвечает на вопросы по теме
2.	Дисперсионный анализ в лингвистике. Репрезентация массива и диапазона данных. Нормальное распределение. Среднее квадратическое отклонение	Опрос	10	10 баллов - студент владеет темой и методикой статистического лингвистического анализа. 7 баллов - студент владеет темой и методикой статистического лингвистического анализа в достаточной степени, допускает незначительные ошибки в ответе. 5 баллов - студент владеет основами темы, затрудняется в статистическом анализе. 2 балла - студент фрагментарно отвечает на вопросы по теме
3.	Регрессионный анализ в лингвистике.	Опрос	10	10 баллов - студент владеет темой и методикой статистического лингвистического анализа. 7 баллов - студент владеет темой и методикой статистического лингвистического анализа в достаточной степени, допускает незначительные ошибки в ответе. 5 баллов - студент владеет основами темы, затрудняется в статистическом анализе. 2 балла - студент фрагментарно отвечает на вопросы по теме
		Контрольная работа(контрольный срез)	10	Контроль владения методикой статистического анализа лингвистических явлений проводится в виде выполнения 5 практических заданий, за правильный ответ начисляется 2 балла. За неточный ответ -1 балл

4.	Кросскорреляционный анализ в лингвистике	Опрос	10	10 баллов - студент владеет темой и методикой статистического лингвистического анализа. 7 баллов - студент владеет темой и методикой статистического лингвистического анализа в достаточной степени, допускает незначительные ошибки в ответе. 5 баллов - студент владеет основами темы, затрудняется в статистическом анализе. 2 балла - студент фрагментарно отвечает на вопросы по теме
5.	Факторный анализ в лингвистике	Опрос	10	10 баллов - студент владеет темой и методикой статистического лингвистического анализа. 7 баллов - студент владеет темой и методикой статистического лингвистического анализа в достаточной степени, допускает незначительные ошибки в ответе. 5 баллов - студент владеет основами темы, затрудняется в статистическом анализе. 2 балла - студент фрагментарно отвечает на вопросы по теме
6.	Кластерный анализ в лингвистике	Опрос	10	10 баллов - студент владеет темой и методикой статистического лингвистического анализа. 7 баллов - студент владеет темой и методикой статистического лингвистического анализа в достаточной степени, допускает незначительные ошибки в ответе. 5 баллов - студент владеет основами темы, затрудняется в статистическом анализе. 2 балла - студент фрагментарно отвечает на вопросы по теме
		Контрольная работа(контрольный срез)	10	Контроль владения методикой статистического анализа лингвистических явлений проводится в виде выполнения 5 практических заданий, за правильный ответ начисляется 2 балла. За неточный ответ -1 балл
7.	Вероятностный анализ языковых явлений в тексте	Опрос	10	10 баллов - студент владеет темой и методикой статистического лингвистического анализа. 7 баллов - студент владеет темой и методикой статистического лингвистического анализа в достаточной степени, допускает незначительные ошибки в ответе. 5 баллов - студент владеет основами темы, затрудняется в статистическом анализе. 2 балла - студент фрагментарно отвечает на вопросы по теме
8.	Посещаемость		10	Начисляются, если студент посетил не менее 80% занятий
9.	Премиальные баллы		20	Начисляются за активную работу на занятиях
10.	Индивидуальные задания, с помощью которых можно набрать дополнительные баллы		100	НАчисляются за выполнение заданий курса
11.	Итого за семестр		100	

Итоговая оценка по зачету выставляется в 100-балльной шкале и в традиционной четырехбалльной шкале. Перевод 100-балльной рейтинговой оценки по дисциплине в традиционную четырехбалльную осуществляется следующим образом:

100-балльная система	Традиционная система
50 - 100 баллов	Зачтено
0 - 49 баллов	Не зачтено

4.2 Типовые оценочные средства текущего контроля

Контрольная работа

Тема 3. Регрессионный анализ в лингвистике.

1. Определение абсолютного количества явлений языка на странице текста (части речи, члены предложения в английском языке, в сравнении с русским языком).
2. Определение среднего арифметического встречаемости явлений языка на странице текста (части речи, члены предложения в английском языке в сравнении с русским языком).
3. Определение частотности встречаемости явлений языка на странице (индекс и % частей речи и членов предложения, сокращений, терминов, синтаксических конструкций). Дисперсионный анализ массива языковых данных на странице текста (части речи и члены предложения).
4. Определения диапазона языковых данных на странице (части речи и члены предложения). Графическая репрезентация диапазона.
5. Регрессионный анализ языковых данных (неологизмы, заимствования, эллиптические конструкции, полные предложения).

Тема 6. Кластерный анализ в лингвистике

1. Определение центральной тенденции в регрессионном анализе (сокращения, термины, ключевые слова).
2. Вычисление коэффициента корреляции языковых явлений (предлоги, союзы, артикли).
3. Кросскорреляционный анализ (части речи, члены предложения, грамматические явления).

Проверка массива данных на «нормальное распределение» (распределение на одной странице слов с количеством букв от 1 до 10).

Опрос

Тема 1. Квантитативная лингвистика. Определение количества языковых явлений в тексте, среднего арифметического и частотности.

Задание 1. Изучите примеры личного письма (Приложение 4). Одно из них опубликовано в 1872 г. в сборнике художественных произведений («Letters from India» by Emily Eden), другое – образец современного письма другу, опубликованного на Интернет-сайте (http://www.publishyourarticles.org). Третий текст представляет собой литературное письмо Е.Блаватской из Индии [2].

Определите количество языковых явлений, характерных для стиля личного письма, в каждом из образцов Приложения 4: идиоматические выражения, сокращенные формы, эллиптические конструкции, вводные слова и выражения, фразовые глаголы, восклицательные предложения, ссылки на других людей (существительные какого класса использованы). Сделайте выводы о разнице в стилях литературного и нелитературного письма; общих чертах, характерных для личного письма как жанра текста; исторических изменениях в языке личного письма на английском языке. Сравните количество предложений в прямой речи

Задание 2. Определите среднее арифметическое употребления местоимения I, времени Past Simple, сложноподчиненных предложений на каждые 300 слов и сделайте вывод о синтаксических особенностях английских текстов личного письма по сравнению с повествовательным текстом примера.

Задание 3. Определите частотность употребления частей речи (существительных, прилагательных, глаголов) в каждом из английских текстов; сделайте вывод о морфологических особенностях личного письма; в связи с этим, прокомментируйте, говорит ли письмо в большей степени о событиях или о фактах. В соответствии с полученными данными составьте круговую диаграмму распределения частей речи в каждом из текстов.

Задание 4. Определите частотность употреблений прямой речи в русском и английском вариантах литературного письма. Сделайте вывод о синтаксических особенностях.

Задание 5. Определите частотность использования лексики, обозначающей время и даты в русском и английском литературном письмах. Сделайте вывод об особенностях повествования.

Задание 6. Определите частотность восклицательных предложений в русском и английском литературном письме, сделайте вывод об особенностях авторского стиля.

Задание 7. Определите частотность и среднее арифметическое буквосочетаний, передающих краткие, долгие гласные звуки и дифтонги в сонетах Шекспира (Приложение 7). Сделайте выводы об особенностях стиля. Составьте круговую диаграмму распределения долготы гласных.

Задание 8. Определите среднее арифметическое использования сонантов, глухих и звонких согласных в каждом из сонетов Шекспира (Приложение 7). Одинаковы ли значения среднего арифметического в разных сонетах? Изучите оригинальный текст и сравните переводы Маршака, Финкеля и Степанова. С какой целью реализуются приемы аллитерации и созвучия гласных в оригинальных текстах сонетов и в переводах? Сделайте выводы, используя количественные методы.

Задание 9. Разделите каждый из текстов Приложения 4 на секторы с одинаковым количеством слов с помощью сносок о количестве слов внизу страницы Word, посчитайте среднее арифметическое обращений к получателю письма и ссылку на авторов. Сделайте выводы о распределении обращений в каждом из текстов, затем сравните два текста.

Задание 10. Представьте образец художественного текста, сформулируйте не менее 10 проблемных вопросов для анализа текста с помощью количественных методов. Представьте результаты в количественном и графическом виде, дайте интерпретацию результатов.

Тема 2. Дисперсионный анализ в лингвистике. Репрезентация массива и диапазона данных. Нормальное распределение. Среднее квадратическое отклонение

Задание 1.

Посчитайте дисперсию и среднее квадратическое отклонение в программе Excel на базе газетных текстов (Приложения 5,6) для: употреблений определенного артикля, determiners, времен Simple Active and Passive, сложноподчиненных предложений, предлогов, прилагательных, существительных, причастий.

Постройте точечные графики рассеивания для каждой из величин и сделайте выводы.

Задание 2. В текстах Приложений 5,6 посчитайте количество сложноподчиненных предложений определительных (relative clauses) и сложноподчиненных предложений обстоятельственных (adverbial clauses).

Постройте точечные диаграммы дисперсии для каждого из видов предложений.

Постройте график нормального распределения с двумя переменными (виды придаточных предложений). Сделайте выводы.

Задание 3.

С помощью функции Word (выделение текста и автоматический подсчет количества слов) определите количество слов в каждом из текстов Приложений 5, 6. Посчитайте дисперсию, составьте точечный график и график нормального распределения. Является ли объем газетного текста закономерной или неограниченной величиной?

Задание 4. Посчитайте дисперсию для слов, обозначающих время и времена года в сонете 104 и трех вариантах перевода. Насколько велика разница этих величин для английского и русского текстов? Каким образом величина дисперсии на лексическом уровне отражает смысл сонета?

Задание 5. Исследуйте дисперсию слов, выражающих понятия «любовь», «тоска», «цвет» в сонетах Шекспира и переводах (Приложение 7). Простройте графики рассеяния для каждой из переменных (для каждого концепта) и сравните их. Проанализируйте расположение точек вокруг оси центральной тенденции, сделайте выводы о степени выраженности каждого из концептов в сонетах.

Тема 3. Регрессионный анализ в лингвистике.

Задание 1. На основе текста Skin (Приложение 3) определите корреляцию сочетаемости глаголов на каждой из страниц текста. В столбцах матрицы укажите части речи и формы, которые следуют за глаголом (инфинитив, другие). В строках матрицы укажите страницы текста последовательно. Постройте графики корреляции переменных. Сделайте выводы.

Задание 2. Посчитайте корреляцию предлогов of, at в тексте Приложения 3. Постройте матрицу и графики, сделайте выводы.

Задание 3. Используя тексты Приложения 2, посчитайте корреляцию глагола make с другими словоформами, сгруппировав их предварительно в столбцах матрицы. В строках матрицы укажите каждый из текстов. Сделайте выводы о наиболее высоких корреляциях. Проверьте данные опытным путем, используя корпусы текстов.

Задание 4. Определите корреляцию буквы t в конце слов в каждом из сонетов (Приложение 7) с последующим звуком. Какие буквосочетания коррелируют в большей степени? Какие фонетические законы сочетаемости звуков они реализуют? Как сказывается корреляция на фонетическом рисунке сонетов?

Задание 5. Определите корреляцию «положительных» и «отрицательных оттенков значений по трем частям речи: глаголам, существительным и прилагательным во всех сонетах и переводах (Приложение 7). Составьте две матрицы соответственно для «положительных» и «отрицательных» значений. Для какой из матриц степень корреляции выше? Сделайте выводы о том, какие морфологические средства использовали переводчики, чтобы выразить смысл сонетов.

Тема 4. Кросскорреляционный анализ в лингвистике

Задание 6. Определите корреляцию глагола have (has) на основе текстов Приложения 2 с другими словоформами. Постройте график, сделайте вывод.

Задание 7. Определите корреляцию слов с одинаковым значением (например, coral – коралл) в каждом из оригинальных сонетов и каждом из приведенных переводов. Для каждого из сонетов составьте матрицу и посчитайте корреляцию переменных (слов). Сделайте выводы о точности перевода каждого сонета для каждого из русских текстов.

Задание 8. На основе текстов Приложения 5 определите корреляцию имен собственных и чисел в каждом из текстов. Сделайте выводы о сочетаемости этих языковых явлений.

Задание 9. В английском и русском текстах Приложения 4 содержится немало сравнений (явных и скрытых), выраженных различными языковыми средствами. Установите корреляцию между способами выражения сравнения во всех трех текстах. Постройте графики.

Задание 10. Представьте образец текста учебника на английском языке. Проанализируйте корреляцию грамматических и лексических явлений. Данные представьте графически.

Тема 5. Факторный анализ в лингвистике

Задание 1. В текстах Приложения 2 проведите факторный анализ сочетаемости исчисляемых и неисчисляемых имен существительных, обозначающих конкретные и абстрактные понятия (анализируйте слова, стоящие перед каждым из существительных).

Задание 2. В английских текстах Приложения 4 проведите факторный анализ употребления предлогов. Сравните данные, полученные в результате вращения и интерпретации факторов, с данными, приведенными в разделе 1.4.

Задание 3. Проведите факторный анализ односложных слов и частиц (ну, не) в русском тексте Приложения 4. Сделайте вывод об особенностях стиля письма.

Задание 4. Используя тексты переводов сонета 50 (Приложение 7), проведите факторный анализ использования слова «конь» и его производных тремя авторами. Используйте слова, стоящие до и после этих словоформ. Сделайте выводы об используемых переводчиками ассоциациях.

Задание 5. Проведите факторный анализ сочетаемости гласных букв в диграфах во всех английских текстах сонетов (Приложение 7). Интерпретируйте полученные данные с точки зрения передаваемых звуков, а также звукового и поэтического эффекта, достигаемого с их помощью.

Тема 6. Кластерный анализ в лингвистике

Задание 1. Используя корпус текстов J.K.Rowling, выделите десять самых частотных слов и проведите кластерный анализ.

Задание 2. Используя Приложения 5 и 6, в каждом из текстов посчитайте количество видо-временных форм глагола и проведите кластерный анализ.

Задание 3. Проанализируйте фонетические окончания рифм в сонетах Приложения 7, добавьте несколько сонетов для полноты выборки и выделите кластеры наиболее характерных рифм.

Задание 4. Используя данные корпуса текстов, внесите в матрицу данные частотности вспомогательных глаголов в 5 выборках. Проведите кластерный анализ.

Задание 5. Используя тексты Приложений 2,5,6, посчитайте частотность различных окончаний имен прилагательных. Проведите кластерный анализ.

Тема 7. Вероятностный анализ языковых явлений в тексте

Задание 1. Используя тексты Приложения 5, посчитайте биномиальное распределение вероятности использования существительных в единственном числе постранично.

Задание 2. Разделите тексты приложения 6 на выборки с одинаковым количеством слов. Посчитайте частотность использования слова said в первой выборке. Определите вероятность использования этого слова в следующих выборках. Сделайте вывод о том, насколько характерна косвенная речь для текстов избранного типа.

Задание 3. Выберите один из сонетов У. Шекспира и посчитайте количество сонантов в первой строке. Определите частотность сонантов в строке по отношению к общему количеству звуков. Определите вероятность использования сонантов во всем сонете. Как полученные данные характеризуют мелодический рисунок сонета? Вычислите вероятность, используя функцию полиномиального распределения (гласные, согласные, сонанты) в оном из сонетов.

Задание 4. На примере текста Приложения 1 определите частотность употребления неопределенного артикля с помощью функций биномиального распределения и гипергеометрического распределения.

Задание 5. Выберите один из сонетов У. Шекспира из Приложения 7. Посчитайте количество словоупотреблений местоимения I. Допуская, что количество слов в каждом из сонетов примерно одинаково, вычислите вероятность его появления во всех сонетах Приложения 7. Сделайте выводы о том, насколько обращения автора персонифицированы.

4.3 Промежуточная аттестация по дисциплине проводится в форме зачета

Типовые вопросы зачета (ОПК-7)

1. Дайте определение квантитативной лингвистики и математической статистики.
2. Дайте подробное описание понятия qualitative vs. quantitative linguistics.
3. Дайте определение выборки и частотности, расскажите о функциях, которые они выполняют в лингвистических исследованиях.
4. Назовите статистические методы, используемые в лингвистике.
5. Дайте определение дисперсии, отклонения, среднего квадратического отклонения и приведите примеры использования этих величин в лингвистике.
6. Дайте определение регрессионного анализа и центральной тенденции.
7. Назовите методы графического представления дисперсии.
8. Дайте определение корреляционного анализа и корреляции в лингвистике. Приведите примеры его использования в анализе текста.

9. Дайте определение факторного анализа. Приведите примеры использования этого метода в анализе семантики письменных текстов.

10. Опишите возможности программы Tropes в анализе семантики текста.

11. Назовите виды распределений и дайте понятие вероятности. Приведите примеры вероятностного анализа текста.

12. Дайте определение языкового корпуса. Опишите возможности использования языкового корпуса как экспериментальной выборки в квантитативной лингвистике.

Типовые задания для зачета (ОПК-7)

Не предусмотрены

4.4. Шкала оценивания промежуточной аттестации

Оценка	Компетенции	Дескрипторы (уровни) – основные признаки освоения (показатели достижения результата)
«зачтено» (50 - 100 баллов)	ОПК-7	Демонстрирует прочные знания по использованию статистических методов в лингвистических исследованиях для решения поставленных задач
«не зачтено» (0 - 49 баллов)	ОПК-7	Не имеет прочных знаний по использованию статистических методов в лингвистических исследованиях для решения поставленных задач

5. Методические указания для обучающихся по освоению дисциплины (модуля)

5.1 Методические указания по организации самостоятельной работы обучающихся:

Приступая к изучению дисциплины, в первую очередь обучающимся необходимо ознакомиться содержанием рабочей программы дисциплины (РПД), которая определяет содержание, объем, а также порядок изучения и преподавания учебной дисциплины, ее раздела, части.

Для самостоятельной работы важное значение имеют разделы «Объем и содержание дисциплины», «Учебно-методическое и информационное обеспечение дисциплины» и «Материально-техническое обеспечение дисциплины, программное обеспечение, профессиональные базы данных и информационные справочные системы».

В разделе «Объем и содержание дисциплины» указываются все разделы и темы изучаемой дисциплины, а также виды занятий и планируемый объем в академических часах.

В разделе «Учебно-методическое и информационное обеспечение дисциплины» указана рекомендуемая основная и дополнительная литература.

В разделе «Материально-техническое обеспечение дисциплины, программное обеспечение, профессиональные базы данных и информационные справочные системы» содержится перечень профессиональных баз данных и информационных справочных систем, необходимых для освоения дисциплины.

5.2 Рекомендации обучающимся по работе с теоретическими материалами по дисциплине

При изучении и проработке теоретического материала необходимо:

- просмотреть еще раз презентацию лекции в системе MOODLe, повторить законспектированный на лекционном занятии материал и дополнить его с учетом рекомендованной дополнительной литературы;

- при самостоятельном изучении теоретической темы сделать конспект, используя рекомендованные в РПД источники, профессиональные базы данных и информационные справочные системы:

- ответить на вопросы для самостоятельной работы, по теме представленные в пункте 3.2 РПД.

- при подготовке к текущему контролю использовать материалы фонда оценочных средств (ФОС).

5.3 Рекомендации по работе с научной и учебной литературой

Работа с основной и дополнительной литературой является главной формой самостоятельной работы и необходима при подготовке к устному опросу на семинарских занятиях, к дебатам, тестированию, экзамену. Она включает проработку лекционного материала и рекомендованных источников и литературы по тематике лекций.

Конспект лекции должен содержать реферативную запись основных вопросов лекции, в том числе с опорой на размещенные в системе MOODLe презентаци, основных источников и литературы по темам, выводы по каждому вопросу. Конспект может быть выполнен в рамках распечатки выдачи презентаций лекций или в отдельной тетради по предмету. Он должен быть аккуратным, хорошо читаемым, не содержать не относящуюся к теме информацию или рисунки.

Конспекты научной литературы при самостоятельной подготовке к занятиям должны содержать ответы на каждый поставленный в теме вопрос, иметь ссылку на источник информации с обязательным указанием автора, названия и года издания используемой научной литературы. Конспект может быть опорным (содержать лишь основные ключевые позиции), но при этом позволяющим дать полный ответ по вопросу, может быть подробным. Объем конспекта определяется самим студентом.

В процессе работы с основной и дополнительной литературой студент может:

- делать записи по ходу чтения в виде простого или развернутого плана (создавать перечень основных вопросов, рассмотренных в источнике);

- составлять тезисы (цитирование наиболее важных мест статьи или монографии, короткое изложение основных мыслей автора);

- готовить аннотации (краткое обобщение основных вопросов работы);

- создавать конспекты (развернутые тезисы).

5.4. Рекомендации по подготовке к отдельным заданиям текущего контроля

Собеседование предполагает организацию беседы преподавателя со студентами по вопросам практического занятия с целью более обстоятельного выявления их знаний по определенному разделу, теме, проблеме и т.п. Все члены группы могут участвовать в обсуждении, добавлять информацию, дискутировать, задавать вопросы и т.д.

Устный опрос может применяться в различных формах: фронтальный, индивидуальный, комбинированный. Основные качества устного ответа подлежащего оценке:

- правильность ответа по содержанию;

- полнота и глубина ответа;

- сознательность ответа;

- логика изложения материала;

- рациональность использованных приемов и способов решения поставленной учебной задачи;

- своевременность и эффективность использования наглядных пособий и технических средств при ответе;

- использование дополнительного материала;

- рациональность использования времени, отведенного на задание.

Устный опрос может сопровождаться презентацией, которая подготавливается по одному из вопросов практического занятия. При выступлении с презентацией необходимо обращать внимание на такие моменты как:

- содержание презентации: актуальность темы, полнота ее раскрытия, смысловое содержание, соответствие заявленной темы содержанию, соответствие методическим требованиям (цели. ссылки на ресурсы. соответствие содержания и литературы), практическая направленность, соответствие содержания заявленной форме, адекватность использования технических средств учебным задачам, последовательность и логичность презентуемого материала;

- оформление презентации: объем (оптимальное количество), дизайн (читаемость, наличие и соответствие графики и анимации, звуковое оформление, структурирование информации, соответствие заявленным требованиям), оригинальность оформления, эстетика, использование возможности программной среды, соответствие стандартам оформления;

- личностные качества: ораторские способности. соблюдение регламента, эмоциональность, умение ответить на вопросы, систематизированные, глубокие и полные знания по всем разделам программы:

- содержание выступления: логичность изложения материала, раскрытие темы, доступность изложения, эффективность применения средств ИКТ, способы и условия достижения результативности и эффективности для выполнения задач своей профессиональной или учебной деятельности, доказательность принимаемых решений, умение аргументировать свои заключения, выводы.

6. Учебно-методическое и информационное обеспечение дисциплины

6.1 Основная литература:

1. Грачева А. С. Статистические методы анализа оригинала и перевода художественного текста : студенческая научная работа. - Уфа: б.и., 2023. - 75 с. - Текст : электронный // ЭБС «Университетская библиотека онлайн» [сайт]. - URL: https://biblioclub.ru/index.php?page=book&id=704252

2. Хруничев Р. В. Прикладные статистические методы анализа : учебное пособие. - Рязань: РГРТУ, 2023. - 80 с. - Текст : электронный // ЭБС "ЛАНЬ" [сайт]. - URL: https://e.lanbook.com/book/380498

3. Долинский В. А. Квантитативная лингвистика и новые информационные технологии : курс лекций. - Москва: Московский государственный лингвистический университет (МГЛУ), 2022. - 352 с. - Текст : электронный // ЭБС «Университетская библиотека онлайн» [сайт]. - URL: https://biblioclub.ru/index.php?page=book&id=710346

4. Шунейко А. А., Авдеенко И. А. Квантитативная лингвистика и новые информационные технологии : учебник для вузов. - Москва: Юрайт, 2025. - 347 с. - Текст : электронный // ЭБС «ЮРАЙТ» [сайт]. - URL: https://urait.ru/bcode/567628

6.2 Дополнительная литература:

1. Моисеева И. Ю. Квантитативная лингвистика и новые информационные технологии : учебное пособие. - Оренбург: Оренбургский государственный университет, 2017. - 103 с. - Текст : электронный // ЭБС «Университетская библиотека онлайн» [сайт]. - URL: http://biblioclub.ru/index.php?page=book&id=481797

2. Агалаков С. А. Статистические методы анализа данных : учебное пособие. - Омск: Омский государственный университет им. Ф.М. Достоевского, 2017. - 92 с. - Текст : электронный // ЭБС «Университетская библиотека онлайн» [сайт]. - URL: http://biblioclub.ru/index.php?page=book&id=562918

3. Шайкевич А.Я. Введение в лингвистику : Учеб. пособие. - М.: Академия, 2005. - 394 с.

4. Кащеева А.В., Мильруд Р.П. Квантитативные методы в лингвистике : учеб. пособие. - Тамбов: [Издат. дом ТГУ им. Г.Р.Державина], 2012. - 123 с.

6.3 Иные источники:

1. Языкознание.ру - ресурс, созданный для изучающих различные лингвистические дисциплины. Информация, представленная на сайте, имеет, прежде всего, справочный характер. Данная информация может быть полезна не только студентам-лингвистам, но и преподавателям лингвистики. - http://yazykoznanie.ru/

2. Электронная лингвистическая библиотека - www.superlinguist.ru

3. Электронная гуманитарная библиотека - http://www.gumfak.ru/

4. Словари и энциклопедии он-лайн - http://dic.academic.ru

5. Сборник статистики - http://uucyc.ru/statistics/

7. Материально-техническое обеспечение дисциплины, программное обеспечение, профессиональные базы данных и информационные справочные системы

Для проведения занятий по дисциплине необходимо следующее материально-техническое обеспечение: учебные аудитории для проведения занятий лекционного и семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, помещения для самостоятельной работы.

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

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

Для проведения занятий лекционного типа используются наборы демонстрационного оборудования, обеспечивающие тематические иллюстрации (проектор, ноутбук, экран/ интерактивная доска).

Лицензионное и свободно распространяемое программное обеспечение:

7-Zip 9.20

Adobe flash player

Kaspersky Endpoint Security для бизнеса - Стандартный Russian Edition. 1500-2499 Node 1 year Educational Renewal Licence

Microsoft Office Профессиональный плюс 2007

Office 2007, 2010, 2016

SPSS Statistic

Statistica Base 10 for Windows RU

Профессиональные базы данных и информационные справочные системы:

1. Цифровой образовательный ресурс IPR SMART. – URL: http://www.iprbookshop.ru
2. Scopus: база данных . – URL: https://www.scopus.com
3. Архив научных журналов зарубежных издательств. – URL: https://arch.neicon.ru
4. ЭБС «Университетская библиотека онлайн» . – URL: http://www.biblioclub.ru
5. Электронная библиотека ТГУ. – URL: https://elibrary.tsutmb.ru/
6. Электронная библиотека. Образовательная платформа «Юрайт». – URL: https://biblio-online.ru/book/sud-prisyazhnyh-442275

Электронная информационно-образовательная среда

https://auth.tsutmb.ru/authorize?response_type=code&client_id=moodle&state=xyz

Взаимодействие преподавателя и студента в процессе обучения осуществляется посредством мультимедийных, гипертекстовых, сетевых, телекоммуникационных технологий, используемых в электронной информационно-образовательной среде университета.