КОЛИЧЕСТВЕННЫЕ МЕТОДЫ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА

Лекции № 4-5

Методика и этапы статистического эксперимента

Курс лекций

доцента кафедры перевода и информационных технологий в лингвистике ЮФУ

Агапова Анатолия Михайловича

ТЕСТ

1. Дата, Лекция 4, Группа, ФИО *(разборчиво)*

2. Назовите первые 3 этапа статистического эксперимента

3. Наиболее приемлемое определение слова в лингвостатистических исследованиях

4. Решение каких задач предусматривает методика создания ВЛС

Вопросы для рассмотрения на лекциях 4-5

«Методика и этапы статистического эксперимента»:

1. Формулирование и дальнейшее уточнение цели исследования.
2. Определение единицы ***анализа*** и единицы ***счёта***.
3. Методика сбора информации: генеральная и выборочная лингвистические совокупности, выборочный метод.
4. Методика создания выборочной лингвистической совокупности (выборки) и репрезентативность выборки.
5. Определение минимально достаточного объёма выборки при заданных относительной ошибке (**δ**) и надёжности (**ρ**).

Литература

1. Турыгина Л.А. Моделирование языковых структур средствами вычислительной техники. – М., Высшая школа, 1988. **сс. 19-24.**

2. Пиотровский Р.Г. и др. Математическая лингвистика. Учебное пособие для пед. ин-тов.– М.: Высшая школа, 1977. **сс. 219-222 и 294-301**

3. Головин Б.Н. Язык и статистика. – М., Просвещение, 1971. **сс. 26-27.**

О методике статистического эксперимента

Литература

1. **Турыгина Л.А**. Моделирование языковых структур средствами вычислительной техники. – М., Высшая школа, 1988. **сс. 19-21 – *изучить к следующей лекции***

2. **Пиотровский Р.Г. и др.** Математическая лингвистика. Учебное пособие для пед. ин-тов.- М.: Высшая школа, 1977. **сс. 219-222 – *изучить к следующей лекции***

Статистический анализ в исследовании языковых структур

Целью статистического анализа является исследование **совокупности** однородных лингвистических объектов (ЛЕ), обладающих признаком/признаками, составляющим предмет нашего анализа (ГЛС). Если ГЛС очень велика, то исследованию подвергается некоторая обозримая ее часть, называемая ***выборкой*** (ВЛС). При этом основные статистические характеристики ВЛС мы рассматриваем как **некое приближение** характеристик ЛЕ в ГЛС, и экстраполируем свойства ЛЕ на всю ГЛС и даже на подобные однородные совокупности.

Например, если признаком ЛЕ является длина словоформы в пушкинском тексте, то в качестве ГЛС выступают все тексты, написанные А.С. Пушкиным. Отдельные же произведения, например «Дубровский», являются ВЛС. Если исследуются длины словоформ в русском литературном языке, то ГЛС – сумма всех русских литературных текстов, а ВЛС при этом – …? (***подумайте!)***

Формулирование цели исследования

В качестве примера рассмотрим «Частотный словарь русского языка» (под ред. Засориной), который составлялся в первую очередь для ***определения границ активного словарного состава***, для выяснения границ живой лексической системы современного РЯ (живого словоупотребления образованного человека) на основе предположения о наличии в лексическом составе языка общеупотребительного и периферического слоя.

Задачи: систематизация лексики, определение ее базы и периферии, более полная инвентаризация и систематизация словарного состава языка, и, следовательно, разграничение в нем активного и пассивного запаса.

Основное назначение словаря – дать достаточно полные сведения о современной лексике с учетом жанровой ее дифференциации (художественная проза, драматургия, научные и публицистические тексты, газетные и журнальные тексты) с целью лингвистических и лексикостатистических интерпретаций.

Статистические данные «Частотного словаря» могут быть использованы для анализа словообразования и определения активных средств словообразования современного РЯ, для усовершенствования графики и орфографии, для практической транскрипции и транслитерации, при решении вопросов автоматизации печатного дела, распознавания и автоматического чтения буквенного текста и …

Определение единицы анализа и единицы счёта

При квантитативном исследовании необходимо выявить **единицы анализа** и определить **единицы счёта**. В зависимости от задач исследования (**единиц анализа)** за **единицу счета** можно принять: букву, фонему, морфему, словоформу, слово, словосочетание, предложение, текст, страницу, печатный знак и т.п. В лексикостатистических работах — слово, словоформу, словоупотребление, лемму.

Массовое статистическое обследование ЛЕ может быть осуществлено только на базе формальной процедуры ➔ наиболее приемлемым в лингвостатистическом исследовании является определение Генри Глúсона: «слово – отрезок текста, заключенный между двумя пробелами». ***Мы*** будем называть: цепочку букв, заключенную между двумя пробелами в тексте и имеющую одно значение, *словоупотреблением*, полностью совпадающие словоупотребления – *словоформами*, сумму семантически и грамматически связанных между собой словоформ – *словом*, словарную словоформу – *леммой* (слово в основной, исходной форме: им. п. ед. ч. – для именных форм и инфинитив – для глагольных форм). Словоупотребление является единицей текста (речь), слово – единицей словаря (язык).

Второй принципиальный вопрос в лингвостатистических исследованиях при массовом обследовании больших массивов текста – принадлежность ЛЕ к той или иной категории (часть речи, тип предложения, лексико-грамматический класс слов и т.п.) из-за различных точек зрения и отсутствия унификации.

Методика сбора информации (ГЛС и ВЛС)

Методика отбора текстов (методика создания ВЛС) по исследуемой проблеме имеет принципиальное значение. Необходимо, исходя из **цели** исследования определить ГЛС и её структуру, т.е. разработать методику, обеспечивающую представление текстов, репрезентативных с точки зрения отображения лингвостатистических свойств языковых объектов.

Методика создания ВЛС предусматривает решение следующих задач:

1) качественное и количественное распределение материала по темам, подтемам, разделам

2) установление хронологических рамок источников и документов.

Выделение тем, подтем, разделов обычно подсказывается композицией и содержанием исследуемой совокупности текстов и консультацией со специалистами (экспертами) данной области знаний. Количественное (%) распределение исходных подтем, разделов осуществляется, как правило, в той пропорции, которая наблюдается в корпусе текстов для моделируемого подъязыка.

Отбор источников непосредственно связан с определением хронологических рамок исследуемых документов. В этом случае должны быть удовлетворены два требования: 1) надежная репрезентация тематических выборок в достаточно широком диапазоне времени и, одновременно, 2) представление материала, отображающего основные свойства данного подъязыка.

Вопросы для рассмотрения на лекции 5

1. Требования к ВЛС – репрезентативность и рациональный объём выборки.

2. Репрезентативность выборки, приёмы, обеспечивающие надёжную репрезентативность тематических выборок.

3. Относительная ошибка **δ** и надежность **ρ** в лингвистических исследованиях.

4. Определение минимально достаточного объёма выборки в грамматических, фонетико-фонологических и лексикологических исследованиях.

Литература

1. Турыгина Л.А. Моделирование языковых структур средствами вычислительной техники. – М., Высшая школа, 1988. **сс. 22-25 – *изучить к следующей лекции***

2. Головин Б.Н. Язык и статистика. – М., Просвещение, 1971. **сс. 26-27 – *выучить!***

3. Пиотровский Р.Г. и др. Математическая лингвистика. Учебное пособие для пед. ин-тов.– М.: Высшая школа, 1977. **сс. 219-222 – *повторить к следующей лекции***

сс. 294-301 – *изучить к следующей лекции*

ТЕСТ

1. Дата, Лекция 5, Группа, ФИО *(разборчиво)*

2. Что понимают под репрезентативностью в лингвостатистических исследованиях

3. Перечислите виды выборочного изучения

4. Укажите вид/виды выборочного изучения, применяемые в лингвистике чаще всего

Репрезентативность и рациональный объём выборки

Ценность выводов всякого лингвистического исследования измеряется степенью **достоверности**. Лучшим средством её оценки является проверка полученных выводов на практике. Но такую проверку можно осуществить зачастую лишь после завершения самого исследования. Между тем хотелось бы уже при постановке эксперимента прогнозировать достоверность получаемых результатов.

Наиболее важными для получения заслуживающих доверия выводов являются вопросы о **репрезентативности** ВЛС и о её **рациональном объеме**. При этом необходимо заранее определиться с такими понятиями, как «**степень точности или относительная ошибка (δ)**» и «**надежность (ρ) наших суждений**».

Под **репрезентативностью** понимают способность ВЛС отражать все исследуемые свойства ЛЕ в той пропорции, которая наблюдается в ГЛС, т. е. частота исследуемых свойств ЛЕ должна быть близка соответствующей частоте в ГЛС.

Проблема определения рационального объёма ВЛС (произвольного содержания) далека от решения. На практике объём выборки определяется возможностями исследователя, исходящего из правила «чем больше выборка, тем достовернее результаты». Тем не менее разработано множество различных процедур по определению объёма выборок тематически ограниченного содержания. К сожалению, многие простые формулы для расчёта достаточного объёма выборки базируются на неявном предположении о репрезентативности ВЛС.

Виды выборочного изучения

Случайным является такой **отбор**, при котором все элементы генеральной совокупности (ГС) имеют равную возможность быть отобранными, т. е. для каждого элемента ГС обеспечена равная вероятность попасть в выборку. Случайность отбора достигается на практике с помощью жребия или таблицы случайных чисел.

Механический отбор производится следующим образом. Если формируется 10%-ная выборка, т. е. из каждых десяти элементов должен быть отобран один, то вся ГС условно разбивается на равные части по 10 элементов. Затем из первой десятки выбирается случайным образом элемент. Например, жеребьевка указала девятый номер. Тогда выборка будет состоять из элементов 9, 19, 29 и т.д. Механическим отбором следует пользоваться осторожно, так как существует реальная опасность возникновения систематических ошибок, если элементы ГС расположены неслучайным образом. Механический отбор, как никакой другой, широко использовался в русской и советской статистике.

При **серийном отборе** вся совокупность разбивается на группы (серии). Затем путем случайного или механического отбора выделяют определенную часть этих серий и производят их сплошную обработку. По сути дела, серийный отбор представляет собой случайный или механический отбор, осуществленный для укрупненных элементов исходной совокупности. В теоретическом плане серийная выборка является самой несовершенной из рассмотренных.

Типический отбор. Следует отличать типический отбор от отбора типичных объектов. При собственно типическом отборе в выборочном методе ГС разбивается на группы, однородные в качественном отношении, а затем внутри каждой группы производится случайный отбор. Типический отбор организовать сложнее, чем собственно случайный, т.к. необходимы определенные знания о составе и свойствах ГС, но зато он дает более точные результаты.

В лингвистике ***типический*** отбор чаще всего сочетается с ***серийным***. Пример: текстовые базы данных (корпуса), где количество серий, извлекаемых из каждой тематической группы определяется удельным весом этой группы в ГС.

Кроме описанных выше ***классических способов*** отбора в практике выборочного метода используются и другие способы.

Приёмы, обеспечивающие надёжную репрезентативность ВЛС

Определение структуры ГЛС, исходя из цели исследования. Решаются вопросы о представлении ГЛС в виде схемы её областей (качественное и количественное распределение материала; хронологические рамки источников). Количественная структура строится либо по принципу равнопропорциональности её областей (?? произвол!!), либо по экспертным или иным оценкам объёма структурных элементов ГЛС (опять-таки произвол, но «коллективный и/или обоснованный»!!).

Виды комплектования ВЛС: 1. Случайный отбор. 2. Механический отбор. 3. Серийный отбор. 4. Типический отбор.

Определение размера минимальной выборки. ВЛС представляет собою сумму «кусков» текстов ГЛС. Очевидно, что нужно, чтобы они были одинаковой длины и их было как можно больше. Какой же должна быть оптимальная длина таких «атомов» ВЛС? Очевидно, что она зависит от общей длины ВЛС. Для каждой задачи эта проблема решается по-своему: для лексико-статистического моделирования чаще всего выбирают тексты в 2000 или 1000 (иногда 500) словоупотреблений при общей длине ВЛС в 1 млн. словоупотреблений, при изучении газетного стиля в качестве минимальной единицы выбирают полный номер газеты. Большой опыт в решении этой проблемы накоплен создателями различных частотных словарей и корпусов текстов.

Относительная ошибка и надежность

Ошибка ли так называемая «**относительная ошибка**»? Нет, в нашем случае это – величина, которая характеризует ширину доверительного интервала, в который попадает относительная частота исследуемого свойства ЛЕ. Говорить о частотах свойств ЛЕ, как о конкретных числах нельзя – можно лишь об интервалах, которые покрывают значения исследуемых параметров. Если мы вычислили частоту **f** какого-либо свойства (параметра и т. п.) ЛЕ с относительной ошибкой **δ**, то это означает, что реальная частота попадает в интервал от **(f - δ*f)** до **(f + δ*f)**.

Например: **f** =0,12, **δ**=0,2 (20%), тогда реальная частота находится в интервале от 0,096 (f - δ*f = 0,12-0,024) до 0,144 (f + δ*f = 0,12+0,024).

«**Надежность (ρ)**», измеряемую в % или в виде десятичной дроби (например: 92% или 0,92), трактуют обычно так. Пусть мы провели один опыт на выборке **A** и получили частоту свойства ЛЕ **f** с относительной ошибкой **δ.** Тогда надёжность **ρ**=0,92 означает, что если мы возьмём 100 аналогичных **A** выборок, то в 92 (**ρ**) из них относительная частота **f** будет находиться в пределах от **(f - δ*f)** до **(f + δ*f)** и лишь в 8 из них может выходить за эти пределы.

Определение минимально достаточного объёма выборки в грамматических, фонетико-фонологических и лексикологических исследованиях

Выбор уровня точности, как и выбор надежности, зависит от той дисциплины, которая использует статистические приемы. Если для техники относительная ошибка **δ** в 2,7% может рассматриваться как предельная, то для лингвистики такая точность приведёт к неразумному увеличению объема выборки и к неоправданному расходованию сил исследователя на механическую нетворческую работу.

Принято считать, что в фонетико-фонологических и грамматических исследованиях относительная ошибка не должна превышать 0,2 (20%), а при анализе лексики и фразеологии может достигать 0,33-0,35 (33–35%).

При исследовании частоты **f** какого-либо свойства ЛЕ при заданных относительной ошибке **δ** и надёжности **ρ** для определения рационального объёма **N** выборки будем придерживаться следующей процедуры: на предварительной, небольшой по объёму, выборке проведём требуемое исследование. Получив приближённое значение частоты $\mathbf{f_1}$, определим предварительную величину минимально достаточного объёма выборки по формуле: $\mathbf{N = z_\rho^2 * (1 - f_1) / \delta^2 * f_1}$. ($\mathbf{z_{\rho=0,95}} \approx 1,96$)

Требуемое исследование проведём затем на выборке объёмом **N**. Если величина $\mathbf{N_1}$ нового минимально достаточного объёма выборки больше **N**, то необходимо повторить последний шаг.

1. Определить относительную частоту **f** употребления Булгаковым буквы «**а**» в следующей выборке из романа «Мастер и Маргарита»:

"В саду было тихо. Но, выйдя из-под колоннады на заливаемую солнцем верхнюю площадь сада с пальмами на чудовищных слоновых…“

2. Определить минимально достаточный объём выборки **N** при заданных относительной ошибке $\delta = 0{,}2$ (20%) и надёжности $\rho = 0{,}95$ (95%).

3. Объяснить, что означают в данном исследовании «относительная ошибка», «надёжность» и «минимально достаточный объём выборки».

Решение:

1. Всего букв в выборке – 100, буквы «**а**» – 11 ➔ $f(\text{«а»}) = 0{,}11$

2. $N = z_\rho^2 * (1 - f) / \delta^2 * f$ ➔ при $\delta = 0{,}2$ и $\rho = 0{,}95$ $(z_\rho \approx 1{,}96)$ (≈ 2)

$N \approx (1{,}96)^2 * (1-0{,}11) / ((0{,}2)^2 * 0{,}11) \approx 2^2 * 0{,}89 / (0{,}04 * 0{,}11) \approx 89 / 0{,}11 \approx$ **810**

3. Если мы возьмём **100** аналогичных выборок из романа по **N** (**810**) букв, то в **95** из них относительная частота буквы «**а**» будет находиться в пределах от **0,088** $(f - \delta * f = 0{,}11 - 0{,}022)$ до **0,132** $(f + \delta * f = 0{,}11 + 0{,}022)$, и лишь в **5** может выходить за эти пределы.

***Реально**: в выборке из романа объёмом* **60 000 букв** $f(\text{«а»}) \approx 0{,}088$

Домашнее задание

ОБРАЗЕЦ ТЕСТА ИТОГОВОГО КОНТРОЛЯ

Исследование употребления М. Булгаковым различных букв русского алфавита

I. Предварительный эксперимент

Определить относительную частоту **f** употребления двух букв «**о, е**» в следующей небольшой выборке из романа М.А. Булгакова «Мастер и Маргарита»: «Она несла в руках отвратительные, тревожные желтые цветы. Черт их знает, как их зовут, но они первые почему-то появляются в Москве. И эти цветы очень отчетливо выделялись на черном ее весеннем пальто. Она несла желтые цветы! Нехороший цвет...».

II. Определение минимально достаточного объёма выборки

По данным предварительного эксперимента определить минимально достаточный объём выборки **N** для получения достоверных сведений об употреблении Булгаковым в художественных произведениях букв русского алфавита «**о, е**» при заданных относительной ошибке $\delta = 0{,}15$ (15%) и надёжности $\rho = 0{,}95$ (95%).

III. Анализ

сформулировать возможные цели данного исследования, определить возможные единицы анализа и счёта;

провести анализ генеральной лингвистической совокупности (ГЛС);

описать методику комплектования выборки (ВЛС) с помощью случайного, механического, серийного или типического отбора для полученного объема выборки;

оценить репрезентативность выборки (ВЛС) с точки зрения целей исследования, используя понятия «относительная ошибка», «надёжность» и «минимально достаточный объём выборки».

Пример (на основе канд. диссертации М.Н. Садовниковой)

Поставим ***цель***: проанализировать функционирование сложноподчинённых предложений (СПП) в газетных заголовках современных СМИ разных стран.

Объект исследования – **1**.современные газетные заголовки российских и французских СМИ (**2.** СПП в современных заголовках русских и французских СМИ).

Предмет исследования – **1.**СПП в заголовках российских и французских СМИ (**2.** типы и разновидности СПП в газетных заголовках, прагматические функции).

Методы анализа: основной – *описательный* (синхронный, диахронический, сопоставительный анализ); вспомогательный – *количественный* (диахроническое изменение частоты СПП в заголовочной позиции, сопоставительный аспект регулярности и продуктивности заголовков-СПП, …).

Единицы анализа: типы и разновидности СПП-заголовков.

Единицы счёта: «условная» газета, «условная» страница газеты, газетный текст, газетный заголовок, предложение в заголовочной позиции, СПП-заголовок, типы и разновидности СПП-заголовков.

ГЛС: 1. все российские и все французские газеты за определённый период (конкретизация нечёткого понятия «современный»).

2. один или группа жанровых типов всех российских и французских газет (напр., информационный или/и рекламный жанр) ➔ уточнить цель!

ВЛС: возможные выборки

ВЛС: Газетные заголовки центральных и местных российских газет: „Комсомольская правда", „Труд", „Российская газета", „Известия", „Мегаполис-Экспресс", „Вечерний Ростов", „Газета Дона" за период 1998-2001 годов.

Газетные заголовки французских периодических печатных изданий: „Gazette", „L'Ami du peuple", „Le Monde", „Le Figaro", „L'Humanite", „Journal des enfants".

В результате сплошной выборки из французских и российских газет извлечено более 3000 примеров заголовков, в которых используются различные конструкции СПП.

НО: Насколько можно доверять выводам, полученным на такой выборке???

Домашнее задание *(бонусное)*

1. Сформулировать несколько лингвистических задач (т.е. цели), которые можно решать с помощью статистического моделирования.
Образец: «ЧАСТОТНЫЙ СЛОВАРЬ РУССКОГО ЯЗЫКА (под ред. Засориной)».

2. Определить в зависимости от задачи исследования, что может быть принято за единицу анализа/счёта (буквы, фонемы, морфемы, словоформы, слова, словоупотребления, словосочетания, лексемы, предложения, текст, маркемы, …)

3. Для одной из сформулированных Вами лингвистических задач разработать методику создания ВЛС (методику отбора из ГЛС текстовой информации) и оценить репрезентативность выборки с точки зрения целей исследования.

Выполнить и сдать на следующих лекциях