ИТМО

Математическая лингвистика

Болгова Екатерина Владимировна, к.т.н.
преподаватель ФИКТ

Лингвистика

(от лат. Lingua — язык), языкозна́ние, языкове́дение —

- наука, изучающая языки.
- наука о естественном человеческом языке вообще и о всех языках мира как индивидуальных его представителях» (Лингвистический энциклопедический словарь, 1990)
- наука,исследующая сущность и природу языка, проблему его происхождения и общие законы его развития и функционирования (Ю.С. Маслов)

Разделы лингвистики

- **Когнитивная лингвистика** – исследует проблемы соотношения языка, сознания и мышления, роль языка в познании, понимании и отражении окружающей действительности.
- **Паралингвистика** – изучает неязыковые (невербальные) средства в речи, передающие совместно с вербальными смысловую информацию в составе речевого сообщения.
- **Психолингвистика** – изучает речь, речеобразование (чаще всего конкретного индивида).
- **Социолингвистика** – изучает роль языка в обществе, воздействие общества на язык, язык в связи с социальными условиями его существования.
- **...**

Лингвистика

Теоретическая (фундаментальная)

объективное установление состояния отдельного языка, его истории и закономерностей.
Отвечает на вопрос: «Каков язык?»

Прикладная

связана с разработкой методов решения практических задач использования языка.
Отвечает на вопрос: «Как лучше использовать язык?»

Прикладная лингвистика – это

- деятельность по применению научных знаний об устройстве и функционировании языка в нелингвистических научных дисциплинах и в различных сферах практической деятельности человека, а также теоретическое осмысление такой деятельности
- комплексная научная дисциплина, изучающая язык в различных ситуациях его применения и разрабатывающая *методы* совершенствования *описания* языковых систем и языковых процессов
- отрасль языкознания, занимающаяся вопросами теории языка с учетом возможностей его использования для решения современных практических задач
- раздел языкознания, направленный на решение практических задач в разных областях науки и техники, в различных сферах деятельности человека и общества на основе теоретических наработок и результатов практических исследований естественного языка

История термина

В западных странах

Появился в конце 20-х годов XX в

аналоги этого термина (например, англ. *applied linguistics*) используются прежде всего для обозначения **теории и практики преподавания языков** (лингводидактики)

В СССР

стал широко употребляться в 50-е годы XX в.

в связи с **разработкой компьютерных технологий** и появлением систем автоматической и автоматизированной обработки текстовой информации

Связи ПЛ с областями знаний

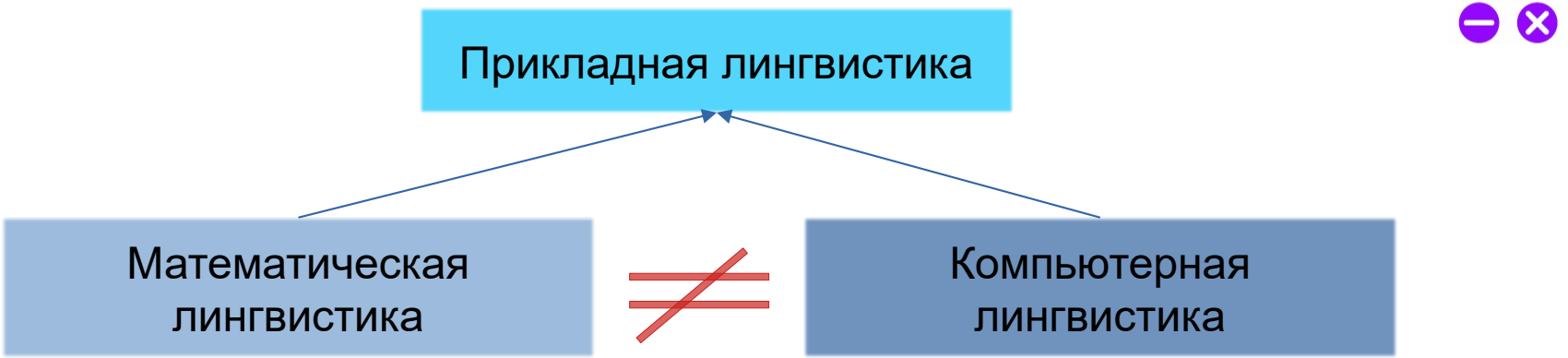
Связи?

ИТМО

Прикладная лингвистика

Математическая лингвистика

Компьютерная лингвистика


Прикладная лингвистика
Математическая лингвистика
Компьютерная лингвистика

Математическая Лингвистика (МЛ)

– дисциплина, предметом которой является разработка формального аппарата для описания строения естественных и некоторых искусственных языков.

- Направление сформировалось в середине 20 века.
- Активное развитие было обусловлено необходимостью решения проблемы автоматической обработки, хранения, поиска и передачи информации на естественном языке.

Квантитативная лингвистика

(англ. QUANTITATIVE LINGUISTICS)

- Это раздел общей лингвистики и, в частности, математической лингвистики.

- Исследует язык при помощи статистических методов; её цель — сформулировать законы, по которым функционирует язык и, в конечном счете, построить общую теорию языка в виде совокупности взаимосвязанных законов функционирования языков.

- использование статистических методов в языкознании позволяет дополнить структурную модель языка вероятностным компонентом, т.е. создать структурновероятностную модель, обладающую значительным объяснительным потенциалом.

Пример КЛ

ИТМО

Объект исследования	Количество слов
Библия (латинская)	5649
Библия (древнееврейская)	5642
Саллюстий	3394
Данте (Божественная комедия)	5860 (1615 имен собственных и географич. названий)
Тассо (Неистовый Орланд)	8474
Милтон	8000 (приблизительно)
Шекспир	15000 (по другим данным 20000)

Пример КЛ

1977 г. - «Частотный словарь русского языка» под ред. Л. Н. Засориной:

- выборка в один миллион словоупотреблений из четырёх жанров (художественная проза, драматургия, научная публицистика, газетножурнальные материалы);
- 40 тысяч слов;
- Самое частотное слово – в (во), служебные слова и местоимения (и, не, на, я, быть, что, он, с, а, как, это).
- Самое частотное существительное – год.

Пример КЛ

Определение авторства: «*Кто является истинным автором романа «Тихий Дон?*»

- Ученые взяли тексты М. Шолохова и тексты донского писателя Ф.Крюкова. Проанализировали их:
- длина предложений
- распределение длины предложений по количеству слов
- распределение частей речи
- сочетание частей речи в начале и в конце предложения
- частота применения союзов
- богатство словарного запаса
- повторяемость лексики и др.
- выборка 12 тыс. фраз, 164637 слов = 250 таблиц, формул и графиков
- ***Автор – М. Шолохов***

Лингвостатистический анализ

Что считать?	Зачем считать?	Как считать?
Определение единицы лингвостатистического исследования	Цель - исследование совокупности однородных языковых единиц	Методики. Статистические инструменты

Что считать

ИТМО

Единица лингвостатистического анализа — языковая единица любого уровня

- буквы
- фонемы
- морфемы
- слова
- словоформы
- словосочетания
- предложения
- ...

Зачем считать

- исследование совокупности однородных лингвистических объектов (лингвистических единиц), обладающих признаками, которые составляют предмет проводимого анализа

- изучаются количественные характеристики лингвистических форм — их употребительность, совместная встречаемость, законы распределения в тексте, их физические размеры

- описываются свойства текста, формулируются гипотезы о механизмах его образования и устройстве системы языка

Как считать

- **Базовые статистические понятия**

Частота

Генеральная совокупность

Выборочная совокупность

Частота –

число появлений факта/явления в наблюдаемом отрезке

Отрезок –

Любая совокупность считаемых единиц и любая среда, в которой появляются или находятся факты, поддающиеся счету

Генеральная совокупность –

вся совокупность однородных языковых единиц

Выборочная совокупность –

(выборка) часть генеральной совокупности, объединенная общим признаком

Виды выборочных совокупностей

По объему:

- Малые (менее 30 ед.)
- Средние (30-100 ед.)
- Большие (более 100)

По способу отбора:

- Случайная (простой случайный отбор)
- Механическая (систематическая)
- Серийная (гнездовая или кластерная)
- Типическая и др.;

Спасибо
за внимание!

ITMO more than a UNIVERSITY

E-mail: ekaterina_bolgova@itmo.ru
telegram: @Katerina_Bolgova

Практическая работа №1

1. Выберете некоторое количество произведений одного писателя. Для удобства упорядочите их хронологически. Для краткости назовем получившуюся ГЛС текстом «данного автора». Таким образом, текст автора (в данном определении) может состоять из нескольких различных произведений — романов, повестей, рассказов и т.п

2. Выделите из этого текста отдельные фрагменты (выборки одинакового объема или ВЛС), состоящие из одного и того же количества слов (фиксированного заранее). Эти равные по объему выборки выделяйте из текста через равные интервалы, т.е. таким образом, чтобы каждые две соседние выборки были отделены друг от друга примерно одним и тем же количеством слов. Этот интервал между соседними выборками называют шагом.

Практическая работа №1

3. Итак, последовательно двигаясь по тексту одного автора, через каждые N страниц книжного текста делайте выборки одного и того же объема. Чем длиннее исследуемый текст, тем больше выборок вы сможете сделать. Для коротких произведений число выборок будет невелико, что усложняет анализ, делает результаты неустойчивыми.

4. Выберите какой-либо лингвистический параметр (например, частоту употребления предлога «в»). Изучите эволюцию этого параметра вдоль всей ГВС. Для этого сделайте последовательные выборки и подсчитайте для каждой из них значение интересующего вас лингвистического параметра. В результате для каждой выборки (порции) получим свое число. От выборки к выборке оно будет меняться.

Практическая работа №1

5. Постройте график, отложив по горизонтали целые числа 1, 2, 3 и т.д., являющиеся номерами последовательных выборок, а по вертикали — значения изучаемой лингвистической характеристики. В результате эволюция данного параметра вдоль всего исследуемого текста изобразится некоторой ломаной линией. Она наглядно показывает поведение исследуемого параметра вдоль произведений данного автора. Такие графики очень удобны при поиске характерных черт данного автора — авторских инвариантов.

Практическая работа №1

6. Для 3-4 ВЛС, используя, например WordStat (http://www.bestfree.ru/soft/obraz/word-count.php), посчитайте частоты употребляемых слов. Сгруппируйте слова по части речи, посчитайте частотность частей речи количественно и в долях к общему числу слов (постройте график).
Далее для каждого слова посчитайте долю встречаемости в совокупности частей речи, к которой оно относится (постройте график)