КОЛИЧЕСТВЕННЫЕ МЕТОДЫ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА

Лекция № 1

Вероятностно-статистическое изучение языка и речи. Основные области приложения структурно-вероятностной модели языка.

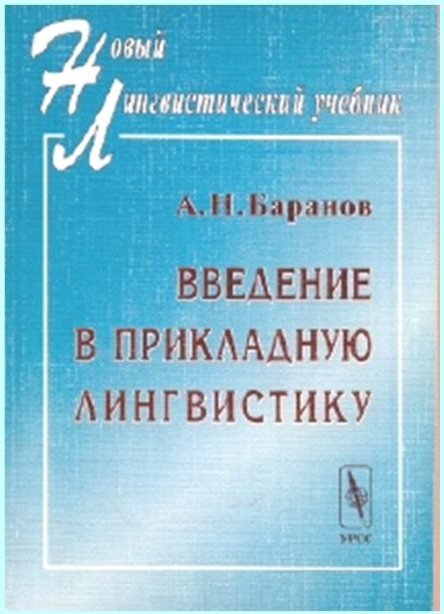
Курс лекций

доцента кафедры перевода и информационных технологий в лингвистике ЮФУ

Агапова Анатолия Михайловича

Материалы по курсу находятся на itlflis.ru

Баранов А.Н. **Введение в прикладную лингвистику:**
Учебное пособие. -М.: Эдиториал УРСС, 2001.



Турыгина Л.А. Моделирование языковых структур средствами вычислительной техники. -М.: Высшая школа, 1988. -(Б-ка филолога)

Пиотровский Р.Г. и др. Математическая лингвистика. Учебное пособие для пед. ин-тов. -М.: Высшая школа, 1977.

Головин Б.Н. Язык и статистика. - М.: Просвещение, 1971.

..... и так далее (см. список основной и доп. литературы в УМК дисциплины «КМЛА»)

ТЕСТ

1. **Дата, Группа, Фамилия, имя, отчество** (полностью, разборчиво и с ударениями при необходимости).

2. **Перечислите наиболее интересные для языкознания определения вероятности**

3. **Является ли *количественный анализ* альтернативой (антонимом) *качественному анализу* и когда исследование языка и речи достигает наибольшей глубины?**

4. Что характеризует название «*квантитативная лингвистика*»?

5. Попробуйте дать определение понятию «**структурно-вероятностная модель языка**»

Вероятность элементарного лингвистического события

Для *языкознания* наибольший интерес представляют **3** определения вероятности: «***субъективное***», «***классическое***» и «***статистическое***» определения.

Субъективное определение вероятности и его использование в лингвистике

Человек, оценивая вероятность наступления события **A**, опирается на все свои знания (тезаурус **Θ**) относительно тех возможностей, которые могут способствовать или не благоприятствовать осуществлению события **A**.

Эта вероятность может быть представлена как **P(A, Θ)**, т. е. как вероятность события **A** при тезаурусе данного человека **Θ**.

Если два человека имеют относительно события **A** одинаковый тезаурус **Θ**, то значения вероятностей события **A** для этих людей будут одни и те же. Однако такая ситуация встречается редко. Чаще вероятность одного и того же события оценивается разными людьми, исходя из разных величин **Θ** и **Θ′.** Даже у одного и того же познающего субъекта со временем величина **Θ** изменяется и превращается в **Θ′**, следовательно, и его оценки вероятности события *A* в разные периоды его жизни являются различными: **P(A, Θ) ≠ P(A, Θ′)**.

Классическое определение вероятности

Если результаты испытания можно представить в виде полной системы **n** равновозможных и попарно несовместимых событий и если случайное событие появляется только в **m** случаях, то вероятность события **A** равна **P(A) = m(A)/n**, т.е. равна отношению количества случаев, благоприятствующих данному событию, к общему числу всех случаев.

Статистическое определение вероятности

К опытам, которые не могут быть исследованы на основе системы событий, применяется ***статистическое*** определение вероятности. Введем некоторые понятия.

Пусть произведена серия из **N** испытаний, в каждом из которых могло произойти или не произойти событие **A**. *Абсолютной частотой* (или *частотой)* **F** назовем число появлений события **A**, а *относительной частотой* (или *частостью)* **f(A)** – отношение абсолютной частоты к общему числу испытаний: **f(A) = F / N**.

При небольшом числе опытов частоты носят случайный характер и могут изменяться от одной группы событий к другой. Но при большом числе испытаний **N** относительная частота **f(A)** обнаруживает все большую устойчивость.

Экспериментальными (приближёнными) значениями вероятности являются относительные частоты ***f(A)*** интересующего нас события ***A*** в определенных сериях **N** испытаний.

Определенную таким образом вероятность случайного события **A** и называют ***статистической вероятностью***.

Общий принцип, называемый «законом больших чисел» даёт понимание вероятности как предела относительной частоты при неограниченном возрастании числа испытаний, т.е. **P(A) = lim f(A)** при **N** $\rightarrow \infty$**.**

Узловым вопросом всех лингвостатистических исследований является выяснение того, насколько далеко отклоняются экспериментальные частоты **f(A)** от вероятности **P(A)**. Не имея обычно возможности обследовать всю совокупность текстов, мы вынуждены обследовать лишь определенную выборку ➔ необходимо решать вопрос об объёме выборки **N**.

Количественный и описательный анализ

Существует два способа выражения информации об объективной реальности – ***описательный*** и ***количественный***, которые сами по себе могут характеризовать лишь видимые черты и свойства исследуемых объектов, но не их внутреннюю, чаще всего скрытую суть. Эта суть, раскрывается в результате сущностно-содержательного, ***качественного анализа*** на основе описательной или количественной информации.

Таким образом, описательный и количественный анализ характеризуют ***явление***, а качественный – ***сущность***. Так как явление и сущность органически взаимосвязаны, то неразрывно взаимосвязаны описательный и количественный анализ, с одной стороны, и качественный – с другой.

В научном исследовании имеет место либо сущностно-описательный, либо сущностно-количественный анализ. Чистого качественного или описательного, или количественного анализа не существует. Может быть, лишь их единство, при этом в действительности количественный анализ является альтернативой не качественному, а описательному анализу.

Объективной основой для широкого распространения количественных методов является потребность в установлении количественной меры явлений и процессов. **Исследование достигает наибольшей глубины тогда, когда установлена количественная мера соответствующего качества.**

Количественный и описательный анализ

Количественный анализ – это выявление и формирование системы численных характеристик изучаемых объектов, явлений и процессов, которые будут подвергнуты определенной математической обработке.

Количественные методы анализа более трудоемки и сложны в применении и понимании, чем *описательные*, но являются более мощными, ибо создают основу для глубокого раскрытия *качественной* сущности.

Количественные методы анализа имеют свои минусы. Главный из них состоит в том, что широко применяемое усреднение данных и отвлечение от частных черт и свойств исследуемых объектов сопряжено с потерей информации и может привести к чрезмерному абстрагированию.

Таким образом, и описательные, и количественные методы представляют собой вполне правомерные и равнонеобходимые формы исследований явлений и процессов. Сильная сторона описательных методов – их конкретность и образность, а количественных – глубина и точность. Эти методы не противостоят один другому, а дополняют друг друга. Поэтому неправомерна абсолютизация того или иного из них и их противопоставление. В практике научных исследований наибольший успех достигается тогда, когда оба эти метода гармонично сочетаются.

Таблица 4

Грамматические категории	Частота				
	общая	*по жанрам*			
		газ.-журн.	драм.	н.-публ.	худ. пр.
1. Существительное	26,65	32,77	20,4	31,03	23,44
2. Глагол	17,12	14,5	20,88	13,50	18,96
3. Прилагательное	9,37	11,97	6,24	12,46	7,37
4. Наречие	8,096	6,95	9,01	7,26	8,98
5. Числительное	1,17	1,55	1,13	1,026	0,9969
6. Местоимение	13,29	10,01	16,18	11,55	14,94
7. Союз	7,39	6,57	6,81	7,61	8,56
8. Предлог	11,1	11,47	11,18	11,23	10,54
9. Частица	1,002	0,57	1,60	0,67	1,078
10. Причастие	0,979	1,05	0,523	1,36	1,05
11. Субстантив. причастие	0,053	0,08	0,06	0,032	0,044
12. Субстантив. прилагат.	0,4457	0,55	0,42	0,518	0,301
13. Омонимы (типа сущ./глаг.)	0,035	0,023	0,029	0,080	0,0489
14. Остальные	3,295	1,92	5,53	1,71	3,69
Итого	*99,996*	*99,983*	*99,992*	*100,036*	*99,999*

***Дом. задание*: заполнить таблицу по НКРЯ + сравнит. анализ**

	Ася		Отрочество		Калина красная		Свой вариант	
Существительное	2 657	19,24%						
Глагол	2 867	20,76%						
Прилагательное								
Наречие (?)								
Числительное								
Местоимение (?)								
Союз								
Предлог								
Частица								
Вв. слово								
Междометие								
Иниц+nonlex	49	0,35%						
Сумма:	13811	100%						
Всего слов	13 811							
Всего предложений	1 093							
Средняя длина предл.	12,64							

Квантитативная (количественная) лингвистика

Название «***квантитативная лингвистика***» достаточно условно, хотя и довольно широко используется в современной научной литературе. Оно характеризует междисциплинарное направление в прикладных исследованиях, в котором в качестве основного инструмента изучения языка и речи используются ***количественные*** или ***статистические*** методы анализа. Иногда **квантитативная лингвистика** противопоставляется **комбинаторной лингвистике**. В последней доминирующую роль занимает «неколичественный» математический аппарат – теория множеств, математическая логика, теория алгоритмов и т. д.

Привлечение методов измерения и подсчета языковых реализаций позволяет существенно модифицировать представление о языковой системе и возможностях ее функционирования ➔ квантитативная лингвистика – важнейший фактор, влияющий на лингвистическую теорию.

Использование статистических методов в языкознании позволяет дополнить структурную модель языка вероятностным компонентом, то есть создать **структурно-вероятностную модель,** обладающую значительным объяснительным потенциалом.

Дома: **А.Н. Баранов, глава 2, § 2, сс. 38-40**

Основные области приложения структурно-вероятностной модели языка

Лингвистический мониторинг функционирования языка. Задача лингвистического мониторинга заключается в выявлении общих особенностей функционирования языковой системы в конкретном типе дискурса (научном, политическом дискурсе, текстах СМИ и т.д.).

Предмет лингвистического мониторинга — феномены естественного языка: типы языковых ошибок, иностранные заимствования, новые слова и значения, новые (креативные, творческие – не конвенциональные) метафоры, тематическое распределение лексики (лексика временных и пространственных отношений, лексика выражения чувств и эмоций, спортивная лексика и т.д.), особенности использования тех или иных грамматических форм, синтаксических конструкций.

Компьютерное моделирование языка и речи – компьютерная лингвистика. Многие программы, связанные с функционированием языка, используют алгоритмы, основывающиеся на данных о частоте употребления фонем, морфем, лексических единиц и синтаксических конструкций. Например, программы автоматической коррекции орфографии содержат словари, как правило, только наиболее частотных лексем. Аналогичные словари используются в программах автоматического распознавания письменного текста и речи (типа Fine Reader). Абсолютная частота появления лексем (особенно терминологической лексики) используется в системах автоматического аннотирования и реферирования.

Дома: А.Н. Баранов, глава 2, § 2, сс. 40-43

Основные области приложения структурно-вероятностной модели языка

Дешифровка кодированного текста. В процессе дешифровки также могут использоваться данные о частоте употребления графем, морфем и слов, а также их взаимном расположении. Уже разработаны продуктивные алгоритмы дешифровки, основанные на частоте и дистрибуции элементов кодированного текста.

Авторизация/атрибуция текста. Проблема авторизации текста относится к числу классических проблем филологического исследования. Однако чисто филологическое направление авторизации не позволяет построить объективные операциональные критерии анализа и атрибуции текста. Авторизация включает как литературную, так и лингвистическую составляющую. Часто она рассматривается в рамках «количественной стилистики» – стилеметрии. Суть – в использовании количественных, статистических методов анализа текста. Пионером в этой области стал Н.А. Морозов, опубликовавший в 1915 г. работу «Лингвистические спектры. Средство для отличия плагиатов от истинных произведений того или другого известного автора. Стилеметрический этюд».

Дома: **А.Н. Баранов, глава 2, § 2, сс. 40-43**