



Квантитативная лингвистика. Лингвостатистический анализ текста

План

- ▶ Понятие квантитативной лингвистики (КЛ)
- ▶ Методы КЛ
- ▶ Лингвостатистический анализ
- ▶ Частота, генеральная и выборочная совокупности
- ▶ Практические задания

Квантитативная лингвистика

- ▶ раздел общей лингвистики
- ▶ исследует язык при помощи статистических методов
- ▶ цель — сформулировать законы функционирования языка
- ▶ связывает языкознание, математику, информатику

Исторические факты

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

- ✓ выборка в **один миллион словоупотреблений** из **четырёх жанров** (художественная проза, драматургия, научная публицистика, газетно-журнальные 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.**

Определение авторства

Кто является истинным автором романа «Тихий Дон»?

Ученые взяли тексты, бесспорно принадлежащие М. Шолохову, и тексты донского писателя Ф. Крюкова, которому приписывалось авторство романа, и проанализировали их, выявляя особенности писательской манеры каждого:

- ❖ **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

Методы КЛ

Количественные

- ▶ учитывают и регистрируют частоту фактов/явлений/объектов
- ▶ подсчитывают единицы любого уровня языка

Статические

- ▶ исследуют факты с целью вскрыть закономерности (правила) появления этих фактов при функционировании языка

Лингвостатистический анализ

Что
считать?

- Единицы
ЛА

Зачем
считать?

- Цель -
исследование
совокупности
однородных
языковых единиц

Как
считать?

- Методики

**Единица ЛА
- языковая
единица
любого уровня**

буквы

фонемы

морфемы

словоформы

слова

словосочетания

предложения

текст

печатный знак

Базовые статистические понятия

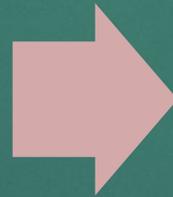
частота

генеральная
совокупность

выборочная
совокупность

Частота

- число появлений
факта/явления в
наблюдаемом
отрезке



Отрезок - любая
среда, в которой
находятся
факты/явления,
поддающиеся счету

Генеральная совокупность

ВСЯ
СОВОКУПНОСТЬ
ОДНОРОДНЫХ
ЯЗЫКОВЫХ
ЕДИНИЦ

Выборочная совокупность (выборка)

часть
генеральной
совокупности,
объединенная
общим
признаком

Виды генеральной совокупности

Совокупность текстов
одинакового жанра,
одного автора,
заданного
временного
интервала и т.д.

Совокупность
языковых единиц
любого уровня

Выборочные совокупности (выборки) – по объему

малые (менее 30 единиц)

средние (30-100)

большие (более 100)

Выборочные совокупности (выборки) – по способу выборки

- случайная выборка – простой случайный отбор
- механическая выборка – вид случайной, упорядочена по к.-л. признаку
- и др.

Практическое задание № 1

Взять в читальном зале (или посмотреть прикрепленные страницы) учебник

А. В. Гребенщиковой «Квантитативная лингвистика и новые информационные технологии»

Стр. 34. Задание 1.

Скачать программу wordstat и обязательно прочитать инструкцию по ссылке

<https://www.bestfree.ru/soft/obraz/word-count.php>

Практическое задание № 2

А. В. Гребенщикова. Квантитативная лингвистика и новые информационные технологии

Стр. 35. Задание 2, п. 1-4.

Результаты лингвостатистического анализа представить в виде графика (п. 4), принести на следующий семинар, в электронном виде.

Уметь прокомментировать процесс, методику и результаты проведенного исследования – устно.

Список литературы

- ▶ Гребенщикова А. В. Квантитативная лингвистика и новые информационные технологии. 2013.
- ▶ Зубов А. В., Зубова И.А. Информационные технологии в лингвистике.
- ▶ Статистика слов
<https://www.bestfree.ru/soft/obraz/word-count.php>