Статистическая модель негативного биномиального распределения в анализе лингвистических данных

Вьюшкова Евгения Александровна, гр. 522

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н., доц. Алексеева Н.П.
Рецензент: к.ф.-м.н., доц. Коробейников А.И.

Санкт-Петербург
2013

Описание данных

- Структура данных: текст с N искусственно созданными главами или в авторской редакции
  - Главы имеют смысл повторностей.
  - Случайное число (X) вхождений слова a_j в главу.
- Модель НБР: $X \sim \beta\_(j|k,p)$ [Alexeyeva, Sotov, 2013]
- Гамма-пуассоновская схема, $j = 1, 2, \ldots$ [Bart, 2003]

$$\mathbf{P}\{X = j\} = \frac{\Gamma(k+j)p^k(1-p)^j}{\Gamma(j+1)\Gamma(k)} = \int_0^\infty P(j|\lambda q)\gamma(\lambda p|k)d(\lambda p) \tag{1}$$

$$P(j|\lambda) = \exp(-\lambda)\frac{\lambda^j}{j!} \text{ и } \gamma(\lambda q|k) = \frac{\lambda^{k-1}}{\Gamma(k)}q^{k-1}\exp(-\lambda q)$$

- Интерпретация параметров в лингвистике
  - k — число потерянных слов или индекс синонимии;
  - p — вероятность замены слова, гибкость.

Математические проблемы

- Формирование глав и структурирование выборки (R);
- *Оценка параметров словоупотребления* k *и* p (C++, R);
- Визуализация результатов (C++);
- *Проверка гипотез согласия* (R);
- Выделение “слов” из текста без пробелов (C++).

Два способа оценивания параметров словоупотребления

1. В C++ оценка максимального правдоподобия параметра $\hat{k}$ является решением уравнения

$$\psi(x_i + k - 1) - \psi(k) + \ln(p) = 0 \tag{1}$$

$$\hat{p} = \frac{\hat{k}}{\hat{k} + \bar{x}},$$

где $\psi(x) = \ln'(\Gamma(x))$.

2. В R оценки $\hat{k}$, $\hat{p}$ получены при помощи встроенного метода максимизации функции правдоподобия (mle2).

Проверка адекватности модели

Критерии согласия для дискретных распределений [Choulakian, 1994]

Рассматриваемые критерии

1. Критерий χ^2;
2. Модифицированный критерий Колмогорова-Смирнова;
3. Модифицированное семейство критериев Крамера-фон Мизеса: W^2, U^2, A^2.

Вычисление p-value:

- Метод Монте-Карло;
- Асимптотические критерии. W^2, U^2, A^2: метод Бокса (взвешенное χ^2 распределение) [Imhof, 1961].

Статистики критериев Крамера-фон Мизеса

Обозначения:

- o_i — количество появлений события i, $Np_i = e_i$ — ожидаемое число появлений события i, N — число независимых наблюдений;
- $S_j = \sum_{i=1}^{j} o_i$, $T_j = \sum_{i=1}^{j} e_i$, $Z_j = S_j - T_j$, $\bar{Z} = \sum_{j=1}^{k} Z_j p_j$;
- $H_j = \sum_{i=1}^{j} p_i$ — теоретическая функция распределения.

Статистики:

$$W^2 = N^{-1} \sum_{j=1}^{k} Z_j^2 p_j, \tag{1}$$

$$U^2 = N^{-1} \sum_{j=1}^{k} \left(Z_j^2 - \bar{Z}\right)^2 p_j, \tag{2}$$

$$A^2 = N^{-1} \sum_{j=1}^{k} \frac{Z_j^2 p_j}{H_j \left(1 - H_j\right)}. \tag{3}$$

Мощность критерия Крамера-фон Мизеса

1. Применимость критерия $(n = 20)$.
2. Проверяемые гипотезы
   - H_0 : НБР
   - H_1 : Пуассон

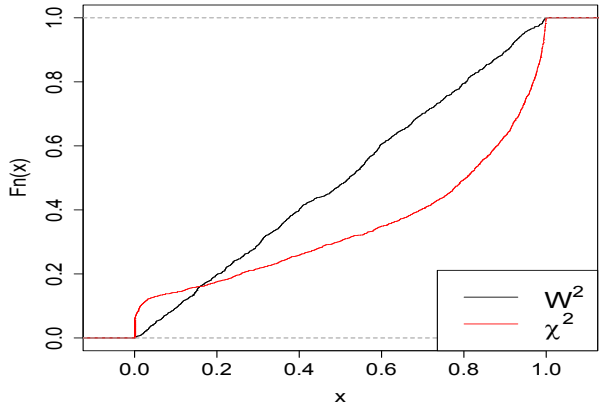
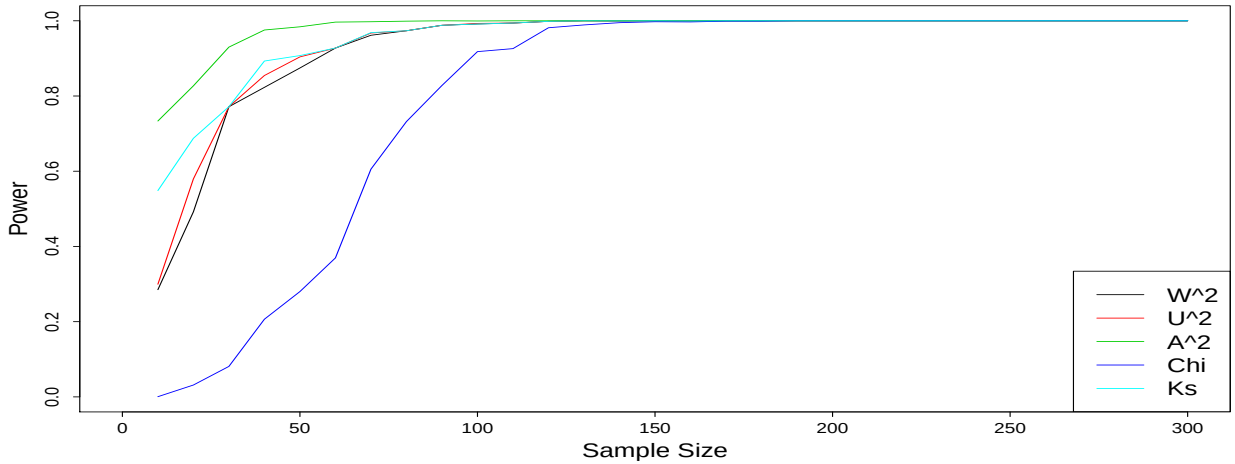


Рис. 1: *Ф. р. p-value для* W^2 *и* χ^2



Визуализация семантических областей

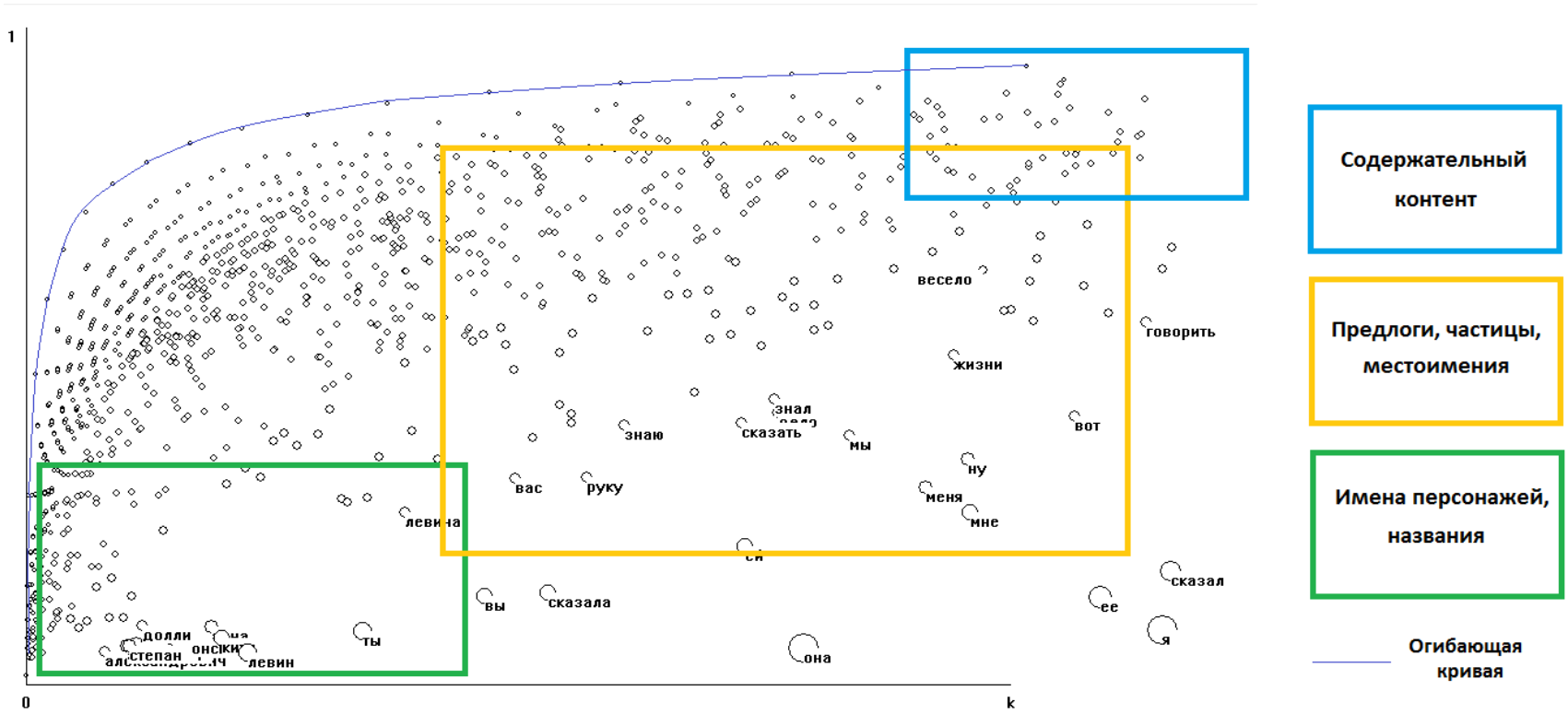


Рис. 1: *Двухмерная диаграмма оценок параметров для романа Л.Н.Толстого "Анна Каренина"*

Огибающая кривая параметрической области

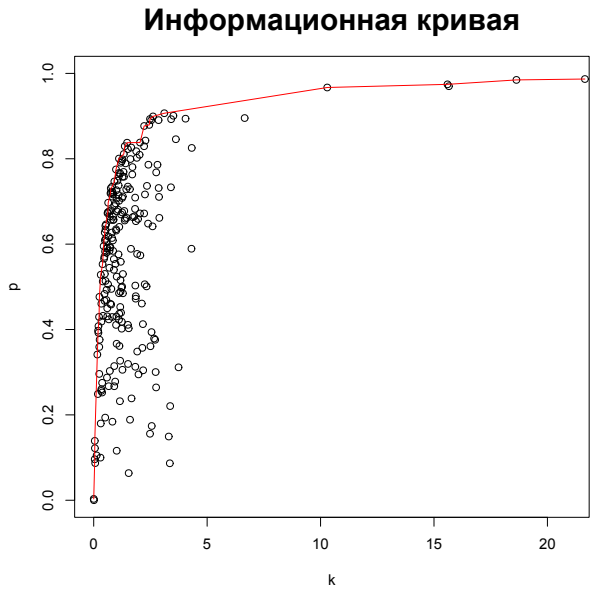
Определение

Пусть (k, p) параметрическая область оценок параметров НБР данного текста. Множество точек таких, что:

$$f(k) = \max(p|k) \text{ или } f(p)^{-1} = \min(k|p)$$

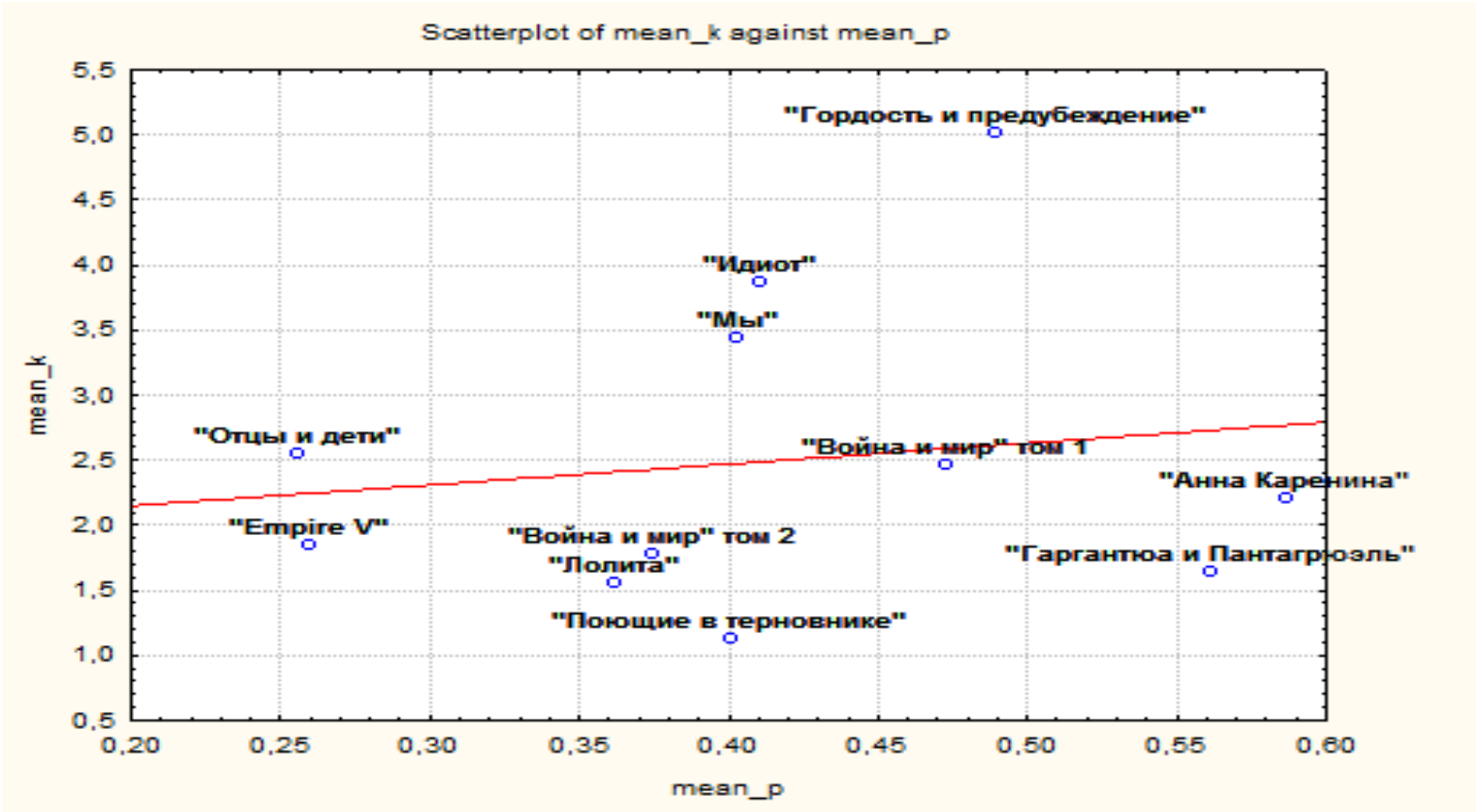
называется *огибающей кривой*.

- Аналитическая аппроксимация:
  $p = 1 - e^{-(a\ln(k)+b)^2}$
- Интерпретация:
  минимальные потери при заданной вероятности замены слова (коммутационный след)



Пример из лингвистики

Рис. 1: *Двухмерная диаграмма средних оценок параметров НБР для различных книг.*

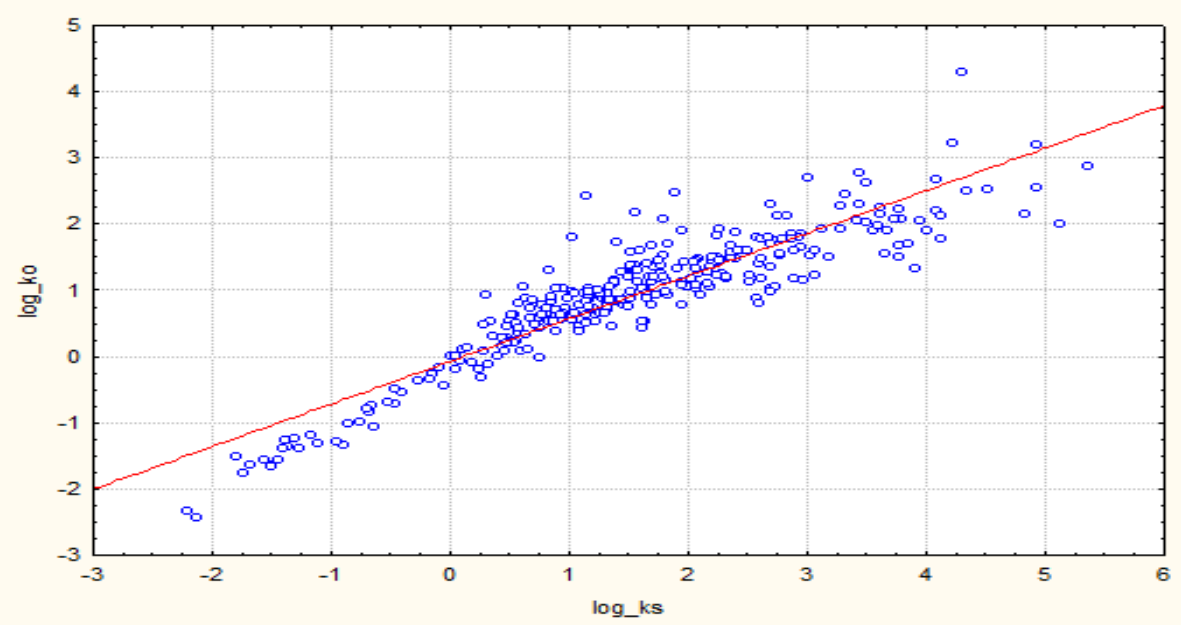


Проблема соизмеримости текстов.

Влияние на оценки способа разбиения текста на главы

Обнаружены значимые корреляции средних параметров при авторском разбиении текста на 60 глав (k_o, p_o) и соответствующим равномерным разбиением (k_s, p_s).

Корреляция	$\log(k_s)$	p_s
$\log(k_o)$	0.764	
p_o		0.615



Влияние на параметры НБР масштаба равномерного разбиения

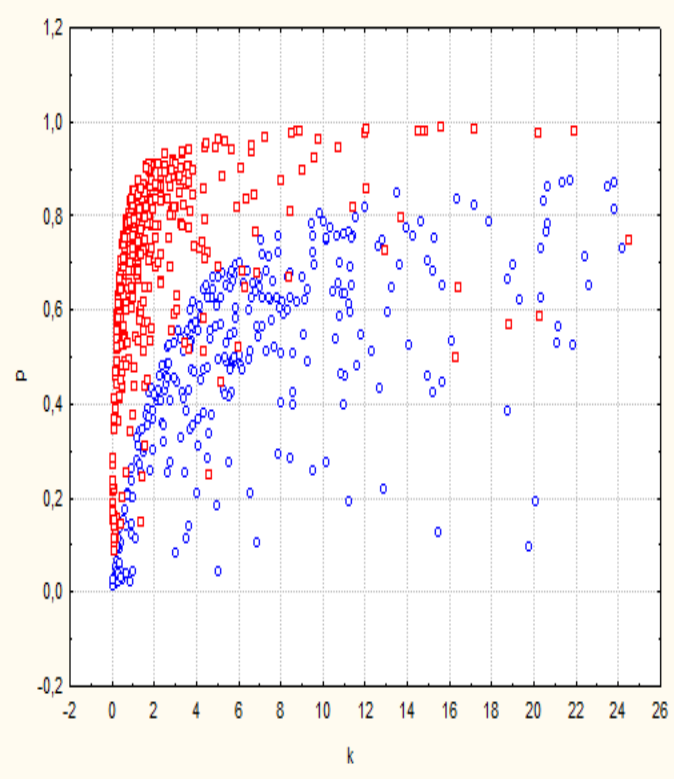


Рис. 1: *Оценки параметров словоупотребления при разном масштабе* $(N = 30,\ N = 400)$

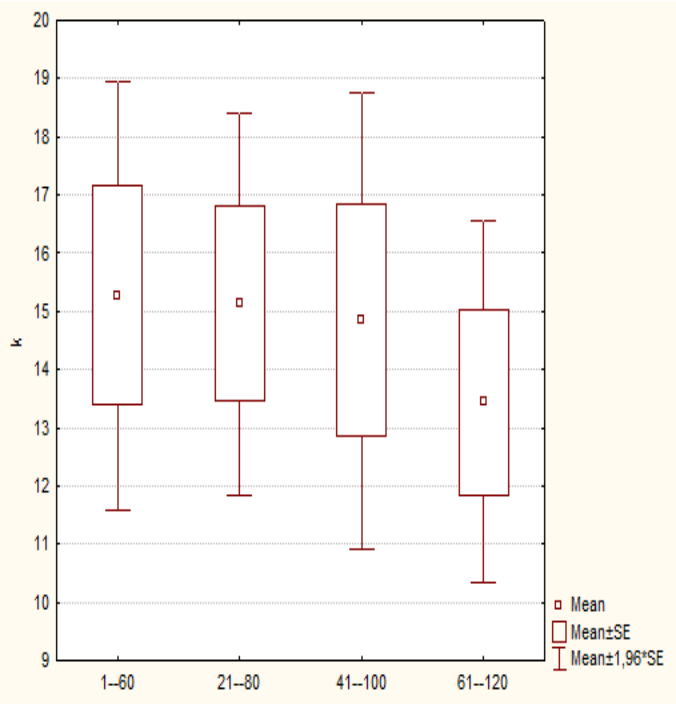
Эксперимент:

- текст разбивается на $N = 30, 60, \ldots$ частей;
- для каждого разбиения вычисляется набор слов с оценкой параметров k, p и проверкой согласия;
- линейная зависимость параметров от масштаба разбиения

$$\ln(k) = a_1 + b_1 \ln(N); \quad (1)$$

$$p = a_2 + b_2 N. \quad (2)$$

Исследование зависимости параметра k от выбора глав



Эксперимент:
N — число глав

- целый текст $N = 120$;
- рассматриваются подвыборки $N = 60$;
- для каждой подвыборки вычисляется набор слов с оценкой параметров k, p и проверкой согласия;
- выделяются одинаковые слова во всех подвыборках.

- Изменение оценок параметра в разных частях текста не значимо.

Применение модели НБР: сравнение 5 переводов

Авторы переводов трилогии “The Lord of the Rings” J. R. R. Tolkien:

1. Н. В. Григорьева, В. И. Грушецкий (1991);
2. В. А. Маторина (1991);
3. М. Каменкович, В. Каррика (1990-е);
4. В. Муравьёв, А. Кистяковский (1982);
5. А. Грузберг (1977–1978).

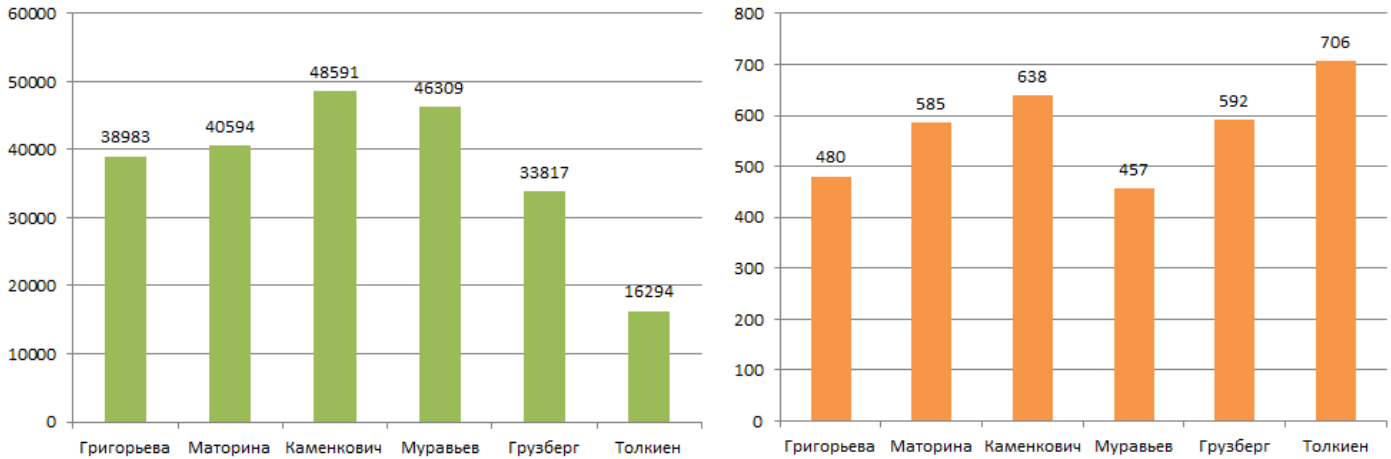


Рис. 1: *Объемы словаря и исследуемые слова*

Дисперсионный анализ сравнения параметров словоупотребления

Несмотря на одинаковую структуру разбиения на главы оценки параметров в разных переводах отличаются значимо.

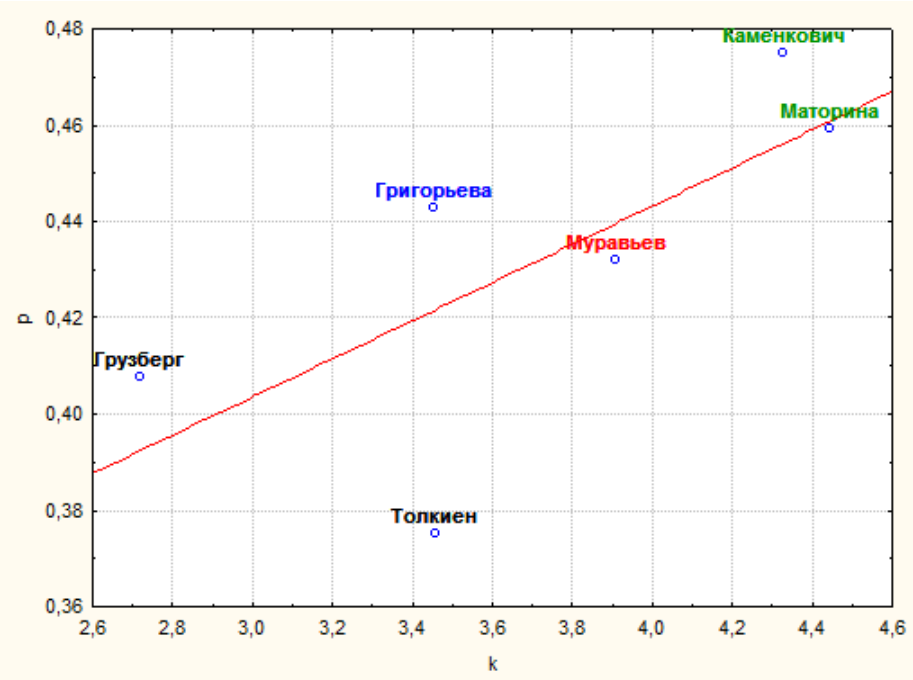


Таблица 1: *Ответы по предпочтению переводов (2013)*

Перевод	"+"	"-"
Григорьева	13	8
Грузберг	5	7
Каменкович	13	0
Маторина	8	1
Муравьев	1	24

Метод множественных сравнений — адекватность перевода Грузберга оригиналу.

Метод множественных сравнений Шеффе

Сравнение параметров НБР в переводах Грузберга и Каменкович

Значимые различия

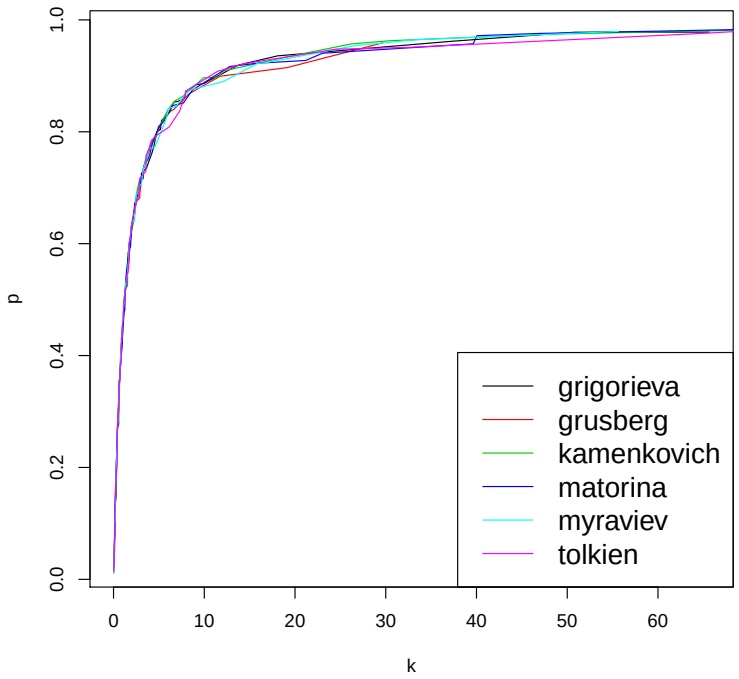
существительные N числительные Num	глаголы V	наречия AV прилагательные AJ

Незначимые различия

имена собственные, названия PN; союзы CJ частицы Part; предлоги PR местоимения NPP; собственные местоимения PP

Инвариантность информационной кривой

- Информационная кривая инвариантна относительно перевода.
- Коммутационный след изменяется не значимо при рассмотрении разных разбиений на равные части одного текста.



Результаты дипломной работы

- Проведен анализ мощности критериев в рамках поставленной задачи, выбраны критерии наилучшим образом удовлетворяющие условиям задачи.
- Разработано ПО, позволяющее оценивать параметры и проверять гипотезы согласия с НБР.
- Изучена применимость разработанного ПО, проведен анализ результатов.