УДК 519.766.48

Статистические закономерности в законе Хипса в художественной литературе на различных языках

Д. И. Смирнова

Московский физико-технический институт (государственный университет)

Институт системного анализа РАН

Закон Хипса является важным эмпирическим статистическим законом, описывающим свойства естественного языка. Обозначим текст на некотором естественном языке T. Рассмотрим последовательную выборку T_n из n слов текста T, тогда закон Хипса будет формулироваться в виде:

$$Q(T_n) = cn^{\alpha}$$

где $Q(x)$ – количество уникальных слов в тексте x, C, α – некоторые коэффициенты. Эмпирически установленный факт, что C, α зависят от многих параметров, таких как язык, на котором написан текст [1], стиль автора, жанр произведения и т.д.

В данной работе произведено исследование по оценке параметров C, αдля художественной литературы в зависимости от автора и языка произведения.

При исследовании зависимости коэффициентов C, α закона Хипса была использована линейная регрессия:

$$\log Q(T_n) = \log C + \alpha \log n$$

при этом показано, что несмотря на то, что метод дает качество оценки ниже, чем метод максимального правдоподобия, увеличение ошибки не существенно, при этом метод наименьших квадратов, использовавшийся для определения коэффициентов закона Хипса, демонстрирует превосходящую на порядок скорость работы, что существенно при анализе больших объемов данных.

Для эксперимента была сформирована выборка, состоящая из 3-6 художественных произведений для каждого автора, суммарно по 30-50 художественных работ (прозы) для каждого из следующих языков: русский, английский, немецкий, французский и итальянский. Для того, чтобы повысить статистическую значимость оценки коэффициентов [2], [3], а также устранить синтаксические особенности некоторых языков (например, склонения в русском языке) была использована лемматизация, предоставленная библиотекой nltk для языка Python.

Целью эксперимента являлась проверка следующих гипотез о постоянстве коэффициентов среди:

1) книг одного автора

2) книг, написанных на одном языке

3) всех художественных произведений

В результате эксперимента все три гипотезы были опровергнуты. При этом была замечена следующая закономерность: коэффициенты $\log C, \alpha$ связаны ярко выраженной линейной зависимостью:

$$\log C \approx \alpha$$

Данный факт невозможно определить теоретически, а значит он отражает эмпирические свойства естественных языков. Предположительно, данная зависимость свидетельствует в пользу стохастических моделей естественного языка.

На рисунках 1-3 представлены графики полученной зависимости коэффициентов $\log C(\alpha)$ для художественной литература на английском, русском и немецком языках.

Рисунок 4 иллюстрирует зависимость $C(\alpha)$ для книг на английском и русском языках.

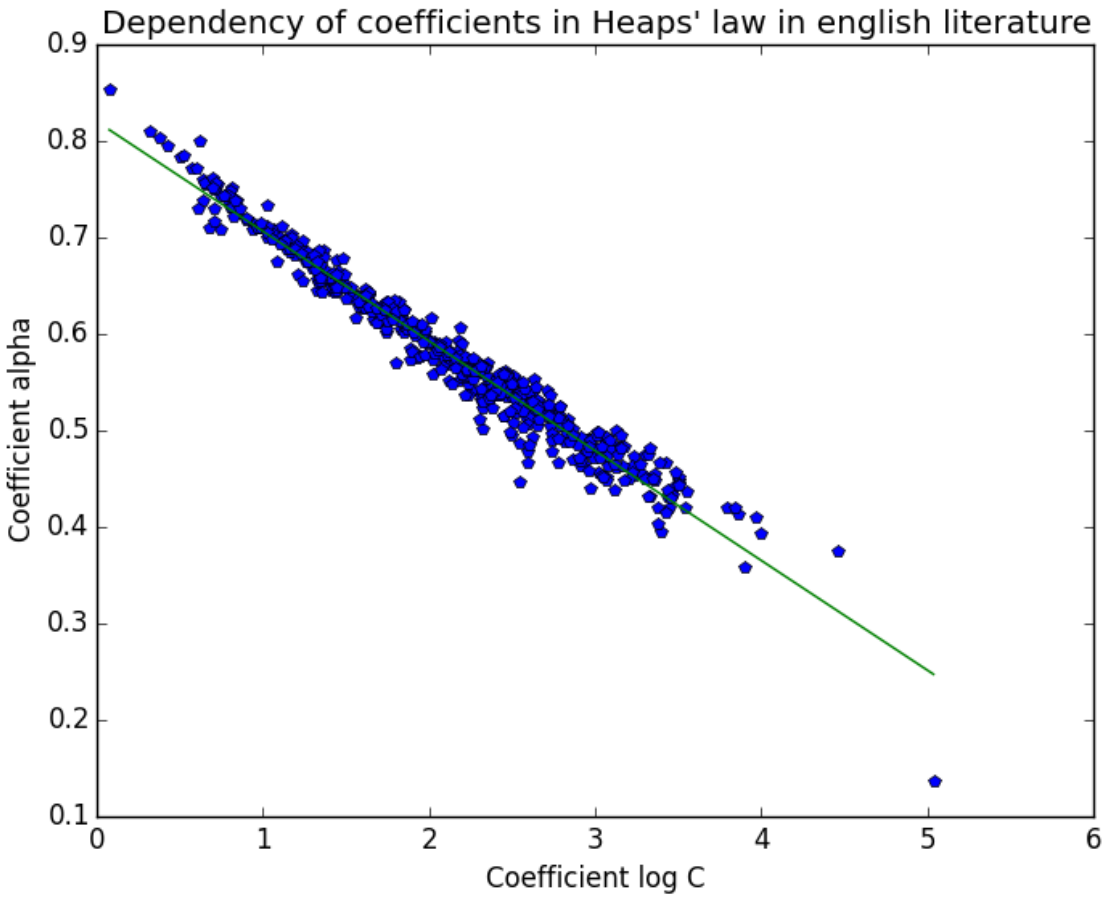

рис. 1 Зависимость коэффициентов $\log C(\alpha)$ для художественной литературы на английском языке

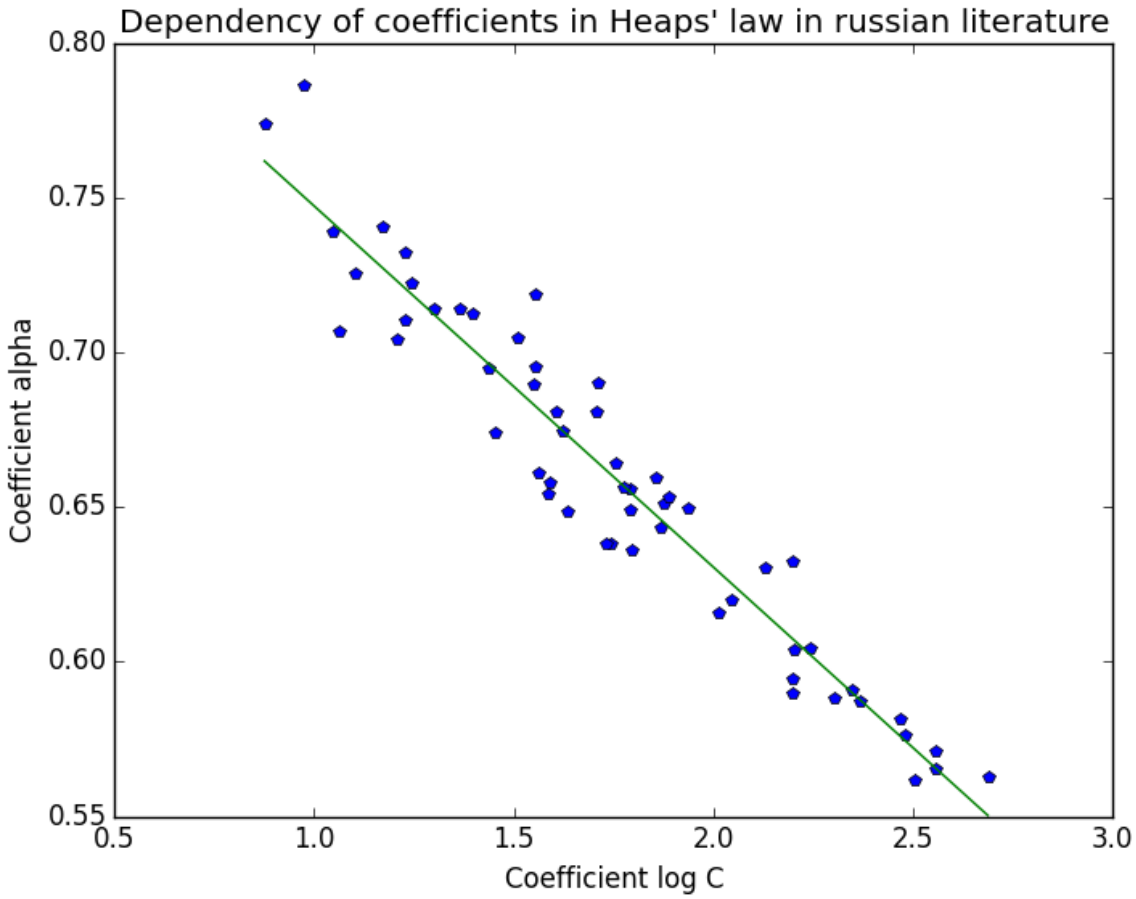

рис. 2 Зависимость коэффициентов $\log C(\alpha)$ для художественной литературы на русском языке

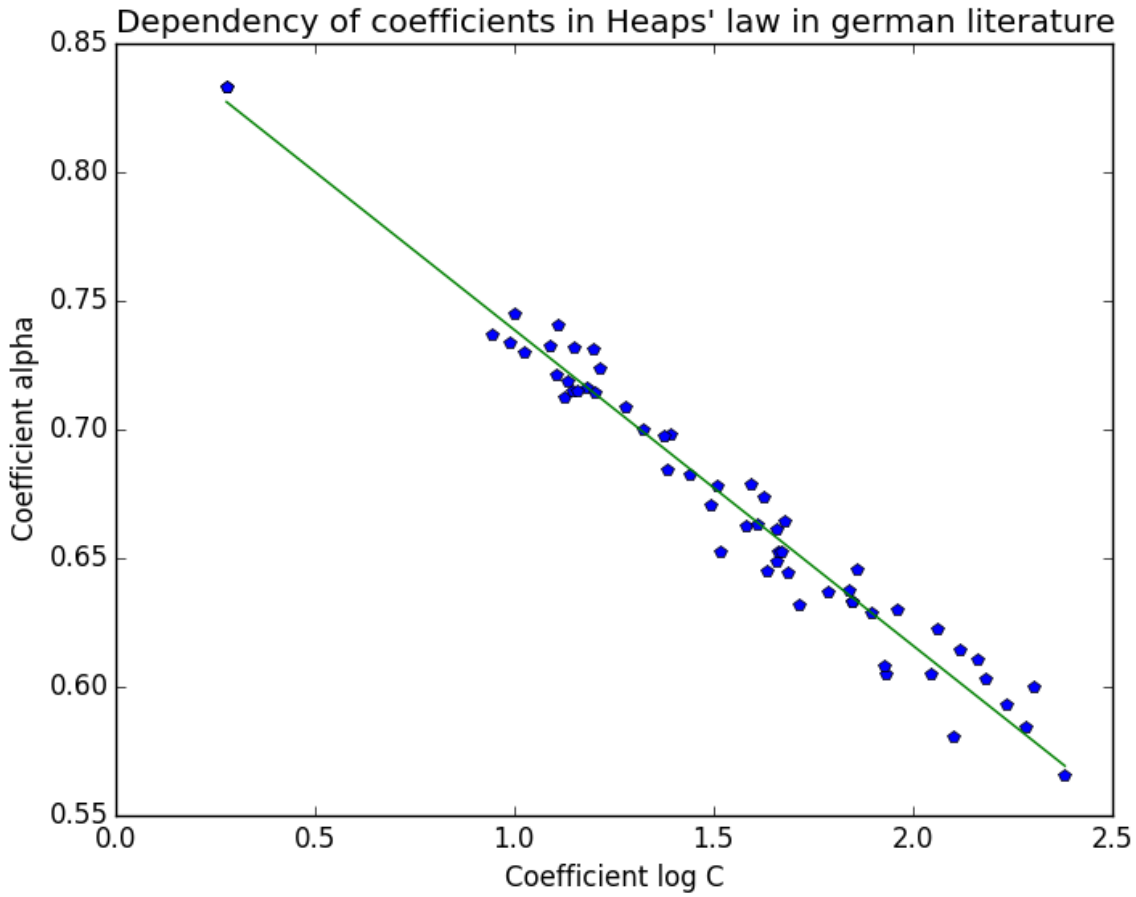

рис. 3 Зависимость коэффициентов $\log C(\alpha)$ для художественной литературы на немецком языке

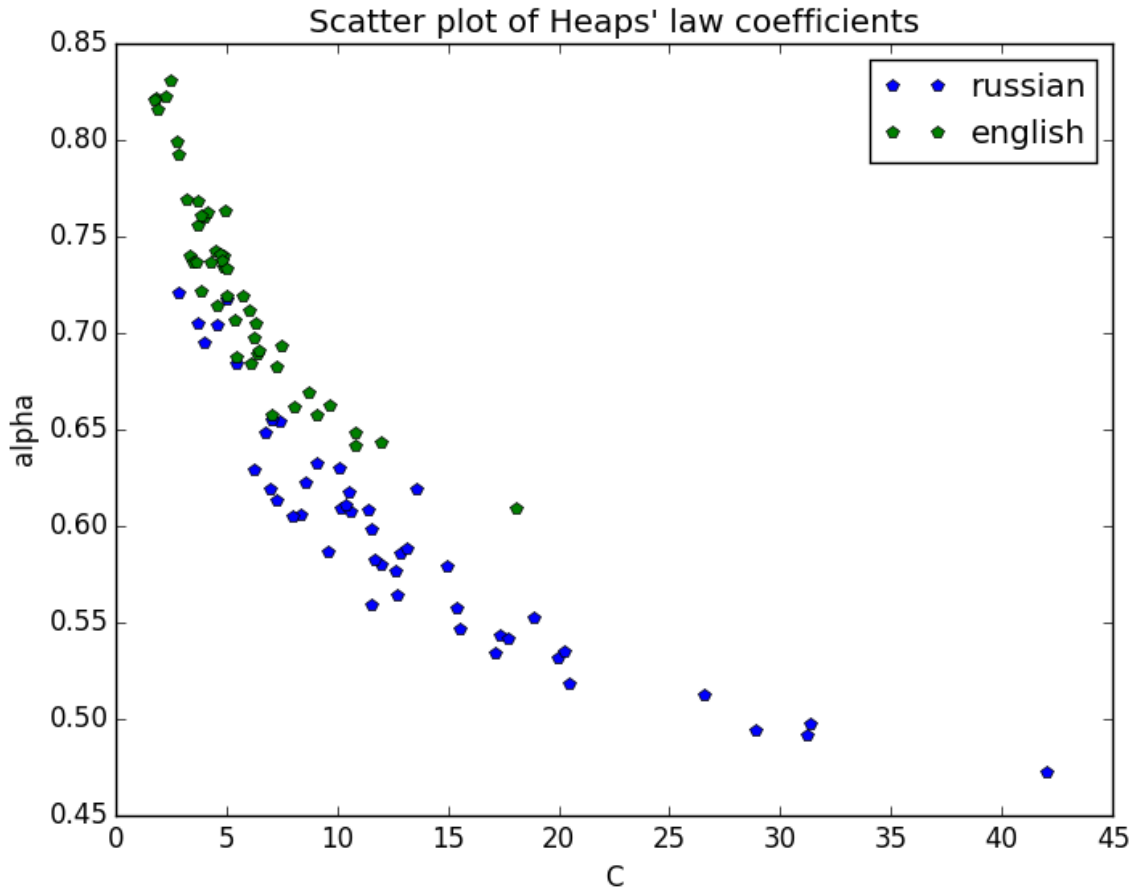

рис.4 Зависимость коэффициентов $C(\alpha)$ для художественной литературы на русском и английском языках

Литература

1. *Gelbukh A., Sidorov G.* Zipf and Heaps Laws' coefficients depend on language //Computational Linguistics and Intelligent Text Processing. – Springer Berlin Heidelberg, 2001. – С. 332-335.
2. *van Leijenhorst D. C., Van der Weide T. P.* A formal derivation of Heaps' Law //Information Sciences. – 2005. – Т. 170. – №. 2. – С. 263-272.
3. *Tudman M.* The Text Vocabulary Size Law. Heaps' Law and Determining Text Vocabulary Size in Croatian Language //Društvena istraživanja. – 2005. – Т. 14. – №. 1-2 (75-76). – С. 227-250.