АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81'32'373[811.111+811.161.1]

А. Ф. Гельбух, Г. О. Сидоров

Коэффициенты законов Ципфа и Хипса для русского и английского языков*

Приводятся экспериментальные данные для английского и русского языков, показывающие, что коэффициенты важных статистических законов естественного языка — закона Ципфа и закона Хипса — зависят от языка. Этот факт имеет как теоретическое, так и практическое значение. С одной стороны, выяснение его причин может пролить свет на природу языка. С другой стороны, например, закон Хипса важен в практических приложениях, например, при разработке полнотекстовых баз данных, для которых он позволяет предсказать размер индексного файла.

ВВЕДЕНИЕ

Интересными статистическими эмпирическими законами, описывающими поведение слов в тексте, являются законы Ципфа и Хипса [1–3].

Напомним, что закон Ципфа состоит в следующем. Посчитаем частоты слов (лемм[1] или словоформ) в тексте. Присвоим словам ранги в соответствии с их частотами, начиная с самых больших значений. Упорядочим ранги от больших к меньшим. Оказывается, что в любом достаточно большом тексте ранги обратно пропорциональны частотам, т. е. можно записать[2]:

$$f_r \approx C/r^z \tag{1}$$

или в логарифмической форме:

$$\log f_r \approx C - z \log r, \tag{2}$$

где f_r — частота в тексте единицы (леммы или словоформы) с рангом r, z — экспоненциальный коэффициент (близкий к 1) и C — константа. В логарифмической шкале график данного распределения близок к прямой под углом –45°.

Другим статистическим эмпирическим законом, описывающим поведение слов в тексте, является закон Хипса. Он гораздо менее известен, чем закон Ципфа, однако не менее значим.

Закон Хипса состоит в том, что количество различных слов (словоформ или лемм) в тексте пропорционально экспоненте его размера:

$$n_i \approx D \cdot i^h \tag{3}$$

или в логарифмической форме:

$$\log n_i \approx D + h \log i, \tag{4}$$

где n_i — это число различных слов (лемм или словоформ), встречающихся перед словом с текущим номером i, h — экспоненциальный коэффициент (между 0 и 1) и D — константа. В логарифмической шкале график данного распределения близок к прямой под углом 45°.

Природа законов типа Ципфа или Хипса неясна. Любопытным фактом является то, что практически любой текст подчиняется подобным эмпирическим законам. С лингвистической точки зрения интересно, что для лемм и для словоформ получаются очень близкие распределения, даже для такого языка с развитой морфологией, как русский. Про закон Ципфа также известно, что многие другие явления, связанные с обыденной жизнью, подчиняются этому закону, например, число жителей в городах.

В принципе, поскольку эти законы свойственны естественному языку, они могут служить для определения языковой природы неизвестных сигналов [4]. Кроме того, на практике полезно знать значения коэффициентов этих законов для разных языков при разработке полнотекстовых баз данных, потому, что это позволяет определить необходимый размер индекса в зависимости от размера базы документов.

В данной работе мы показываем, что коэффициент z для закона Ципфа и коэффициент h для закона Хипса существенно зависят от языка. Мы

*Эта работа была выполнена при частичной поддержке CONACyT, REDII и SNI, Мексика. Мы благодарим проф. R. Baeza-Yates, проф. E. Atwell и проф. И. Большакова за полезное обсуждение. (The work done under partial support of CONACyT, REDII, and SNI, Mexico. We thank Prof. R. Baeza-Yates, Prof. E. Atwell, and Prof. I. Bolshakov for useful discussion.)

[1] Напомним, что словоформа — это грамматическая форма слова, непосредственно встречающаяся в тексте; лемма — это нормализованная форма слова, обычно дающаяся в качестве заглавной в словаре, например, для словоформ *стола*, *столу* и т. д., лемма будет *стол*.

[2] Здесь мы игнорируем поправки Мандельброта к закону Ципфа [1], поскольку они касаются только крайних значений и не влияют на дальнейшее обсуждение.

рассматриваем русский и английский языки. Эксперименты проводились на достаточно большом объеме текстов. Тексты принадлежат к разным жанрам художественной литературы.

ЭКСПЕРИМЕНТАЛЬНЫЕ ДАННЫЕ

Мы обработали по 39 литературных произведений для русского и английского языков, выбранных случайным образом из произведений различных жанров. При этом дополнительным ограничением было требование, чтобы размер текста был не менее 100 КВ (минимум 10 тыс. слов); всего не менее 2.5 млн словоформ (24.8 МВ) для английского и 2 млн словоформ (20.2 МВ) для русского.

Наши эксперименты проводились как с леммами, так и со словоформами. Лемматизация проводилась автоматически. Омонимия при этом не разрешалась, т. е. в результирующий файл заносились все возможные леммы. Результаты вычислений коэффициентов для лемм и словоформ оказались зависящими от языка.

Мы пользовались программой, строящей на экране графики распределений по закону Ципфа и Хипса, где для закона Ципфа использовались точки:

$$x_r = \log r, \quad y_i = \log f_r \tag{5}$$

и для закона Хипса точки:

$$x_i = \log i, \quad y_i = \log n_i. \tag{6}$$

Примеры таких графиков приведены на рис. 1 и на рис. 2. По оси X отложены ранги, а по оси Y — частоты. При построении графиков распределений для текстов на разных языках различия в угле наклона прямой (графика распределения) оказывались визуально заметными, что подтверждается ниже точными вычислениями.

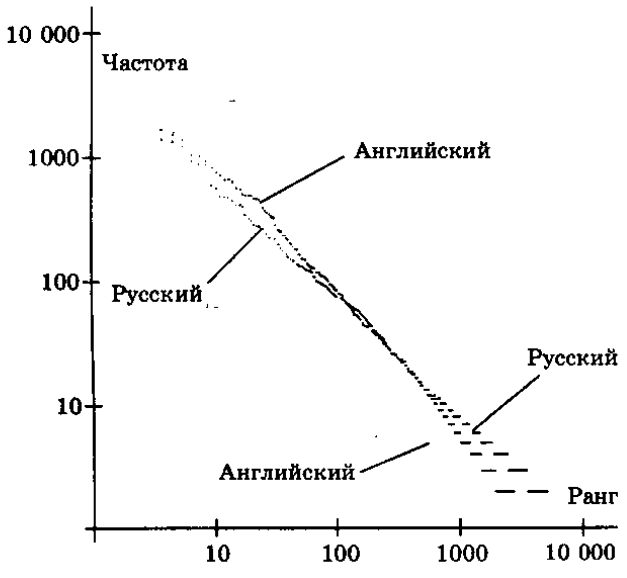

Рис. 1. Графики распределения по закону Ципфа для английского текста № 10 (автор *A. Hope*) и русского текста № 13 (автор *В. Пелевин*) (логарифмическая шкала)

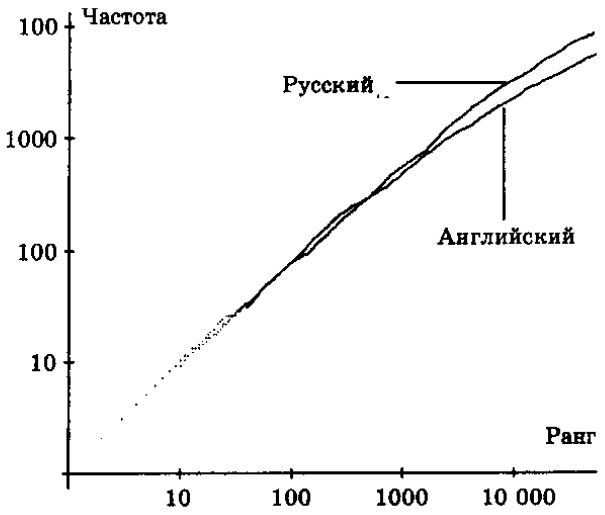

Рис. 2. Графики распределения по закону Хипса для английского текста № 10 (автор *A. Hope*), и русского текста № 20 (автор *Марина Наумова*) (логарифмическая шкала)

Для точных вычислений был использован метод линейной регрессии. График аппроксимировался прямой $y = ax + b$, где a и b соответствуют z и C в законе Ципфа и h и D в законе Хипса. Поскольку плотность точек (x_i, y_i) возрастает экспоненциально с ростом x_i, мы умножаем все значения на c^{-x_i}, что дает равномерный учет влияния каждого отрезка графика на расстояние, где c есть основание логарифма, используемое для вычисления (x_i, y_i), в нашем случае равное 10. Для вычисления a и b мы используем следующие формулы:

$$b = \frac{\sum_i \frac{x_i}{c^{x_i}} \sum_i \frac{x_i y_i}{c^{x_i}} - \sum_i \frac{x_i^2}{c^{x_i}} \sum_i \frac{y_i}{c^{x_i}}}{\left(\sum_i \frac{x_i}{c^{x_i}}\right)^2 - \sum_i \frac{x_i^2}{c^{x_i}} \sum_i \frac{1}{c^{x_i}}},$$

$$a = \frac{\sum_i \frac{y_i}{c^{x_i}} - b \sum_i \frac{1}{c^{x_i}}}{\sum_i \frac{x_i}{c^{x_i}}}, \tag{7}$$

Визуальные наблюдения подтвердили, что данная формула лучше аппроксимирует графики, чем обычная формула линейной регрессии. Результаты приводятся в *Приложении 1*, список обработанных текстов дан в *Приложении 2*. Приводятся только значения z и h, потому что параметры C и D не существенны. В качестве меры точности мы пользовались общепринятым правилом 3σ, где σ — стандартное отклонение. Для английского языка z=0.97±0.06 на словоформах и z=0.98±0.07 на леммах. Для русского z=0.89±0.07 на словоформах и z=0.91±0.09 на леммах. Разница между русским и английским языками для закона Ципфа на словоформах составляет 8.3% и на леммах 9.9%. Для закона Хипса, для английского h=0.79±0.05 на леммах и h=0.80±0.05 на словоформах. Для русского h=0.84±0.06 на леммах и h=0.89±0.05 на словоформах. Разница между русским и английским языками составляет 5.9% на леммах и 5.6% на словоформах. Как для закона Ципфа, так и для закона Хипса, уровень значимости разницы значительно больше 1%.

ВОЗМОЖНОЕ ОБЪЯСНЕНИЕ ЯВЛЕНИЯ

Первое, и самое естественное, объяснение различий в значениях коэффициентов, это попытка обратиться к грамматическому строю языка. В русском языке есть развитая морфология (т. е. — это язык синтетический), а в английском ее нет (т. е. — это язык аналитический). Другое соображение — число различных словоформ в русском тексте должно быть больше, чем в английском за счет многообразия морфологических вариантов. Однако эта идея опровергается тем, что коэффициенты для лемм и словоформ в русском языке оказываются очень близкими.

Мы предполагаем, что различия связаны с понятием "лексического богатства" языка. Известно, например, что язык часто требует выразить смысл, излишний в какой-то ситуации, просто потому, что в нем есть соответствующие слова. Например, в английском нужно сказать *The table was near the wall.* — букв. *Стол находился около стены.* В русском же естественно и идиоматично сказать *Стол стоял у стены*, выбрав из нескольких возможных слов, характеризующих способ нахождения (*лежать, сидеть*, и пр.). Т. е. в данном случае количество возможных слов, зависящих от субъекта, больше в русском языке. Данный вопрос, безусловно, требует дальнейших исследований.

ВЫВОДЫ

Мы показали, что экспоненциальный коэффициент законов Ципфа и Хипса существенно зависит от языка, сравнивая коэффициенты для 39 различных достаточно больших художественных текстов на русском и английском языках. Для вычисления коэффициентов был использован метод линейной регрессии.

В дальнейшем мы планируем уточнить понятие "лексического богатства", например, сравнивая оригинал и перевод. Кроме того, интересно посчитать коэффициенты для законов Ципфа и Хипса с учетом частей речи. Кроме того, нам хотелось бы включить в рассмотрение тексты на различных языках, но этому препятствует отсутствие доступных массивов текстов.

СПИСОК ЛИТЕРАТУРЫ

1. Manning C. D. and Shutze H. Foundations of statistical natural language processing. Cambridge, MA, The MIT press, 1999, 680 p.

2. Zipf G. K. Human behavior and the principle of least effort. Cambridge, MA, Addison-Wesley, 1949.

3. Лингвистический Энциклопедический Словарь, М.: Советская Энциклопедия, 1991.

4. Elliott J., Atwell E. and Whyte B. Language identification in unknown signals. In COLING'2000, ACL and Morgan Kaufmann Publishers, 2000, pp. 1021–1026.

ПРИЛОЖЕНИЕ 1

ЗНАЧЕНИЯ КОЭФФИЦИЕНТОВ ДЛЯ ЗАКОНОВ ЦИПФА И ХИПСА

В нижеследующих таблицах приводятся значения коэффициентов законов Хипса и Ципфа для английского и русского языков. Номер текста в таблицах соответствует номеру текста в *Приложении 2*, где и даются данные о тексте. Таблицы упорядочены по значению коэффициента закона Ципфа для словоформ.

Таблица 1

Значения коэффициентов законов Ципфа и Хипса для английского языка

Текст	Жанр	Ципф (словоформы)	Ципф (леммы)	Хипс (леммы)	Хипс (словоформы)
1	детектив	1.037639	1.034344	0.759330	0.773328
2	приключения	1.004620	0.998473	0.788285	0.802263
3	роман	0.999033	0.991512	0.794793	0.808854
4	роман	0.996945	0.987663	0.777628	0.790161
5	детектив	0.991697	0.973819	0.793684	0.802822
6	детектив	0.991656	0.986506	0.784293	0.794182
7	приключения	0.991037	0.979161	0.795032	0.805405
8	роман	0.988051	0.988768	0.801261	0.811563
9	фантастика	0.984583	0.979749	0.790036	0.803747
10	фантастика	0.984467	0.972981	0.798092	0.807740
11	роман	0.983066	0.994065	0.800523	0.812804
12	фантастика	0.982076	0.983231	0.810374	0.821457
13	детектив	0.982069	0.954409	0.804559	0.812377
14	детектив	0.981934	0.972231	0.806420	0.816998
15	роман	0.978492	0.968451	0.815062	0.825980
16	роман	0.978363	0.966682	0.798223	0.807001
17	детектив	0.978101	0.967885	0.809228	0.819173
18	детская лит.	0.976800	1.012762	0.742432	0.756829
19	фантастика	0.976773	0.966636	0.784674	0.796484
20	приключения	0.971846	0.961520	0.823809	0.831446
21	роман	0.971531	0.958468	0.806512	0.815702
22	приключения	0.971082	0.989721	0.792677	0.802851
23	роман	0.970900	0.962113	0.794577	0.804060
24	роман	0.968299	0.993697	0.803362	0.815941
25	детская лит.	0.968028	0.959380	0.777983	0.793339

Продолжение табл. 1

Текст	Жанр	Ципф (словоформы)	Ципф (леммы)	Хипс (леммы)	Хипс (словоформы)
26	роман	0.967511	0.974234	0.754915	0.767074
27	роман	0.966305	1.001287	0.778061	0.790588
28	фантастика	0.965116	0.950745	0.794937	0.804610
29	фантастика	0.961867	0.949584	0.813870	0.825393
30	роман	0.961286	0.952750	0.799193	0.809003
31	фантастика	0.955980	0.945660	0.803026	0.810366
32	фантастика	0.955516	0.940502	0.809863	0.820718
33	роман	0.954731	1.026885	0.741586	0.753864
34	роман	0.952700	0.991605	0.795840	0.811328
35	фантастика	0.952088	0.941467	0.780060	0.788162
36	детская лит.	0.950748	0.972238	0.771153	0.781493
37	детектив	0.948861	0.967911	0.792331	0.801062
38	фантастика	0.948237	0.945391	0.801813	0.814089
39	роман	0.930612	0.972905	0.816378	0.824606
Среднее:		0.973863	0.975318	0.792458	0.803458
$3 \times \sigma$:		0.057036	0.065021	0.055954	0.053281

σ — стандартное отклонение

Таблица 2

Значения коэффициентов законов Ципфа и Хипса для русского языка

Текст	Жанр	Ципф (словоформы)	Ципф (леммы)	Хипс (леммы)	Хипс (словоформы)
1	детская лит.	0.936576	0.964813	0.787141	0.841100
2	роман	0.935878	0.964046	0.825040	0.871886
3	роман	0.929603	0.955567	0.839364	0.889200
4	детектив	0.928132	0.939130	0.839518	0.886388
5	детектив	0.924204	0.944139	0.858930	0.894042
6	детектив	0.917411	0.942821	0.822190	0.873935
7	приключения	0.916674	0.960386	0.793264	0.855948
8	роман	0.912970	0.931723	0.842878	0.885869
9	роман	0.912406	0.940216	0.822597	0.871927
10	детектив	0.909435	0.927857	0.839980	0.889580
11	роман	0.908496	0.963706	0.814065	0.864963
12	роман	0.906881	0.922668	0.838711	0.886990
13	фантастика	0.903534	0.919563	0.816362	0.868314
14	роман	0.902698	0.927154	0.846717	0.894226
15	фантастика	0.902272	0.915499	0.842399	0.885195
16	детская лит.	0.901783	0.916074	0.844565	0.886987
17	фантастика	0.899720	0.911501	0.821493	0.871524
18	фантастика	0.892304	0.907987	0.853072	0.896268
19	роман	0.890569	0.946387	0.846493	0.891929
20	роман	0.890088	0.902435	0.859763	0.900825
21	детектив	0.887773	0.909617	0.838548	0.889677
22	роман	0.886602	0.898627	0.856025	0.897606
23	роман	0.884160	0.963282	0.818838	0.864900
24	роман	0.883826	0.896010	0.832264	0.885477
25	детектив	0.883621	0.880983	0.872263	0.910767
26	детская лит.	0.883044	0.885564	0.856513	0.895081
27	фантастика	0.881713	0.889017	0.848118	0.891209
28	приключения	0.880597	0.899939	0.834420	0.882924
29	роман	0.879422	0.887770	0.873361	0.905620
30	фантастика	0.876683	0.885460	0.858251	0.899792
31	роман	0.874849	0.888930	0.852379	0.897232
32	детектив	0.873471	0.907970	0.830596	0.882299
33	детектив	0.870795	0.863837	0.876895	0.915232
34	роман	0.867954	0.885425	0.871117	0.907745
35	фантастика	0.867008	0.870758	0.870979	0.903001
36	фантастика	0.863004	0.879573	0.841957	0.884644
37	приключения	0.859045	0.894258	0.834773	0.885242
38	детектив	0.857402	0.871889	0.850555	0.896164
39	фантастика	0.839270	0.840562	0.881458	0.912924
Среднее:		0.892869	0.912901	0.842406	0.887555
$3 \times \sigma$:		0.068292	0.094028	0.063054	0.046417

σ — стандартное отклонение

СПИСОК ОБРАБОТАННЫХ ТЕКСТОВ

В наших экспериментах использовались нижеперечисленные тексты. Номер текста в *Приложении 1* соответствует номеру текста в следующих таблицах. Наборы текстов для русского и английского языков примерно эквиваленты.

Таблица 3

Английские тексты

Текст	Автор	Название	Жанр
1.	Arthur Conan Doyle	Novels and Stories	детектив
2.	Walter Scott	Ivanhoe	приключение
3.	Herman Melville	Moby Dick	роман
4.	Harriet Beecher Stowe	Uncle Tom's Cabin	роман
5.	Arthur Conan Doyle	The Case Book of Sherlock Holmes	детектив
6.	Arthur Conan Doyle	The Memoirs of Sherlock Holmes	детектив
7.	Edgar Rice Burroughs	Tarzan of The Apes	приключения
8.	Thomas Hardy	Far from the Madding Crowd	роман
9.	Winn Schwartau	Terminal Compromise	фантастика
10.	Anthony Hope	The Prisoner of Zenda	фантастика
11.	Mark Twain	Life on the Mississippi	роман
12.	Jules Verne	From the Earth to the Moon	фантастика
13.	Arthur Conan Doyle	His Last Bow	детектив
14.	G. K. Chesterton	The Innocence of Father Brown	детектив
15.	Nathaniel Hawthorne	The Scarlet Letter	роман
16.	Mark Twain	The Adventures of Tom Sawyer	роман
17.	G. K. Chesterton	The Wisdom of Father Brown	детектив
18.		Laddie. A True Blue Story	детская лит.
19.	Richard J. Denissen	The Europa Affair	фантастика
20.	Ambrose Bierce	Can Such Things Be	приключения
21.	Jules Verne	Around the World in Eighty Days	роман
22.	Edgar Rice Burroughs	The Mucker	приключения
23.	Arthur Conan Doyle	Valley of Fear	роман
24.	Walter Scott	Chronicles of the Canongate	роман
25.	R. Kipling	The Jungle Book	детская лит.
26.	Jane Austin	Pride and Prejudice	роман
27.	D. H. Lawrence	Sons and Lovers	роман
28.	Douglas K. Bell	Jason the Rescuer	фантастика
29.	William Gibson	Neuromancer	фантастика
30.	Baroness Orczy	The Scarlet Pimpernel	роман
31.	Douglas Adams	The Restaurant at the End of the Universe;	фантастика
32.	Douglas K. Bell	Van Gogh in Space	фантастика
33.	Mark Twain	The Adventures of Huckleberry Finn	роман
34.		Walden & on The Duty of Civil Disobedience	роман
35.	Lawrence Dworin	Revolt of the Cyberslaves	фантастика
36.	Lucy Maud Montgomery	Anne of Green Gables	детская лит.
37.	Arthur Conan Doyle	Hound of Baskervilles	детектив
38.	Bruce Sterling	The Hacker Crackdown	фантастика
39.	Nathaniel Hawthorne	The House of the Seven Gables	роман

Таблица 4

Русские тексты

Текст	Автор	Название	Жанр
1.	Николай Носов	Приключения Незнайки	детская лит.
2.	Василий Аксенов	Сборник	роман
3.	А. Солженицын	Архипелаг ГУЛаг	роман
4.	Анатолий Степанов	День гнева	детектив
5.	Виктор Федоров, Виталий Щигельский	Бенефис двойников	детектив
6.	Юлиан Семенов	Семнадцать мгновений весны	детектив
7.	Генри Райдер Хаггард	Дочь Монтесумы	приключения
8.	Вл. Кунин	Повести	роман
9.	Александр Покровский	"...Расстрелять"	роман
10.	Марина Наумова	Констрикторы	детектив
11.	Федор Достоевский	Неточка Незванова	роман

Продолжение табл. 4

Текст	Автор	Название	Жанр
12.		Азюль	роман
13.	В. Пелевин	Сборник рассказов и повестей	фантастика
14.	М. Горький	Автобиографические рассказы	роман
15.	Сергей Михайлов	Шестое чувство	фантастика
16.	Л. Лагин	Старик Хоттабыч	детская лит.
17.	Дмитрий Громов	Сборник рассказов и повестей	фантастика
18.	Вячеслав Рыбаков	Рассказы	фантастика
19.	Евгений Козловский	Киносценарии и повести	роман
20.	Александр Мелихов	Во имя четыреста первого, или Исповедь еврея	роман
21.	Андрей Курков		детектив
22.	Всеволод Иванов	Голубые пески	роман
23.	Михаил Мишин	Почувствуйте разницу	роман
24.	Андрей Платонов	Котлован	роман
25.	Виктор Черняк	Выездной!	детектив
26.	Александр Некрасов	Приключения капитана Врунгеля	детская лит.
27.	Игорь Федоров	Рассказы	фантастика
28.	Ульрих Комм	Фрегаты идут на абордаж	приключения
29.	Наталья Галкина	Ночные любимцы	роман
30.	Б. Иванов, Ю. Щербатых	Случай контрабанды	фантастика
31.	Владимир Набоков	Рассказы	роман
32.	Виктор Суворов	Аквариум	детектив
33.	Виктор Черняк	Жулье	детектив
34.	Сергей Дышев	До встречи в раю	роман
35.	Ник Перумов	Рассказы, Русский меч	фантастика
36.	Антон Первушин	Рассказы	фантастика
37.	Т. Майн Рид	Американские партизаны	приключения
38.	Михаил Болтунов	"Альфа" — сверхсекретный отряд КГБ	детектив
39.	Виталий Бабенко	Игоряша "Золотая рыбка"	фантастика

Материал поступил в редакцию 13.06.01.