Zipf and Heaps Laws' Coefficients Depend on Language*

Alexander Gelbukh and Grigori Sidorov

Center for Computing Research (CIC), National Polytechnic Institute (IPN),
Av. Juan Dios Batiz s/n esq. Mendizabal, col. Zacatenco, CP 07738, DF, Mexico.
{gelbukh, sidorov}@cic.ipn.mx

Abstract. We observed that the coefficients of two important empirical statistical laws of language – Zipf law and Heaps law – are different for different languages, as we illustrate on English and Russian examples. This may have both theoretical and practical implications. On the one hand, the reasons for this may shed light on the nature of language. On the other hand, these two laws are important in, say, full-text database design allowing predicting the index size.

Introduction. Perhaps the most famous statistical distribution in linguistics is Zipf law [1, 2]: in any large enough text, the frequency ranks (starting from the highest) of wordforms or lemmas are inversely proportional to the corresponding frequencies:[1]

$$\log f_r \approx C - z \log r \tag{1}$$

where f_i is the frequency of the unit (wordform or lemma) having the rank r, z is the exponent coefficient (near to 1), and C is a constant. In a logarithmic scale, it is a straight line with about – 45° angle.

Another, less famous but probably not less important empirical statistical law of language is the Heaps law: the number of different wordforms or lemmas in a text is roughly proportional to an exponent of its size:

$$\log n_i \approx D + h \log i \tag{2}$$

where n_i is the number of different units (wordforms or lemmas) occurring before the running word number i, h is the exponent coefficient (between 0 and 1), and D is a constant. In a logarithmic scale, it is a straight line with about 45° angle.

The nature of these laws is not clear. They seem to be specific for natural languages in contrast to other types of signals [3]. In practice, knowing the coefficients of these laws is important in, for example, full-text database design, since it allows predicting some properties of the index as a function of the size of the database.

In this paper, we present the data that show that the coefficients of both laws – z and h – depend on language. For our experiments, we use English and Russian texts. Experiments with Spanish (which we do not discuss here) gave the results between those for English and Russian.

* The work was 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] We ignore Mandelbrot's improvements to Zipf law [1] since they do not affect our discussion.

Experimental data. We processed 39 literature texts for each language, see Appendix 2, chosen randomly from different genres, with the requirement that the size be greater than 10,000 running words (100 KB); total of 2.5 million running words (24.8 MB) for English and 2.0 million (20.2 MB) for Russian.

We experimented with wordforms and lemmas, with very similar results. We plotted on the screen the graphs for pairs of texts (one English and one Russian), using for Zipf law the points: $x_r = \log r$, $y_i = \log f_r$ ($x_i = \log i$, $y_i = \log n_i$ for Heaps law). The difference in the angle was in most cases clearly visible.

We used linear regression to approximate such a graph by a straight line $y = ax + b$, where a and b correspond to $-z$ and C for Zipf law, or h and D for Heaps law. Since the density of the points (x_i, y_i) increases exponentially with x_i, we scaled the distance penalty for regression by c^{-x_i} (we have to omit here the details; obviously, the results do not depend on c), which gave the following formulae for a and b:

$$p = \sum_i \frac{x_i}{c^{x_i}}, \quad s = \sum_i \frac{x_i^2}{c^{x_i}}, \quad t = \sum_i \frac{y_i}{c^{x_i}}, \quad u = \sum_i \frac{1}{c^{x_i}}, \quad b = \frac{p \sum_i \frac{x_i y_i}{c^{x_i}} - st}{p^2 - su}, \quad a = \frac{t - bu}{p}.$$

Visual control proved that these weighted formulae approximate the graphs much better than the standard linear regression ones. The results are shown in Appendix 1 (ordered by z); we give the values of z (Zipf, on wordforms) and h (Heaps, on lemmas) and omit C and D since they are less important. The difference between the two languages is obvious. For English $z = 0.97 \pm 0.06$ and for Russian $z = 0.89 \pm 0.07$, the difference being 8.3% (as a measure of precision, we use 3σ, where σ is the standard deviation); for English $h = 0.79 \pm 0.05$ and for Russian $h = 0.84 \pm 0.06$, the difference being 5.9%.

Discussion. Two properties of the languages in question might be involved in the explanation of this phenomenon. First, Russian is a highly inflective language while English is analytical. Our experiments with Spanish seem to favor this consideration: Spanish, having “inflectivity” intermediate between Russian and English, showed intermediate results as to the coefficients. On the other hand, counting lemmas instead of wordforms nearly did not change our results. Second, it is well known that lexical richness of Russian is greater than that of English (and Spanish).

Conclusions. Exponential coefficients of Zipf and Heaps laws depend on language. This can have both theoretical and practical implications (the latter, for example, in full-text database design). Explanation of this phenomenon needs more investigation.

References

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

Appendix 1. Experimental Results

English				Russian			
Text	**Genre**	**Zipf**	**Heaps**	**Text**	**Genre**	**Zipf**	**Heaps**
1	detective	1.037639	0.759330	1	children	0.936576	0.787141
2	adventure	1.004620	0.788285	2	novel	0.935878	0.825040
3	novel	0.999033	0.794793	3	novel	0.929603	0.839364
4	novel	0.996945	0.777628	4	detective	0.928132	0.839518
5	detective	0.991697	0.793684	5	detective	0.924204	0.858930
6	detective	0.991656	0.784293	6	detective	0.917411	0.822190
7	adventure	0.991037	0.795032	7	adventure	0.916674	0.793264
8	novel	0.988051	0.801261	8	novel	0.912970	0.842878
9	SF/fantasy	0.984583	0.790036	9	novel	0.912406	0.822597
10	SF/fantasy	0.984467	0.798092	10	detective	0.909435	0.839980
11	novel	0.983066	0.800523	11	novel	0.908496	0.814065
12	SF/fantasy	0.982076	0.810374	12	novel	0.906881	0.838711
13	detective	0.982069	0.804559	13	SF/fantasy	0.903534	0.816362
14	detective	0.981934	0.806420	14	novel	0.902698	0.846717
15	novel	0.978492	0.815062	15	SF/fantasy	0.902272	0.842399
16	novel	0.978363	0.798223	16	children	0.901783	0.844565
17	detective	0.978101	0.809228	17	SF/fantasy	0.899720	0.821493
18	children	0.976800	0.742432	18	SF/fantasy	0.892304	0.853072
19	SF/fantasy	0.976773	0.784674	19	novel	0.890569	0.846493
20	adventure	0.971846	0.823809	20	novel	0.890088	0.859763
21	novel	0.971531	0.806512	21	detective	0.887773	0.838548
22	adventure	0.971082	0.792677	22	novel	0.886602	0.856025
23	novel	0.970900	0.794577	23	novel	0.884160	0.818838
24	novel	0.968299	0.803362	24	novel	0.883826	0.832264
25	children	0.968028	0.777983	25	detective	0.883621	0.872263
26	novel	0.967511	0.754915	26	children	0.883044	0.856513
27	novel	0.966305	0.778061	27	SF/fantasy	0.881713	0.848118
28	SF/fantasy	0.965116	0.794937	28	adventure	0.880597	0.834420
29	SF/fantasy	0.961867	0.813870	29	novel	0.879422	0.873361
30	novel	0.961286	0.799193	30	SF/fantasy	0.876683	0.858251
31	SF/fantasy	0.955980	0.803026	31	novel	0.874849	0.852379
32	SF/fantasy	0.955516	0.809863	32	detective	0.873471	0.830596
33	novel	0.954731	0.741586	33	detective	0.870795	0.876895
34	novel	0.952700	0.795840	34	novel	0.867954	0.871117
35	SF/fantasy	0.952088	0.780060	35	SF/fantasy	0.867008	0.870979
36	children	0.950748	0.771153	36	SF/fantasy	0.863004	0.841957
37	detective	0.948861	0.792331	37	adventure	0.859045	0.834773
38	SF/fantasy	0.948237	0.801813	38	detective	0.857402	0.850555
39	novel	0.930612	0.816378	39	SF/fantasy	0.839270	0.881458
	Average:	0.973863	0.792458		**Average:**	0.892869	0.842406
	3 × deviation:	0.057036	0.055954		**3 × deviation:**	0.068292	0.063054

For both Zipf and Heaps, levels of significance of difference are much better than 1%.

Appendix 2. Sources

The following texts were used for our experiments. The text number in Appendix 1 corresponds to the number in the corresponding list below.

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

Russian sources:[2] 1. Николай Носов. *Приключения Незнайки;* 2. Василий Аксенов. *Сборник;* 3. А.Солженицын. *Архипелаг ГУЛаг;* 4. Анатолий Степанов. *День гнева;* 5. Виктор Федоров, Виталий Щигельский. *Бенефис двойников;* 6. Юлиан Семенов. *Семнадцать мгновений весны;* 7. Генри Райдер Хаггард. *Дочь Монтесумы;* 8. Вл. Кунин. *Повести;* 9. Александр Покровский. *"...Расстрелять";* 10. Марина Наумова. *Констрикторы;* 11. Федор Достоевский. *Неточка Незванова;* 12. *Азюль;* 13. В. Пелевин. *Сборник рассказов и повестей;* 14. М. Горький. *Автобиографические рассказы;* 15. Сергей Михайлов. *Шестое чувство;* 16. Л. Лагин. *Старик Хоттабыч;* 17. Дмитрий Громов. *Сборник рассказов и повестей;* 18. Вячеслав Рыбаков. *Рассказы;* 19. Евгений Козловский. *Киносценарии и повести;* 20. Александр Мелихов. *Во имя четыреста первого, или Исповедь еврея;* 21. Андрей Курков. 22. Всеволод Иванов. *Голубые пески;* 23. Михаил Мишин. *Почувствуйте разницу;* 24. Андерй Платонов. *Котлован;* 25. Виктор Черняк. *Выездной!;* 26. Александр Некрасов. *Приключения капитана Врунгеля;* 27. Игорь Федоров. *Рассказы;* 28. Ульрих Комм. *Фрегаты идут на абордаж;* 29. Наталья Галкина. *Ночные любимцы;* 30. Б. Иванов, Ю. Щербатых. *Случай контрабанды;* 31. Владимир Набоков. *Рассказы;* 32. Виктор Суворов. *Аквариум;* 33. Виктор Черняк. *Жулье;* 34. Сергей Дышев. *До встречи в раю;* 35. Ник Перумов. *Рассказы, Русский меч;* 36. Антон Первушин. *Рассказы;* 37. Т. Майн Рид. *Американские партизаны;* 38. Михаил Болтунов. *"Альфа" - сверхсекретный отряд КГБ;* 39. Виталий Бабенко. *Игоряша "Золотая рыбка".*

[2] Since most of these titles do not have any standard English translations and many of the authors are not known outside Russia, we give the titles and names in Russian. Their understanding is not relevant for our discussion. The mixture roughly corresponds to English one.