Scaling Laws in Human Language

Linyuan Lü, Zi-Ke Zhang, and Tao Zhou

Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 610054, People’s Republic of China
Department of Physics, University of Fribourg, Chemin du Muse, Fribourg 1700, Switzerland
Beijing Computational Science Research Center, Beijing 100084, People’s Republic of China

(Dated: February 15, 2012)

Zipf’s law on word frequency is observed in English, French, Spanish, Italian, and so on, yet it does not hold for Chinese, Japanese or Korean characters. A model for writing process is proposed to explain the above difference, which takes into account the effects of finite vocabulary size. Experiments, simulations and analytical solution agree well with each other. The results show that the frequency distribution follows a power law with exponent being equal to 1, at which the corresponding Zipf’s exponent diverges. Actually, the distribution obeys exponential form in the Zipf’s plot. Deviating from the Heaps’ law, the number of distinct words grows with the text length in three stages: It grows linearly in the beginning, then turns to a logarithmical form, and eventually saturates. This work refines previous understanding about Zipf’s law and Heaps’ law in language systems.

PACS numbers: 89.20.Hh, 89.75.Hc

Uncovering the statistics and dynamics of human language helps in characterizing the universality, specificity and evolution of cultures [1][11]. Two scaling relations, Zipf’s law [12] and Heaps’ law [13], have attracted much attention from academic community. Denote $r$ the rank of a word according to its frequency $Z(r)$, Zipf’s law is the relation $Z(r) \sim r^{-\alpha}$, with $\alpha$ being the Zipf’s exponent. Zipf’s law was observed in many human languages, including English, French, Spanish, Italian, and so on [12,14,15]. Heaps’ law is formulated as $N_t \sim t^\lambda$, where $N_t$ is the number of distinct words when the text length is $t$, and $\lambda \leq 1$ is the so-called Heaps’ exponent. These two laws coexist in many language systems. Gelbukh and Sidorov [16] observed these two laws in English, Russian and Spanish texts, with different exponents depending on languages. Similar results were recently reported for the corpus of web texts [17], including the Industry Sector database, the Open Directory and the English Wikipedia. The occurrences of tags for online resources [18][19], keywords for scientific publications [20] and words contained by web pages resulted from web searching [21] also simultaneously display the Zipf’s law and Heaps’ law. Interestingly, even the identifiers in programs by Java, C++ and C languages exhibit the same scaling laws [22].

The Zipf’s law in language systems could result from a rich-get-richer mechanism as suggested by the Yule-Simon model [23][24], where a new word is added to a text with probability $q$ and an appeared word is randomly chosen and copied with probability $1-q$. A word appears more frequently thus has high probability to be copied, leading to a power-law word frequency distribution $p(k) \sim k^{-\beta}$ with $\beta = 1 + 1/(1-q)$. Dorogovtsev and Mendes modeled the language processing as evolution of a word web with preferential attachment [25].

| Books                  | $V$  | $N_T$  | $k_{max}$ | $k_{min}$ | $\beta$ |
|------------------------|------|--------|-----------|-----------|--------|
| The Story of the Stone | 727601 | 4239 | 21054 | 1 | 1.09 |
| The Battle Wizard      | 1020336 | 4178 | 20028 | 1 | 1.03 |
| Into the White Night   | 420935 | 2182 | 18992 | 1 | 1.00 |
| History of the Three Kingdoms | 157201 | 1139 | 5929 | 1 | 1.07 |

Zanette and Montemurro [26] as well as Cattuto et al. [27] accounted for the memory effects, say the recently used words have higher probability to be chosen than the words occurred long time ago. These works can be considered as variants of the Yule-Simon model. Meanwhile, the Heaps’ law may originate from the memory and bursty nature of human language [28][30].

Real language systems to some extent deviate from these two scaling laws and display more complicated statistical regularities. Wang et al. [31] analyzed representative publications in Chinese, and showed that the character frequency distribution exhibits an exponential feature. Lü et al. [32] pointed out that in a growing system, if the appearing frequencies of elements obey the Zipf’s law with stable exponent, then the number of distinct elements grows in a complicated way with the Heaps’ law only an asymptotical approximation. This deviation from the Heaps’ law was further emphasized and mathematically proved by Eliazar [33].

Empirical analyses on real language systems showed similar deviation [34]. Via extensive analysis on individual Chinese, Japanese and Korean books, as well as a collection of more than 5 x 10^4 Chinese books, we found even more complicated phenomena: (i) the character frequency distribution follows a power law yet it decays exponentially in the Zipf’s plot; (ii) with the increasing of text length, the number of
distinct characters grows in three different stages: linear, logarithmical and saturated. All these unreported regularities may result from the finite vocabulary size, which is further verified by a simple theoretical model.

We first show some experimental results about the statistical regularities on Chinese, Japanese and Korean literatures, which are typical examples generated from a vocabulary of very limited size if we look at the character level. There are in total more than $9 \times 10^4$ Chinese characters, yet only 3000 to 4000 of which are used frequently (Taiwan and Hong Kong respectively identify 4808 and 4759 frequently used characters, while mainland China has two versions of the list of frequently used characters, one contains 2500 characters and the other contains 3500 characters), and the number of Japanese and Korean characters are even smaller. We start with four famous books, the first two are in Chinese, the third one is in Japanese and the last one is in Korean: (i) The Story of the Stone (aliases: The Dream of the Red Chamber, A dream of Red Mansions and Hong Lou Meng), written by Xueqin Cao in the mid-eighteenth century during the reign of Emperor Chien-lung of the Qing Dynasty; (ii) The Battle Wizard (aliases: Tian Long Ba Bu and Demi-Gods and Semi-Devils), a kung fu novel written by Yong Jin; (iii) Into the White Night, a modern novel written by Higashino Keigo; (iv) The History of the Three Kingdoms, a very famous history book by Shou Chen in China and then translated into Korean. These books cover disparate topics and types and were accomplished in far different dates. The basic statistics of these books are presented in Table 1.

Figure 1 reports the character frequency distribution $p(k)$, the Zipf’s plot on character frequency $Z(r)$ and the growth of the number of distinct characters $N_t$ versus $t$. The number of distinct words versus the text length of The Story of the Stone in (a3) log-log scale and (a4) linear-log scale. Similar plots in (b1-b4), (c1-c4) and (d1-d4) are for the books The Battle Wizard, Into the White Night and The History of the Three Kingdoms, respectively. The power-law exponent $\beta$ is obtained by using the maximum likelihood estimation $[33, 34]$, while the exponent in the Zipf’s plot is obtained by the least square method excluding the head (i.e., $r > 500$ for Chinese books and $r > 200$ for Japanese and Korean books).
FIG. 2: (Color online) The growth of distinct characters in the collection of 57755 Chinese books. The result is obtained by averaging over 100 randomly determined orders of these books.

the total number of characters appeared in the text. As shown in figure 1, the character frequency distributions are power-law, meanwhile the frequency decays exponentially in the Zipf’s plot, which is in conflict to the common sense that a power-law probability density function always corresponds to a power-law decay in the Zipf’s plot. Actually, there exists a relation between two exponents $\alpha$ and $\beta$ as $\alpha = \frac{1}{\beta - 1}$, and thus when $\beta$ gets close to 1, the exponent $\alpha$ will diverge and thus the decaying function in Zipf’s plot could not be well characterized by a power law. Therefore, if we observe a non-power-law decaying in the Zipf’s plot, we cannot immediately deduce that the corresponding probability density function is not a power law – it is possibly a power law with exponent close to 1. Note that, in the Zipf’s plots, the turned-up head contains a few hundreds of characters, majority of which play the similar role to the auxiliary words, conjunctions or prepositions in English.

Figure 1 also indicates that the growth of distinct characters cannot be described by the Heaps’ law. Indeed, there are two distinguishable stages: In the early stage, $N_t$ grows approximately linearly with the text length $t$, and in the later stage, $N_t$ grows logarithmically with $t$. Figure 3 presents the growth of distinct characters for a large collection of 57755 Chinese books consisting of about $3.4 \times 10^9$ characters and 12800 distinct characters. In addition to those observed in figure 1 and figure 2, $N_t$ displays a strongly saturated behavior when the text length $t$ is much bigger than the total distinct characters in the vocabulary. In summary, the experiments on Chinese, Japanese and Korean literature show us some unreported phenomena: the character frequency obeys a power law with exponent close to 1 yet it decays exponentially in the Zipf’s plot, and the number of distinct characters grows in three distinguishable stages. We next propose a theoretical model to explain these observations.

Consider a vocabulary with finite number, $V$, of distinct characters or words. At each time step, one character will be selected from the vocabulary to form the text. Motivated by the rich-get-richer mechanism of the Yule-Simon model, at time step $t$, if the character $i$ has been used $k_i$ times, it will be selected with the probability

$$f(k_i) = \frac{k_i + \varepsilon}{\sum_{j=1}^{V} (k_j + \varepsilon)} = \frac{k_i + \varepsilon}{V\varepsilon + t - 1}, \quad (1)$$

where $\varepsilon$ is the initial attractiveness of each character. Assuming that at time $t$, there are $N_t$ distinct characters in the text, and we first investigate the dependence of $N_t$ on the text length $t$. The selection at time $t+1$ can be equivalently divided into two complementary yet repulsive actions: (i) to select a character from the original vocabulary with probability proportional to $\varepsilon$, or (ii) to select a character from the $N_t$ words in the created text with probability proportional to its appeared frequency. Therefore the probability to choose a character from the original vocabulary is $\frac{V\varepsilon}{V\varepsilon + t}$, whereas $\frac{k_i + \varepsilon}{V\varepsilon + t}$ from the created text. A character chosen from the created text is always old, while a character chosen from the vocabulary could be new with probability $1 - \frac{N_t}{V}$. Accordingly, the probability that a new character appears at the $t+1$ time step, namely the growing rate of $N_t$, is

$$\frac{dN_t}{dt} = \frac{V\varepsilon}{V\varepsilon + t} \left( 1 - \frac{N_t}{V} \right). \quad (2)$$

With the boundary conditions $N_0 = 0$ and $N_\infty = V$, we derive the solution of Eq. (2) as

$$N_t = V \left[ 1 - \left( \frac{V\varepsilon}{V\varepsilon + t} \right)^\varepsilon \right]. \quad (3)$$

This solution embodies three stages of growth of $N_t$. (i) In the very early stage, when $t$ is much smaller than $V\varepsilon$, $(\frac{V\varepsilon}{V\varepsilon + t})^\varepsilon \approx 1 - \frac{t}{V\varepsilon}$ and thus $N_t \approx t$, corresponding
to a short period of linear growth. (ii) When \( t \) is of the same order of \( V \varepsilon \), if \( \varepsilon \) is very small, \( N_t \) could be much smaller than \( V \). Then Eq. 2 can be approximated as

\[
\frac{dN_t}{dt} \approx \frac{V \varepsilon}{V \varepsilon + t},
\]

leading to a logarithmical solution

\[
N_t \approx V \varepsilon \ln \left(1 + \frac{t}{V \varepsilon}\right).
\]

Indeed, expanding \( \left(\frac{V \varepsilon}{V \varepsilon + t}\right)^\varepsilon \) by Taylor series as

\[
\left(\frac{V \varepsilon}{V \varepsilon + t}\right)^\varepsilon = \sum_{m=0}^{\infty} \frac{1}{m!} \varepsilon \cdot \ln \left(\frac{V \varepsilon}{V \varepsilon + t}\right)^m
\]

and neglecting the high-order terms \((m \geq 2)\) under the condition \( \varepsilon \ll 1 \), one can also arrive to the solution Eq. 5.

(iii) When \( t \) gets larger and larger, \( N_t \) will approach to \( V \) and thus both \( \frac{V \varepsilon}{V \varepsilon + t} \) and \( 1 - \frac{V \varepsilon}{V \varepsilon + t} \) are very small, leading to a very slow growing of \( N_t \) according to Eq. 2. These three stages predicted by the analytical solution are in good accordance with the above empirical observations.

Figure 3 reports the numerical results on Eq. 3. In accordance with the analysis, when \( t \) is small, \( N_t \) grows in a linear form as shown in Fig. 3(a) and 3(c), and from Fig. 3(b) and 3(d), straight lines appear in the middle region, indicating a logarithmical growth predicted by Eq. 5.

Denote by \( n(t,k) \) the number of distinct characters that appeared \( k \) times until time \( t \), then \( n(t,k) = N_t p(k) \). According to the master equations, we have

\[
n(t+1,k+1) = n(t,k+1) [1 - f(k+1)] + n(t,k) f(k).
\]

Substituting Eq. 11 into Eq. 12 we obtain

\[
N_{t+1} p(k+1) = N_t p(k+1) \left[1 - \frac{k+1 + \varepsilon}{V \varepsilon + t}\right] + \frac{N_t p(k+1) + N_t p(k)}{V \varepsilon + t}.
\]

Via continuous approximation, it turns to be the following differential equation

\[
\frac{dp}{dt} = -\left[1 + \frac{V \varepsilon + t}{N_t} (N_{t+1} - N_t)\right] \frac{dk}{k + \varepsilon}.
\]

Substituting \( N_{t+1} = N_t = dN_t/dt \) and Eq. 12 we get the solution

\[
p(k) = B(k + \varepsilon)^{1 [1 + \varepsilon (\frac{k}{N_t} - 1)]},
\]

where \( B \) is the normalized factor. The result shows that the character frequency follows a power-law distribution with exponent changing in time. Considering the finite vocabulary size, in the large limit of \( t, N_t \to V \) and thus the power-law exponent, \( \beta = 1 + \varepsilon \left(\frac{V}{N_t} - 1\right) \), approaches 1. Under the continuous approximation, the cumulative distribution of character frequency can be written as

\[
P(k > k_0) = 1 - \int_{k_0}^{k_0} p(k) dk = 1 - B \frac{k^{1 - \beta}}{1 - \beta} \frac{k_0^{1 - \beta}}{1 - \beta},
\]

where \( k_{\text{min}} \) is the smallest frequency. When \( \beta \to 1 \), \( k^{1 - \beta} \approx 1 + (1 - \beta) \ln k \), and thus

\[
P(k > k_0) = 1 - B \ln \frac{k_0 + \varepsilon}{k_{\text{min}} + \varepsilon},
\]

where \( B \approx \left(\frac{\ln k_{\text{max}} + \varepsilon}{k_{\text{min}} + \varepsilon}\right)^{-1} \) according to the normalization condition \( \int_{k_{\text{min}}}^{k_{\text{max}}} p(k) dk = 1 \) and \( k_{\text{max}} \) is the highest frequency. According to Eq. 12, there are \( 1 - B \ln \frac{k_{\text{max}} + \varepsilon}{k_{\text{min}} + \varepsilon} \) \( N_t \) characters having appeared more than \( k \) times. That is to say, a character having appeared \( k \) times will be ranked at \( r = 1 + \left(1 - B \ln \frac{k + \varepsilon}{k_{\text{min}} + \varepsilon}\right) N_t \).

Therefore

\[
Z(r) = k = (k_{\text{min}} + \varepsilon) \exp \left[\frac{1}{B} \left(1 - \frac{r - 1}{N_t}\right)\right] - \varepsilon,
\]

and \( Z(1) = k_{\text{max}}, Z(N_t) = k_{\text{min}} \). In a word, this simple model accounting for the finite vocabulary size results in a power-law character frequency distribution \( p(k) \sim k^{-\beta} \) with exponent \( \beta \) close to 1 and an exponential decay of \( Z(r) \) in the Zipf’s plot, which perfectly agree with the empirical observations on Chinese, Japanese and Korean books.

Figure 4 reports the simulation results for typical parameters. The power-law frequency distribution, the exponential decay of frequency in the Zipf’s plot and the
linear to logarithmic transition in the growth of the distinct number of characters are all clearly observed in the simulation. The simulation results agree very well with the analytical solutions presented in Eq. 3, Eq. 10 and Eq. 13.

Previous statistical analyses about human language overwhelmingly concentrate on Indo-European family, where each language consists of a huge number of words. In contrast, languages consisting of characters, though cover more than a billion people, obtained less attention. These languages include Chinese, Japanese, Korean, Vietnamese, Jurchen language, Khitan language, Makhi language, Tangut language, and many others. Empirical studies here show remarkably different scaling laws of character-formed from word-formed languages. Salient features include an exponential decay of character frequency in the Zipf’s plot associated with a power-law frequency distribution with exponent close to 1, and a frequency in the Zipf’s plot for a power-law distribution with exponent around 1, such as the cluster size distribution in two-dimensional self-organized critical systems, the inter-event time distribution in human activities, the family name distribution in Korea, species lifetime distribution, and so on. Meanwhile, we cannot deny a possibly power-law distribution just from a non-power-law decay in the Zipf’s plot.

The currently reported scaling laws can be reproduced by considering finite vocabulary size in a rich-get-richer process. Different from the well-known finite-size effects that vanish in the thermodynamic limit, the effects caused by finite vocabulary size get stronger as the increasing of the system size. Finite choices must be a general feature in selecting dynamics, but not a necessary ingredient in growing systems. For example, also based on the rich-get-richer mechanism, neither the linear growing model nor the accelerated growing model (treating total degree as the text length and nodes as distinct characters, the accelerated networks grow in the Heaps’ manner) has considered such ingredient. The present model could distinguish the selecting dynamics from general dynamics for growing systems.

This work is partially supported by the Swiss National Science Foundation (Project 200020-132253) and the Fundamental Research Funds for the Central Universities.

[1] J. A. Hawkins and M. Gell-Mann, *The Evolution of Human Languages* (Addison-Wesley, Reading, Massachusetts, 1992).
[2] D. Caplan, *Language: Structure, Processing and Disorders* (MIT Press, Cambridge, 1994).
[3] D. Lightfoot, *The Development of Language: Acquisition, Changes and Evolution* (Blackwell, Oxford, 1999).
[4] M. A. Nowak and D. C. Krakauer, Proc. Natl. Acad. Sci. 96, 8028 (1999).
[5] Ramon Ferrer i Cancho and R. V. Solé, Proc. R. Soc. Lond. B 268, 2261 (2001).
[6] M. A. Nowak, N. L. Komarova and P. Niyogi, Nature 417, 611 (2002).
[7] M. D. Hauser, N. Chomsky and W. T. Fitch, Science 298, 1569 (2002).
[8] D. Abrams and S. Strogatz, Nature 424, 900 (2003).
[9] E. Lieberman, J.-B. Michel, J. Jackson, T. Tang and M. A. Nowak, Nature 449, 713 (2007).
[10] R. Lambiotte, M. Ausloos and M. Thelwall, J. Informetrics 1, 277 (2007).
[11] A. M. Petersen, J. Tenenbaum, S. Havlin, and H. E. Stanley, arXiv: 1107.3707.
[12] G. K. Zipf, *Behavior and the Principal of Least Effort* (Addison-Wealey, Cambridge, MA, 1949).
[13] H. S. Heaps, *Information Retrieval-Computational and Theoretical Aspects* (Academic Press, 1978).
[14] I. Kanter, D. A. Kesseler, Phys. Rev. Lett. 74, 4559 (1995).
[15] Ramon Ferrer i Cancho and R. V. Solé, Proc. Natl. Acad. Sci. U.S.A. 100, 788 (2002).
[16] A. Gelbukh and G. Sidorov, Lect. Notes Comput. Sci. 2004, 332 (2001).
[17] M. A. Serrano, A. Flammini, and F. Menczer, PLoS ONE 4, e5372 (2009).
[18] C. Cattuto, V. Loreto, and L. Pietronero, Proc. Natl. Acad. Sci. U.S.A. 104, 1461 (2007).
[19] C. Cattuto, A. Barrat, A. Baldassarri, G. Schehr, and V. Loreto, Proc. Natl. Acad. Sci. U.S.A. 106, 10511 (2009).
[20] Z.-K. Zhang, L. Lü, J.-G. Liu and T. Zhou, Eur. Phys. J. B 66, 557 (2008).
[21] J. C. Lasney and B. Bukiet, J. Quant. Linguistics 16, 40 (2009).
[22] H.-Y. Zhang, Inf. Process. Manage. 45, 477 (2009).
[23] H. A. Simon, Biometrika 42, 425 (1955).
[24] M. V. Simkin and V. P. Roychowdhury, Phys. Rep. 502, 1 (2011).
[25] S. N. Dorogovtsev and J. F. F. Mendes, Proc. R. Soc. Lond. B 268, 2603 (2001).
[26] D. H. Zanette and M. A. Montemurro, J. Quant. Linguistics 12, 29 (2005).
[27] C. Cattuto, V. Loreto, and V. D. P. Servedio, Europhys. Lett. 76, 208 (2006).
[28] W. Ebeling and T. Pöschel, Europhys. Lett. 26, 241 (1994).
[29] J. Kleinberg, Data Min. Knowl. Disc. 7, 373 (2003).
[30] G. G. Altmann, J. B. Pierrehumbert, and A. E. Motter, PLoS ONE 4, e7678 (2009).
[31] D.-H. Wang, M.-H. Li, and Z.-R. Di, Physica A 358, 545 (2005).
[32] L. Lü, Z.-K. Zhang, and T. Zhou, PLoS ONE 5, e14139 (2010).
[33] I. Eliazar, Physica A 390, 3189 (2011).
[34] S. Bernhardsson, L. E. C. da Rocha, and P. Minnhagen, New J. Phys. 11, 123015 (2009).
[35] M. L. Goldstein, S. A. Morris, and G. G. Yen, Eur. Phys. J. B 41, 255 (2004).
[36] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Rev. 51, 661 (2009).
[37] P. Bak, C. Tang, and K. Wiesenfeld, Phys. Rev. A 38, 364 (1988).
[38] A.-L. Barabási, Nature 435, 207 (2005).
[39] B. J. Kim ans S. M. Park, Physica A 347, 683 (2005).
[40] S. Pigolotti, A. Flammini, M. Marsili, and A. Martian, Proc. Natl. Acad. Sci. U.S.A. 102, 15747 (2005).
[41] A.-L. Barabási and R. Albert, Science 286, 509 (1999).
[42] S. N. Dorogovtsev and J. F. F. Mendes, Phys. Rev. E 63, 025101(R) (2001).