

RESEARCH ARTICLE

Model of the Dynamic Construction Process of Texts and Scaling Laws of Words Organization in Language Systems

Shan Li¹, Ruokuang Lin¹, Chunhua Bian^{1,2,3*}, Qianli D. Y. Ma^{2,4,5}, Plamen Ch. Ivanov^{2,5,6*}

1 Institute for Biomedical Electronics Engineering, School of Electronic Science and Engineering, Ministry of Education Key Laboratory of Modern Acoustics, Nanjing University, Nanjing 210093, China, **2** Keck Laboratory for Network Physiology, Department of Physics, Boston University, Boston, MA 02215, United States of America, **3** Department of Neurology and Program in Neuroscience, Harvard Medical School, Beth Israel Deaconess Medical Center, Boston, MA 02215, United States of America, **4** College of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing 210003, China, **5** Harvard Medical School and Division of Sleep Medicine, Brigham and Women's Hospital, Boston, MA 02115, United States of America, **6** Institute of Solid State Physics, Bulgarian Academy of Sciences, Sofia 1784, Bulgaria

* bch@nju.edu.cn (CB); plamen@buphy.bu.edu (PChI)



OPEN ACCESS

Citation: Li S, Lin R, Bian C, Ma QDY, Ivanov PC. (2016) Model of the Dynamic Construction Process of Texts and Scaling Laws of Words Organization in Language Systems. PLoS ONE 11 (12): e0168971. doi:10.1371/journal.pone.0168971

Editor: Boris Podobnik, University of Rijeka, CROATIA

Received: September 20, 2016

Accepted: December 11, 2016

Published: December 22, 2016

Copyright: © 2016 Li et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data of Chinese and English language books are available from the Project Gutenberg database (<http://www.gutenberg.org>).

Funding: This work is supported by the National Natural Science Foundation of China (61271082, 61201029, recipient, Q.D.Y.M.); Fundamental Research Funds for the Central Universities of China, Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD, recipient, C.B.); Nanjing University Scholar

Abstract

Scaling laws characterize diverse complex systems in a broad range of fields, including physics, biology, finance, and social science. The human language is another example of a complex system of words organization. Studies on written texts have shown that scaling laws characterize the occurrence frequency of words, words rank, and the growth of distinct words with increasing text length. However, these studies have mainly concentrated on the western linguistic systems, and the laws that govern the lexical organization, structure and dynamics of the Chinese language remain not well understood. Here we study a database of Chinese and English language books. We report that three distinct scaling laws characterize words organization in the Chinese language. We find that these scaling laws have different exponents and crossover behaviors compared to English texts, indicating different words organization and dynamics of words in the process of text growth. We propose a stochastic feedback model of words organization and text growth, which successfully accounts for the empirically observed scaling laws with their corresponding scaling exponents and characteristic crossover regimes. Further, by varying key model parameters, we reproduce differences in the organization and scaling laws of words between the Chinese and English language. We also identify functional relationships between model parameters and the empirically observed scaling exponents, thus providing new insights into the words organization and growth dynamics in the Chinese and English language.

Introduction

Scaling laws have been discovered and investigated in many fields such as physics, biology, finance, geology, and sociology. Examples include urban growth [1], population distribution of

Exchange Project (recipient, C.B.); National Institutes of Health (NIH Grant 1R01-HL098437, recipient, P.Ch. I.); the US-Israel Binational Science Foundation (BSF Grant 2008137, recipient P.Ch. I.); and the W.M. Keck Foundation (recipient, P.Ch. I.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

cities [2, 3], clusters in DNA structure [4, 5], financial markets [6–8], spatial distribution of words in texts [9–13], structures of rocks and geological formations [14], citation networks [15], social interactions [16], and stochastic physical systems [17, 18]. Zipf's and Heaps' law, two classical representatives of scaling laws, have been studied in various systems [19–22]. In the context of natural language structure and organization, Zipf's law indicates the inversely proportional relationship in log-log scale between the descending words frequency $Z(r)$ and the words frequency rank r [23]. Heaps' law reveals a different aspect of words organization in natural languages, which indicates that the vocabulary size $N(t)$ of a given text grows roughly as a sub-linear function in log-log plot with increasing text length t [24]. The initial Zipf's study focused on the English language [23], and later researchers extended his work to other natural languages, including Hebrew, Greek, Spanish and Irish [25, 26], reporting that different languages are characterized by distinct scaling exponents, reflecting differences in words organization.

Although Chinese is a widely used language, relatively few studies have focused on the Chinese lexical organization [21, 27], and the laws which govern words frequency, words rank and growth of distinct words with increasing text length in the Chinese language remain not well understood. The existence of Chinese polysemic words and the complex word segmentation rules in the Chinese language pose challenges to systematic and consistent studies [28]. Currently, it is an open question whether Zipf's law, Heaps' law or other scaling laws adequately describe the structure and organization of the Chinese language system.

In this study, we analyze scaling properties of words organization in the Chinese language and compare with English, using a database of classic Chinese and English books. Our analyses show that: (i) the probability distribution of words frequency in the Chinese and English language obeys a power-law, and the word organization of both languages conforms to Zipf's law and Heaps' law; (ii) the scaling exponents and crossover behavior of these two languages are significantly different, reflecting the different role and organization of Chinese characters compared to English words. Furthermore, we propose a model of words organization and text growth mechanism which accounts for the empirically observed scaling laws. The introduced model parameters provide new insight into the scaling dynamics and construction mechanisms of words organization in the Chinese and English written texts. By varying key model parameters, we successfully reproduce the differences in the organization and scaling properties of words in Chinese and English texts, and we establish functional relationship between the empirically observed scaling exponents and our model parameters.

Methods and Results

Database

We compare two categories of data sets obtained from Project Gutenberg [29]:

1. Ten classic Chinese books written by different authors during the period from the 14th to late 18th century, including novels, Chinese mythologies and history texts, where each book has approximately 0.4×10^4 distinct Chinese characters and on average 5×10^5 total number of Chinese characters per book (a Chinese character can appear many times in a book). In our analyses, we treat each Chinese character as a separate word because in contrast to western languages where each word is composed of letters, characters in the Chinese language do not correspond to letters but often indicate separate words, and the same Chinese character can play role as a verb, noun, or adverb depending on the context in the sentence.
2. Ten classic English books written by different authors covering different themes and genres. Each book contains approximately 10^4 distinct words and 10^5 text length, a corpus of words comparable in size with the database of Chinese classic books analyzed in this paper.

Table 1. Detailed information on the database of Chinese and English books and corresponding words statistics for each book. T is the total number of words, and N_T is the vocabulary size of each book.

Language	No.	Book Title	T	N_T
Chinese	01	Flowers in the Mirror 镜花缘 (by <i>Ruzhen Li</i> 李汝珍)	338469	4229
	02	The Scholars 儒林外史 (by <i>Jingzi Wu</i> 吴敬梓)	231950	3338
	03	Sigong'an 施公案 (by <i>Anonymous</i> 佚名)	967788	4077
	04	Amazing Tales 二刻拍案惊奇 (by <i>Mengchu Ling</i> 凌初)	52559	4241
	05	Yuewei Cottage Notes 阅微草堂笔记 (by <i>Xiaolan Ji</i> 纪晓岚)	307714	4980
	06	Penggong'an 彭公案 (by <i>Tan Meng Dao Ren</i> 贪梦道人)	732458	3481
	07	Stories to Caution the World 警世通言 (by <i>Menglong Feng</i> 冯梦龙)	312085	4333
	08	Journey to the West 西游记 (by <i>Chengen Wu</i> 吴承恩)	589705	4483
	09	Dream of Red Mansions 红楼梦 (by <i>XueQin Cao</i> 曹雪芹)	730112	4636
	10	Romance of the Three Kingdoms 三国演义 (by <i>Guangzhong Luo</i> 罗贯中)	485914	3950
English	01	Alices adventures in wonderland (by <i>Lewis Carroll</i>)	30083	4285
	02	Ulysses (by <i>James Joyce</i>)	268843	29020
	03	Hamlet (by <i>William Shakespeare</i>)	218510	16963
	04	David Crockett (by <i>John S. C. Abbott</i>)	78802	7343
	05	A Christmas carol (by <i>Charles Dickens</i>)	57314	4778
	06	The adventures of Tom Sawyer (by <i>Mark Twain</i>)	27341	2572
	07	Moby-Dick; Or, the Whale (by <i>Herman Melville</i>)	59909	5546
	08	The hound of the Baskerville (by <i>Sir Arthur Conan Doyle</i>)	73908	7172
	09	An enquiry concerning human understanding (by <i>David Hume</i>)	30979	4834
	10	The origin of species by means of natural selection (by <i>Charles Darwin</i>)	156815	6930

doi:10.1371/journal.pone.0168971.t001

Detailed information on the books included in our analyses, their length and the vocabulary size of each book is shown in [Table 1](#).

Fundamental scaling laws of words organization in Chinese language

We first obtain the distribution of word frequency for each Chinese and English language book in our database. We find that the probability distribution $P(k)$ of word frequency k for both Chinese and English books follows a power-law with scaling exponent β ([Fig 1\(a\)](#)),

$$P(k) \sim k^{-\beta}. \tag{1}$$

Our analyses show that Chinese language texts exhibit higher percentage of high frequency words compared to English texts. This is reflected by the significantly lower value of the scaling exponent $\beta = 1.55 \pm 0.06$ (group mean \pm standard deviation) for Chinese books, compared to $\beta = 1.83 \pm 0.04$ for English books with Student's t -test $p < 0.01$, showing statistically significant difference ([Fig 2](#), fitting range $k \in [1, 10^3]$).

We next perform Zipf's rank analysis, and we find that words organization in both Chinese and English language texts obeys the Zipf's law, i.e. the normalized word frequency $Z(r)$ exhibits a power-law behavior as a function of the word frequency rank r characterized by a scaling exponent α ,

$$Z(r) \sim r^{-\alpha}. \tag{2}$$

Our analyses show that while English texts exhibit a power-law with a single exponent $\alpha = 1.05 \pm 0.03$ for the entire fitting range of words frequency rank $r \in [3, 2 \times 10^3]$, the Chinese language texts are characterized by a clear crossover in the Zipf's scaling from regime with $\alpha_1 = 0.60 \pm 0.07$ for high and intermediate frequency ranks $r \in [3, 100]$ to a second regime with

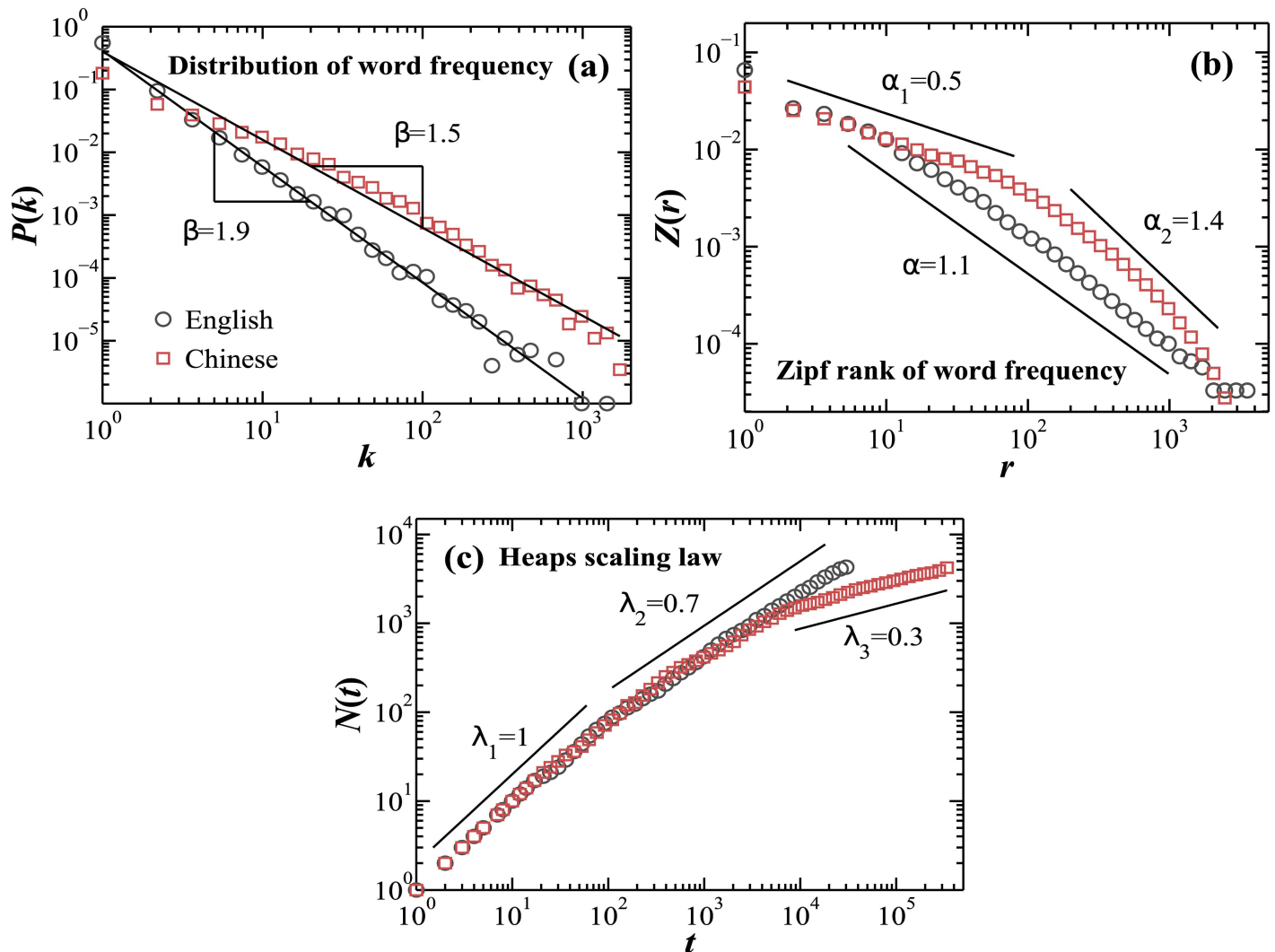


Fig 1. Scaling analyses of words organization in Chinese and English books. Words organization in Chinese and English language exhibits scaling laws with different characteristics. Log-log plots of (a) probability distribution $P(k)$ of the word frequency k , (b) Zipf's law $Z(r)$ of the word frequency rank r , and (c) Heaps' scaling law of the number of distinct words $N(t)$ vs. the number of words t in the text, obtained for the first book in Chinese and in English language listed in Table 1. Straight lines indicate the fitting range where the scaling exponents β , α and λ are obtained. Our analyses show that (a) Chinese language books exhibit lower exponent β compared to English books; (b) while English books exhibit a single scaling regime over the entire range of frequency ranks r , Chinese texts are characterized by a clear crossover in the Zipf's scaling of the normalized word frequency $Z(r)$ vs. word frequency rank r ; (c) the number of distinct words $N(t)$ vs. text length t exhibits a crossover with two different scaling exponents at small and intermediate scales for both Chinese and English books. However, Chinese texts are characterized by a third saturation regime for large scales t that is not observed for English books.

doi:10.1371/journal.pone.0168971.g001

scaling exponent $\alpha_2 = 1.48 \pm 0.14$ for lower frequency ranks $r \in [2 \times 10^2, 2 \times 10^3]$ shown in Fig 1(b) (with Student's t -test $p < 0.01$, showing significant differences when comparing α_1 and α_2 of Chinese language texts, as well as when comparing the scaling exponents between the Chinese and English language, Fig 2). In Table 2, we list the top 20 most frequently used English words and Chinese characters and their frequencies.

We also find that words organization of both Chinese and English language texts obeys the Heaps' law. Specifically, we find the number of distinct words $N(t)$ grows as a power-law with increasing text length t for all Chinese and English books in our database (Fig 1(c)):

$$N(t) \sim t^\lambda. \tag{3}$$

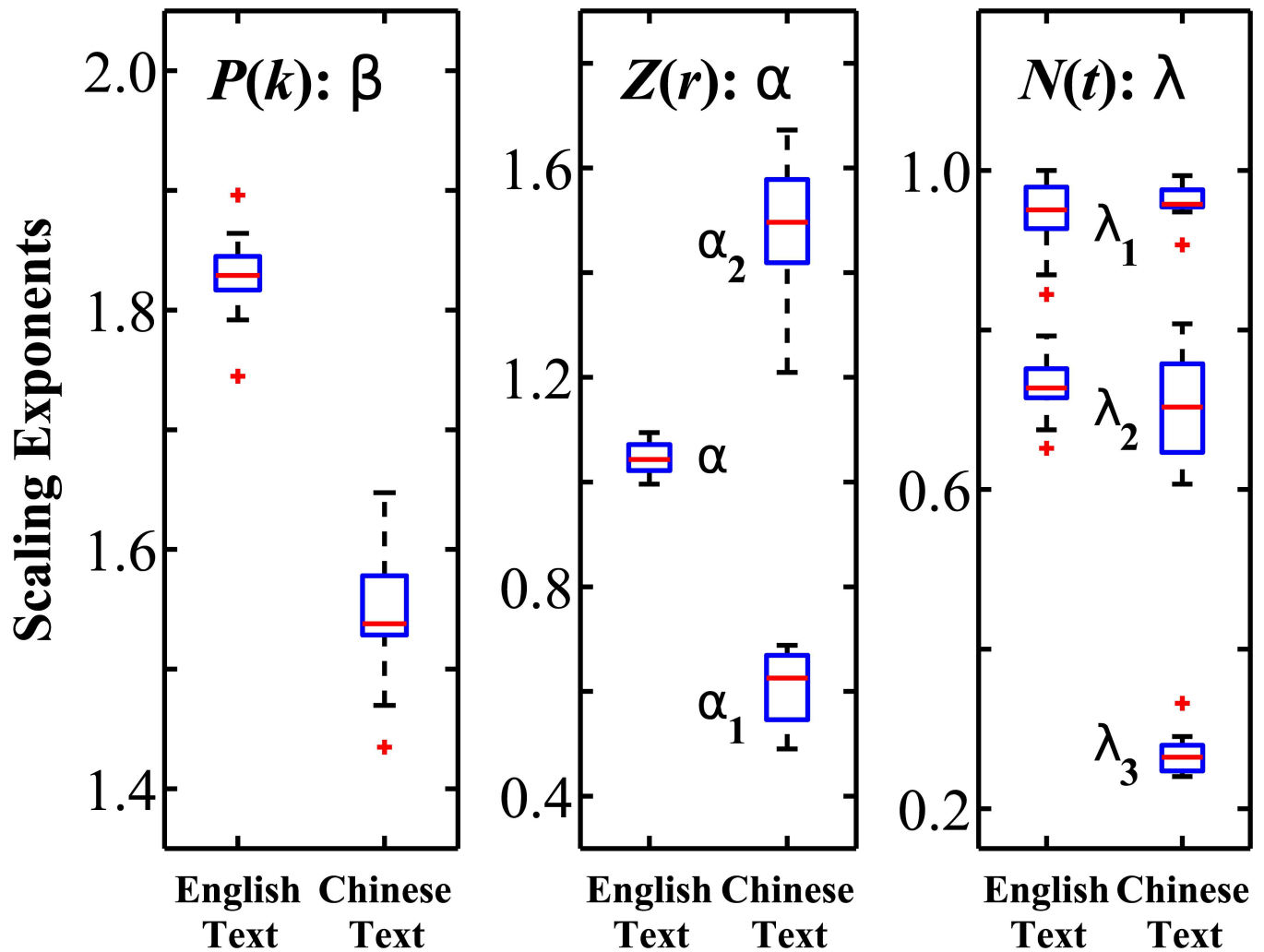


Fig 2. Statistical comparison of the scaling exponents derived from different scaling laws. Data show the group averaged values obtained for all books analyzed in Table 1. The central red mark in each box indicates the median value for the corresponding scaling exponents β , α and λ shown in three panels. Bottom and upper edges of each box denote the 25th and 75th percentiles correspondingly, while the whiskers extend to the most extreme data points not considered outliers; outliers are plotted individually by the symbol “+” in red color. The exponent β of the probability distribution $P(k)$ is obtained in the range of word frequency $k \in [1, 10^3]$ for both Chinese and English texts. The Zipf’s law scaling exponent α for English books is obtained in the range of word frequency rank $r \in [3, 2 \times 10^3]$, while for Chinese books α_1 is obtained for small and intermediate scales $r \in [3, 100]$ and α_2 for large scales $r \in [2 \times 10^2, 2 \times 10^3]$. Heaps’ scaling law for both languages exhibits several scaling regimes with the exponent λ_1 obtained for short scales of text length of $t \in [1, 100]$ for both languages, λ_2 in the scales of $t \in [10^2, 2 \times 10^4]$ for English texts and $t \in [10^2, 3 \times 10^3]$ for Chinese texts, and the exponent λ_3 in long scales of $t > 10^4$ for Chinese texts only. Our analyses indicate significant differences in all three scaling exponents β , α and λ between written texts in the Chinese and English language.

doi:10.1371/journal.pone.0168971.g002

Our results show that at short scales $t \in [1, 10^2]$ both Chinese and English texts exhibit practically the same scaling behavior with exponent $\lambda_1 \approx 1$ (Fig 1(c)) with: $\lambda_1 = 0.96 \pm 0.02$ for Chinese language books and $\lambda_1 = 0.94 \pm 0.05$ for English (Student’s t -test indicates no statistical difference with $p = 0.34$). To estimate the range of this linear growth regime for each Chinese and English texts, we determine the scale t_1 at which the scaling exponent λ reaches 0.95, and we find significant difference in the range of linear growth between the Chinese and English texts ($t_1 = 130.2 \pm 43.2$ for Chinese and $t_1 = 78.6 \pm 16.2$ for English, with $p < 0.01$).

Further, we find that the Heap’s scaling law exhibits a crossover at $t \approx 100$ from a linear to a sub-linear scaling regime characterized by a scaling exponent $\lambda_2 \approx 0.7$ for the intermediate

Table 2. Top 20 most frequently used English words and Chinese characters and their frequencies. A Chinese character can have different functions in the structure of a sentence and carry different meanings depending on the context, as shown in brackets following each Chinese character in the table. The frequencies are calculated using pooled data of all books in our database.

Rank <i>r</i>	English		Chinese	
	Word	Frequency $Z(r)$	Character	Frequency $Z(r)$
1	the	0.0595	一(a, one, whole)	0.0174
2	of	0.0328	人(human, fellow)	0.0163
3	and	0.0304	了(end, finish, understand)	0.0155
4	to	0.0223	不(no, non-, without)	0.0154
5	a	0.0219	来(come, ever since)	0.0121
6	in	0.0187	说(say, persuade, theory)	0.0110
7	that	0.0130	道(path, say, doctrine, Tao)	0.0106
8	I	0.0115	是(is, correct, this, yes)	0.0106
9	it	0.0115	的(of, target)	0.0102
10	he	0.0107	有(have, exist)	0.0084
11	his	0.0096	他(he, other)	0.0083
12	was	0.0084	我(I)	0.0081
13	with	0.0081	在(at, in, be, exist)	0.0074
14	as	0.0072	你(you)	0.0071
15	for	0.0071	这(this, now)	0.0071
16	is	0.0070	去(go, leave, remove)	0.0071
17	on	0.0064	大(big, important, old)	0.0070
18	you	0.0061	个(a, individual, size)	0.0069
19	but	0.0058	上(up, good, superior, on)	0.0062
20	be	0.0057	见(see, meet with, opinion)	0.0061

doi:10.1371/journal.pone.0168971.t002

and large scales: group average $\lambda_2 = 0.70 \pm 0.07$ for Chinese texts in the range of $t \in [100, 3 \times 10^3]$, and $\lambda_2 = 0.73 \pm 0.04$ for English in the full range of intermediate and large scales $t \in [100, 2 \times 10^4]$ (Fig 1(c)). Student's t -test $p = 0.28$ indicates no significant difference between Chinese and English language in this intermediate scaling regime.

Notably, our analyses show that all Chinese texts exhibit a second crossover at large scale $t \approx 4 \times 10^3$, corresponding to $N(t) \approx 2 \times 10^3$ with a transition to a saturation regime characterized by a distinct scaling exponent $\lambda_3 = 0.27 \pm 0.03$. This saturation regime is not observed in the English texts, but is typical for Chinese texts (Fig 1(c)) and results from the limited number of distinct Chinese characters (considered as words in our analyses) that are predominantly used in Chinese language texts.

Stochastic feedback model of words organization and text growth mechanism in Chinese language

To understand the mechanism underlying words organization leads to the empirically observed scaling laws and study the difference between Chinese and English languages, we introduce a stochastic feedback model that accounts for the probability of word occurrence and growth of new word (that has not appeared in the text yet) with increasing text length.

To guide our model construction, we perform an additional statistical analysis on all books in our database to determine how the frequency of words in a new part of the text depends on the frequency of words in the previous part of the text. First, we divide a text into two equal parts (Part I and Part II), and calculate the frequency k for each distinct word in Part I, then we count $n(k)$ the number of distinct words that appears k times in Part I. We next count the total

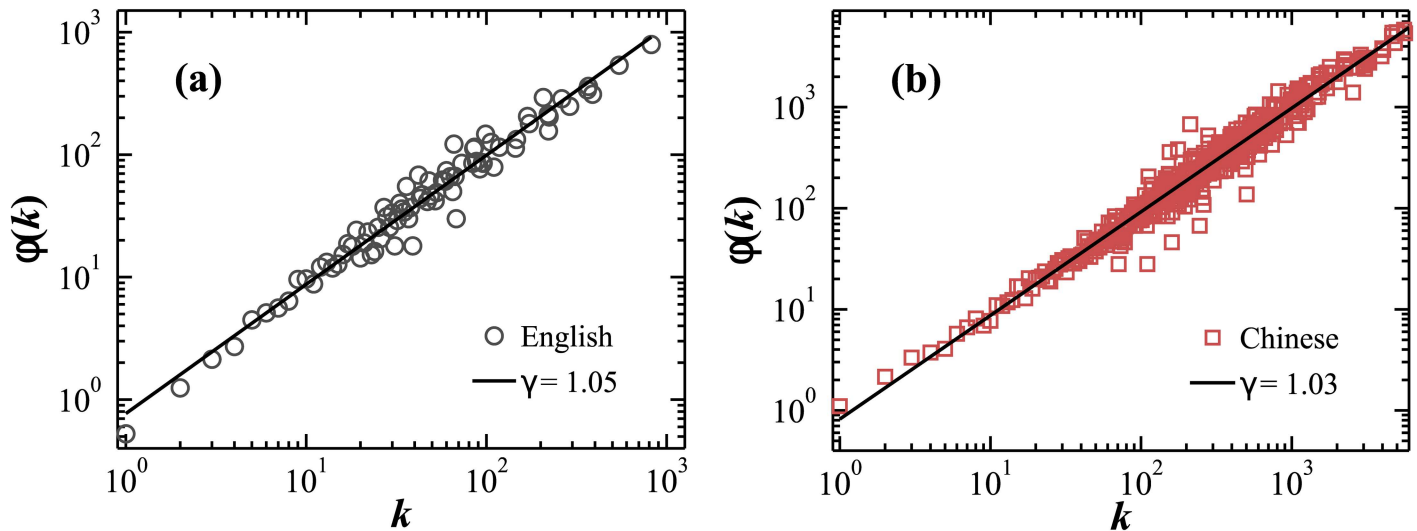


Fig 3. Empirical analyses of the word growth mechanism. Data show results for the first book of (a) English and (b) Chinese language listed in Table 1, revealing a scaling relation of the average number of occurrence $\phi(k)$ of a given word in the second half of a text, provided the frequency of occurrence of this word in the first half of the text is k . Both languages are characterized by a scaling exponent $\gamma \approx 1$, indicating that words which appear with high frequency k in the first part of the text have also high-average occurrence in the rest of the text.

doi:10.1371/journal.pone.0168971.g003

number of times $N(k)$ when all these $n(k)$ words in Part I would also appear in Part II of the text. Finally, we calculate $\phi(k)$ the average number of occurrence in Part II of words that appeared k times each in Part I:

$$\phi(k) = N(k)/n(k). \tag{4}$$

For both Chinese and English language books in Table 1, we find that $\phi(k)$ scales with k with an exponent $\gamma \approx 1$ (Fig 3):

$$\phi(k) \sim k^\gamma. \tag{5}$$

Hence, this analysis reveals a “rich-get-richer” mechanism in words growth, i.e., high frequency word in the “old” text tends to appears with similar high frequency in the “new” text. This behavior is consistently observed in all Chinese and English books in our database, and thus we introduce in our modeling approach this empirically derived functional relation for the organization of distinct words in written text.

To understand the origin of the observed scaling laws describing words structure in Chinese language texts, we simulate the process of language construction using the following stochastic feedback model of text growth. Given a text with a number of words, we grow the text following two different procedures:

Procedure I: with probability p we generate a new word that does not appear in the previous text and we add this word to the new text,

or

Procedure II: with probability $1 - p$ we select a word that is already present in the previous text and we add it to the new text.

Further, to model text growth dynamics based on the process of generating new distinct words that are not present in the prior text (Procedure I), we introduce a probability p that

gradually changes with text length t :

$$p = k_0 t^{-k_t}, \tag{6}$$

where t is the growing text length. In our model simulation, for each book, we generate data length $t = T$ that equals to the length of the corresponding empirical texts in Table 1.

At short text length t , the functional form of p in Eq (6) provides for a close to linear growth of the number of distinct new words. The values of the parameter $k_0 > 1$ determine the range of the linear growth region where the text growth process includes only new words that are not present in the prior text. With text length t increasing, the probability p of adding distinctly new words to the text decreases, which is controlled by the scaling parameter $k_t > 1$ for Chinese texts in order to account for the saturation regime observed in the Heaps' scaling law for very large scales of t shown in Fig 1(c) (for English texts the simulation parameter $k_t \in (0, 1)$).

In parallel to the text growth process that involves generating new distinct words with probability p (Eq 6), natural language text growth in our model involves also a second process of adding words that are already present in the prior text (Procedure II) with probability $1 - p$. Guided by the empirical observations shown in Fig 3, in this procedure, our model incorporates a probability $\tilde{p}(i)$ to add a word i that is already in the text, where $\tilde{p}(i)$ has a positive dependence on the number of times $n(i)$ that the word i already appears in the text:

$$\tilde{p}(i) = \frac{n(i)^{k_p}}{\sum_i n(i)^{k_p}}. \tag{7}$$

The values of the parameter k_p determine the degree of dependence between the probability of a word being selected and its number of occurrence in the already generated text.

Fig 4 shows the scaling analyses results from our model simulation outputs, where model parameters were chosen to reproduce the empirical scaling properties we found for the first book in Table 1 in Chinese and English language. All fundamental scaling laws for $P(k)$, $Z(r)$ and $N(t)$ with their specific scaling regimes are very well captured by our model. Moreover, our separate simulations for each book in the database reveal that:

1. There is a significant difference in the model parameter k_0 values between the Chinese ($k_0 = 5.51 \pm 1.56$, group ave. \pm std. dev.) and English language texts ($k_0 = 2.93 \pm 1.01$) (Table 3). This corresponds to our empirical findings that Chinese and English language have different linear growth regimes in $N(t)$.
2. The model parameter k_t for Chinese texts has significantly higher values compared to English texts: $k_t = 1.16 \pm 0.17$ for Chinese and $k_t = 0.33 \pm 0.05$ for English, indicating a much lower growth rate of new distinct words with increasing text length in the Chinese language. Scatter plot diagram of the model parameter k_t vs. the empirical scaling exponent λ shown in Fig 5(b) for each Chinese and English book in our database reveals an exponential relationship between k_t and λ : $\lambda = e^{-1.2k_t}$.
3. Chinese language texts have significantly lower values of the model parameter k_p compared to English texts ($k_p = 1.02 \pm 0.02$ for Chinese books and $k_p = 1.10 \pm 0.02$ for English books, Table 3), indicating that Chinese texts exhibit weaker dependence between the probability $\tilde{p}(i)$ of a word i being selected and added to the text and its number of occurrence $n(i)$ in prior text, which accounts for the scaling exponent β of the power-law probability distribution $P(k)$. Scatter plot diagram of the model parameter k_p vs. the empirically derived scaling exponent β for each Chinese and English book in our database reveals a linear functional dependence between k_p and β : $\beta = 3.5k_p - 2$ (Fig 5(a)).

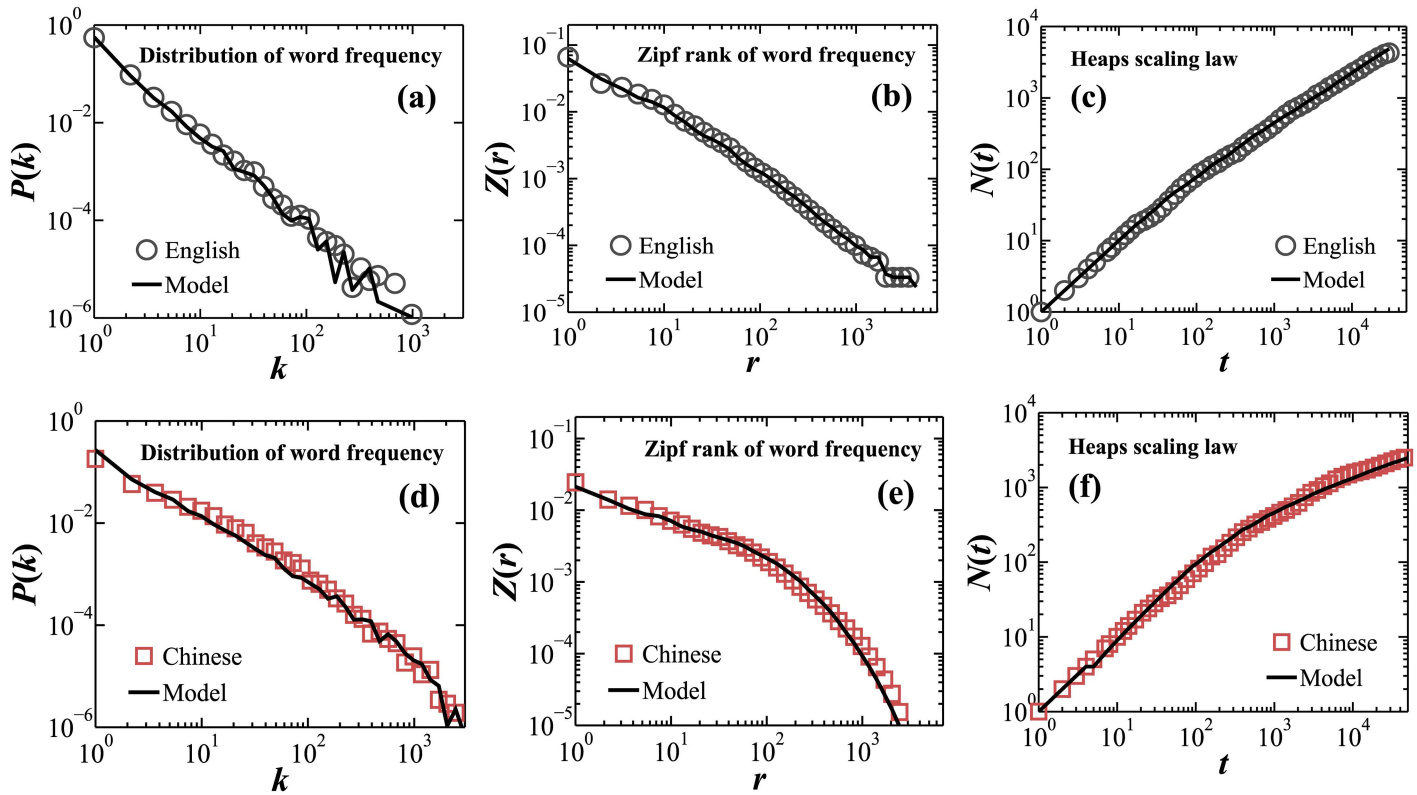


Fig 4. Empirical results and modeling simulation for Chinese and English language books. Scaling laws and model simulations for English book No. 1 (Table 1) are shown in panels (a), (b) and (c), and for Chinese book No. 1 (Table 1) are shown in panels (d), (e) and (f). Modeling parameters for all Chinese and English language books are given in Table 3.

doi:10.1371/journal.pone.0168971.g004

Table 3. Model parameters of Chinese and English books. The statistics show significant differences in the model parameters k_0 , k_t and k_p between Chinese and English texts, indicating differences in the dynamic process underlying the language structure, words organization and the occurrence of new words with text growth.

No.	k_0		k_t		k_p	
	Chinese	English	Chinese	English	Chinese	English
01	4.68	2.34	1.10	0.30	1.01	1.14
02	4.05	2.16	1.05	0.26	1.03	1.09
03	5.51	2.76	1.22	0.30	1.00	1.10
04	4.60	2.85	1.09	0.32	1.04	1.12
05	5.50	3.28	1.16	0.36	1.02	1.09
06	5.95	3.11	1.27	0.38	1.01	1.08
07	4.16	2.79	1.03	0.33	1.00	1.10
08	5.17	2.54	1.15	0.31	1.01	1.11
09	5.93	1.94	1.23	0.27	1.05	1.14
10	9.52	5.54	1.48	0.44	1.01	1.07
group mean	5.51	2.93	1.18	0.33	1.02	1.10
± std. dev.	± 1.56	± 1.01	± 0.13	± 0.05	± 0.02	± 0.02
Student's <i>t</i> -test	$p < 0.01$		$p < 0.01$		$p < 0.01$	

doi:10.1371/journal.pone.0168971.t003

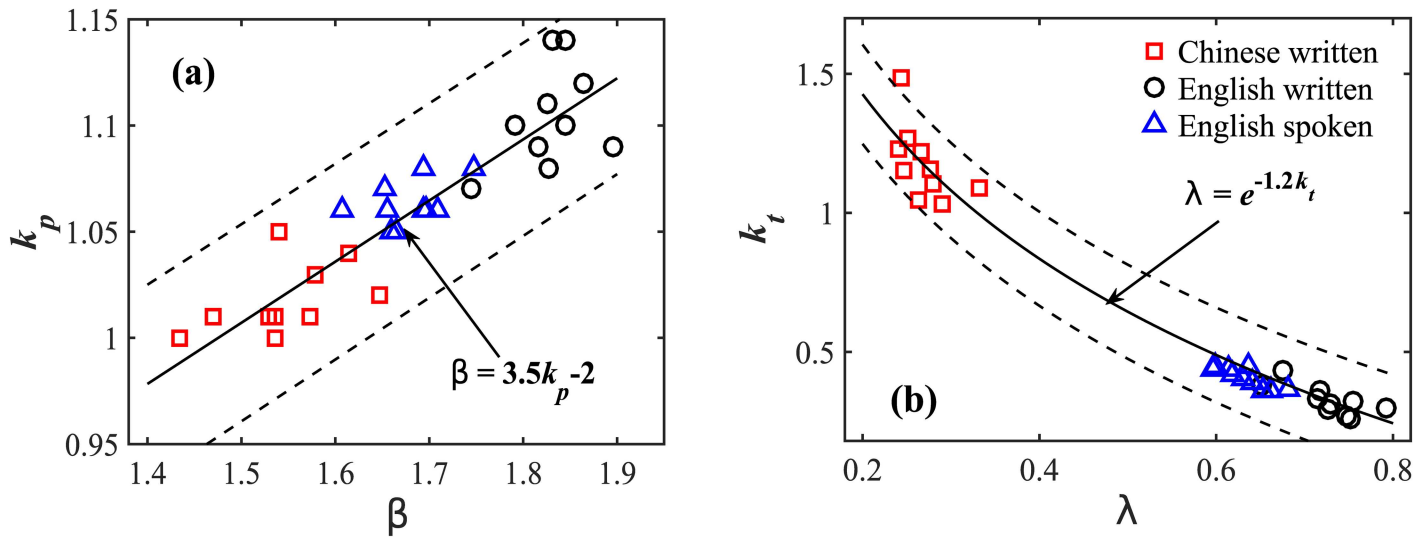


Fig 5. Functional relations between the empirically observed scaling exponents and model parameters. (a) Exponent β of the empirical probability distribution $P(k)$ vs. model parameter k_p indicating linear functional dependence. (b) Heaps' law scaling exponents λ_2 for English books and spoken transcriptions, and λ_3 for Chinese books vs. model parameter k_t indicating exponential functional dependence. Data points are obtained from the scaling analyses and simulation of all ten Chinese and English language books listed in Table 1, and English spoken language from Ref. [30]. The dotted lines indicate 95% confidence intervals of the data points obtained from empirical and model parameters for each separate book.

doi:10.1371/journal.pone.0168971.g005

Note that, in Fig 5 we also include the results for English spoken language we analyzed in a previous study [30]. The data points obtained from English spoken language fall nicely on the fitting curves, which further validate the functional relation we establish in this study between the empirically observed scaling exponents and the model parameters.

Conclusion

Our analyses of Chinese and English books indicate that both language forms obey power-law in the probability distribution of word frequency, as well as Zipf's law and Heaps' law, however with different scaling characteristics. Specifically, words organization in the Chinese language (i) is characterized by a significantly lower exponent for the probability distribution of words frequency, (ii) exhibits a pronounced crossover in the Zipf's scaling law of words frequency rank, and (iii) obeys the Heaps' scaling law exhibiting a similar growth rate in the number of new distinct words with increasing text length for short and intermediate scales compared to English texts, however with a unique for the Chinese language saturation regime when the number of distinct words (Chinese characters) reaches about 4000. The total number of Chinese vocabulary is approximately 90 thousand, and the number of commonly used Chinese characters is about 4500, which accounts for the Heaps' law saturation regime at long text length. In contrast, the vocabulary size of the English language [31] is much larger than that of the Chinese language, thus the average text length of English books may not be sufficient to manifest a saturation regime as observed for Chinese language books. Our empirical findings and model simulations indicate a clear difference in the dynamics of language construction process between Chinese and English language forms, associated with higher concentration of high frequency words and lower occurrence of distinct new words in Chinese texts. The presented model successfully accounts for the different lexical generation mechanisms in Chinese and English. Our simulation results on words generation and text growth are in agreement with the empirical findings of three distinct scaling laws, and confirm that these laws

accurately represent the complex dynamics of words organization in the Chinese and English language. The proposed here stochastic feedback model for text generation can be generally applied to other dynamic growth processes in complex systems characterized by scaling laws.

Acknowledgments

We acknowledge the support from the National Natural Science Foundation of China (61201029, 61271082), the Fundamental Research Funds for the Central Universities of China, the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), the Nanjing University Scholar Exchange Project, the National Institutes of Health (NIH Grant 1R01-HL098437), the US-Israel Binational Science Foundation (BSF Grant 2008137) and the W.M. Keck Foundation.

Author Contributions

Conceptualization: CB QDYM PChI.

Data curation: SL RL.

Formal analysis: SL RL.

Funding acquisition: CB QDYM PChI.

Investigation: SL RL CB QDYM.

Methodology: CB QDYM PChI.

Project administration: CB.

Resources: QDYM CB.

Software: SL RL.

Supervision: PChI.

Validation: CB QDYM.

Visualization: SL CB QDYM PChI.

Writing – original draft: SL CB QDYM PChI.

Writing – review & editing: CB QDYM PChI.

References

1. Makse HA, Havlin S, Stanley HE. Modelling urban growth. *Nature*. 1995; 377(1912):779–782.
2. Blank A, Solomon S. Power laws in cities population, financial markets and internet sites (scaling in systems with a variable number of components). *Physica A*. 2000; 287(1-2):279–288. doi: [10.1016/S0378-4371\(00\)00464-7](https://doi.org/10.1016/S0378-4371(00)00464-7)
3. Newman MEJ. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*. 2005; 46(5):323–351. doi: [10.1080/00107510500052444](https://doi.org/10.1080/00107510500052444)
4. Arneodo A, Bacry E, Graves PV, Muzy JF. Characterizing long-range correlations in DNA sequences from wavelet analysis. *Physical Review Letters*. 1995; 74(16):3293–3296. doi: [10.1103/PhysRevLett.74.3293](https://doi.org/10.1103/PhysRevLett.74.3293) PMID: [10058160](https://pubmed.ncbi.nlm.nih.gov/10058160/)
5. Hackenberg M, Rueda A, Carpena P, Bernaola-Galván P, Barturen G, Oliver JL. Clustering of DNA words and biological function: a proof of principle. *Journal of Theoretical Biology*. 2012; 297:127–136. doi: [10.1016/j.jtbi.2011.12.024](https://doi.org/10.1016/j.jtbi.2011.12.024) PMID: [22226985](https://pubmed.ncbi.nlm.nih.gov/22226985/)
6. Levy M, Solomon S. New evidence for the power-law distribution of wealth. *Physica A*. 1997; 242(1-2):90–94. doi: [10.1016/S0378-4371\(97\)00217-3](https://doi.org/10.1016/S0378-4371(97)00217-3)

7. Romero NE, Ma QDY, Liebovitch LS, Brown CT, Ivanov PCh. Correlated walks down the Babylonian markets. *EPL (Europhysics Letters)*. 2010; 90(1):18004. doi: [10.1209/0295-5075/90/18004](https://doi.org/10.1209/0295-5075/90/18004)
8. Ivanov PCh, Yuen A, Perakakis P. Impact of stock market structure on intertrade time and price dynamics. *PloS one*. 2014; 9(4):e92885. doi: [10.1371/journal.pone.0092885](https://doi.org/10.1371/journal.pone.0092885) PMID: [24699376](https://pubmed.ncbi.nlm.nih.gov/24699376/)
9. Ebeling W, Neiman A. Long-range correlations between letters and sentences in texts. *Physica A*. 1995; 215(3):233–241. doi: [10.1016/0378-4371\(95\)00025-3](https://doi.org/10.1016/0378-4371(95)00025-3)
10. Putnins T, Signoriello DJ, Jain S, Berryman MJ, Abbott D. Advanced text authorship detection methods and their application to biblical texts. *Proc SPIE 6039, Complex Systems*. 2005; p. 60390J.
11. Berryman MJ, Allison A, Abbott D. Statistical techniques for text classification based on word recurrence intervals. *Fluctuation and Noise Letters*. 2003; 03(01):L1–L10. doi: [10.1142/S021947750300104X](https://doi.org/10.1142/S021947750300104X)
12. Carpena P, Bernaola-Galván P, Hackenberg M, Coronado AV, Oliver JL. Level statistics of words: Finding keywords in literary texts and symbolic sequences. *Physical Review E*. 2009; 79(3):035102. doi: [10.1103/PhysRevE.79.035102](https://doi.org/10.1103/PhysRevE.79.035102) PMID: [19392005](https://pubmed.ncbi.nlm.nih.gov/19392005/)
13. Ortuno M, Carpena P, Bernaola-Galván P, Munoz E, Somoza AM. Keyword detection in natural languages and DNA. *EPL (Europhysics Letters)*. 2002; 57(5):759–764. doi: [10.1209/epl/i2002-00528-3](https://doi.org/10.1209/epl/i2002-00528-3)
14. Makse HA, Davies GW, Havlin S, Ivanov PCh, King PR, Stanley HE. Long-range correlations in permeability fluctuations in porous rock. *Physical Review E*. 1996; 54(4):3129–3134. doi: [10.1103/PhysRevE.54.3129](https://doi.org/10.1103/PhysRevE.54.3129) PMID: [9965453](https://pubmed.ncbi.nlm.nih.gov/9965453/)
15. Golosovsky M, Solomon S. Stochastic dynamical model of a growing citation network based on a self-exciting point process. *Physical Review Letters*. 2012; 109(9):098701. doi: [10.1103/PhysRevLett.109.098701](https://doi.org/10.1103/PhysRevLett.109.098701) PMID: [23002894](https://pubmed.ncbi.nlm.nih.gov/23002894/)
16. Perc M, Gomez-Gardenes J, Szolnoki A, Floria LM, Moreno Y. Evolutionary dynamics of group interactions on structured populations: a review. *Journal of The Royal Society Interface*. 2013; 10(80):20120997. doi: [10.1098/rsif.2012.0997](https://doi.org/10.1098/rsif.2012.0997) PMID: [23303223](https://pubmed.ncbi.nlm.nih.gov/23303223/)
17. Solomon S, Levy M. Spontaneous scaling emergence in generic stochastic systems. *International Journal of Modern Physics C*. 1996; 07(05):745–751. doi: [10.1142/S0129183196000624](https://doi.org/10.1142/S0129183196000624)
18. Ivanov PCh, Amaral LAN, Goldberger AL, Stanley HE. Stochastic feedback and the regulation of biological rhythms. *Europhysics Letters (EPL)*. 1998; 43(4):363–368. doi: [10.1209/epl/i1998-00366-3](https://doi.org/10.1209/epl/i1998-00366-3) PMID: [11542723](https://pubmed.ncbi.nlm.nih.gov/11542723/)
19. Solomon S, Richmond P. Stability of Pareto-Zipf Law in non-stationary economies. In: Kirman A, Zimmermann JB, editors. *Economics with Heterogeneous Interacting Agents*. vol. 503 of *Lecture Notes in Economics and Mathematical Systems*. Springer Berlin Heidelberg; 2001. p. 141–159. doi: [10.1007/978-3-642-56472-7_10](https://doi.org/10.1007/978-3-642-56472-7_10)
20. Font-Clos F, Boleda G, Corral Á. A scaling law beyond Zipf's law and its relation to Heaps' law. *New Journal of Physics*. 2013; 15:093033. doi: [10.1088/1367-2630/15/9/093033](https://doi.org/10.1088/1367-2630/15/9/093033)
21. Lü L, Zhang ZK, Zhou T. Zipf's law leads to Heaps' law: Analyzing their relation in finite-size systems. *PloS One*. 2010; 5:e14139. doi: [10.1371/journal.pone.0014139](https://doi.org/10.1371/journal.pone.0014139) PMID: [21152034](https://pubmed.ncbi.nlm.nih.gov/21152034/)
22. Van Leijenhorst DC, Van der Weide TP. A formal derivation of Heaps' Law. *Information Sciences*. 2005; 170:263–272. doi: [10.1016/j.ins.2004.03.006](https://doi.org/10.1016/j.ins.2004.03.006)
23. Zipf GK. *Human behaviour and the principle of least-effort: an introduction to human ecology*. Addison-Wesley Press; 1949.
24. Heaps HS. *Information retrieval: Computational and theoretical aspects*. Academic Press Inc.; 1978.
25. Hatzigeorgiu N, Mikros G, Carayannis G. Word length, word frequencies and Zipf's law in the Greek language. *Journal of Quantitative Linguistics*. 2001; 8:175–185. doi: [10.1076/jqul.8.3.175.4096](https://doi.org/10.1076/jqul.8.3.175.4096)
26. Petersen AM, Tenenbaum J, Havlin S, Stanley HE. Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports*. 2012; 2:313. doi: [10.1038/srep00313](https://doi.org/10.1038/srep00313) PMID: [22423321](https://pubmed.ncbi.nlm.nih.gov/22423321/)
27. Lü L, Zhang ZK, Zhou T. Deviation of Zipf's and Heaps' laws in human languages with limited dictionary sizes. *Scientific Reports*. 2013; 3:1082. doi: [10.1038/srep01082](https://doi.org/10.1038/srep01082) PMID: [23378896](https://pubmed.ncbi.nlm.nih.gov/23378896/)
28. Shtrikman S. Some comments on Zipf's law for the Chinese language. *Journal of Information Science*. 1994; 20:142–143. doi: [10.1177/016555159402000208](https://doi.org/10.1177/016555159402000208)
29. <http://www.gutenberg.org/>;
30. Bian C, Lin R, Zhang X, Ma QDY, Ivanov PCh. Scaling laws and model of words organization in spoken and written language. *EPL (Europhysics Letters)*. 2016; 113:18002. doi: [10.1209/0295-5075/113/18002](https://doi.org/10.1209/0295-5075/113/18002)
31. The Oxford English Dictionary (Second Edition, 1989) contains 291,500 entries;.