Rank-frequency relation for Chinese characters

W.B. Deng,[1, 2, 3] A.E. Allahverdyan,[1, 4, *] B. Li,[5] and Q. A. Wang[1, 3]

[1]*Laboratoire de Physique Statistique et Systèmes Complexes,*
ISMANS, 44 ave. Bartholdi, Le Mans 72000, France
[2]*Complexity Science Center and Institute of Particle Physics,*
Hua-Zhong Normal University, Wuhan 430079, China
[3]*IMMM, UMR CNRS 6283, Université du Maine, 72085 Le Mans, France*
[4]*Yerevan Physics Institute, Alikhanian Brothers Street 2, Yerevan 375036, Armenia*
[5]*Department of Chinese Literature, University of Heilongjiang, Harbin 150080, China*

We show that the Zipf's law for Chinese characters perfectly holds for sufficiently short texts (few thousand different characters). The scenario of its validity is similar to the Zipf's law for words in short English texts. For long Chinese texts (or for mixtures of short Chinese texts), rank-frequency relations for Chinese characters display a two-layer, hierarchic structure that combines a Zipfian power-law regime for frequent characters (first layer) with an exponential-like regime for less frequent characters (second layer). For these two layers we provide different (though related) theoretical descriptions that include the range of low-frequency characters (hapax legomena). The comparative analysis of rank-frequency relations for Chinese characters versus English words illustrates the extent to which the characters play for Chinese writers the same role as the words for those writing within alphabetical systems.

PACS numbers: 89.75.Fb, 89.75.Da, 05.65.+b

I. INTRODUCTION

Rank-frequency relations provide a coarse-grained view on the structure of a text: one extracts the normalized frequencies of different words $f_1 > f_2 > ...$, orders them in a non-increasing way and studies the frequency f_r as a function of its rank r. One widely known aspect of this rank-frequency relation that holds for texts written in many alphabetical languages is the Zipf's law; see [1–4] for reviews, [5–8] for modern instances of the law, and [9] for extensive lists of references on the subject. This regularity was first discovered by Estoup [10]:

$$f_r \propto r^{-\gamma} \quad \text{with} \quad \gamma \approx 1. \tag{1}$$

The message of a power-law rank-frequency relation is that there is no a single group of dominating words in a text, they rather hold some type of hierarchic, scale-invariant organization. This contrasts to the exponential-like form of the rank-frequency relation that would display a dominant group of words that is representative for the text.

The simple form of the Zipf's law hides the mechanism behind it. Hence there is no consensus on the origin of the law, as witnessed by different theories proposed to explain it [11–17]. An influential group of theories explain the law from certain general premises of the language [11–14], e.g. that the language trades-off between maximizing the information transfer and minimizing the speaking-hearing effort [11], or that the language employs its words via the optimal setting of information theory [12]. The general problem of derivations from this group is that explaining the Zipf's law for the language (and verifying it for a frequency dictionary) does not yet mean to explain the law for a concrete text, where the frequency of the same word varies widely from one text to another and is far from its value in a frequency dictionary.

It was held once that the Zipf's law is not especially informative, since it is recovered by very simple stochastic models, where words are generated through random combinations of letters and space symbol seemingly reproducing the $f_r \propto r^{-1}$ shape of the law [15]. But the reproduction is elusive, since the model is based on features that are certainly unrealistic for natural languages, e.g. it predicts a huge redundancy (many words have the same frequency and length) [18]. More recent opinions reviewed in [19] indicate that the Zipf's law *is* informative and *not* reducible to any trivial statistical regularity. These opinions are confirmed by a recent derivation of the Zipf's law from the ideas of latent semantic analysis [17]. The derivation accounts for generalizations of the Zipf's law for high and low frequencies, and also describes (simultaneously with the Zipf's law) the hapax legomena effect [1]; see Appendix A for the glossary of the used linguistic terms.

However, the Zipf's law was so far found to be absent for the rank-frequency relation of Chinese characters [20–25], which play—sociologically, psychologically and (to some extent) linguistically—the same role for Chinese

[*] Email: armen.allahverdyan@gmail.com

[1] Hapax legomena means literally the set of words that appear in the text only once. We shall employ this term in a broader sense as the set of words that appear few times, so that sufficiently many words have the same frequency. The description of this set is sometimes referred to as the frequency spectrum.

readers and writers as the words do in Indo-European languages [26–28].

Rank-frequency relations for Chinese characters were first studied by Zipf and coauthors who did not find the Zipf's law [29]. They claimed to find another power law with exponent $\gamma = 2$ [29], but this result was later on shown to be incorrect [21], since it was not based on any goodness of fit measure. It was also proposed that the data obtained by Zipf are reasonably fit with a logarithmic function $f_r = a + b\ln(c+r)$ with constant a, b and c [21]. The result on the absence of the Zipf's law was then confirmed by other studies [22–25, 30]. All these authors agree that the proper Zipf's law is absent (more generally a power law is absent), but have different opinions on the (non-power-law) form of the rank-frequency relation for Chinese characters: logarithmic [21], exponential $f_r \propto e^{-dr}$ (where $d > 0$ is a constant) [22–24, 30] or a power-law with exponential cutoff [20, 25]. In [31], the authors describe two different classes of rank-frequency relations for English and Chinese literacy works, they also proposed a model to generate such different situations.

The Zipf's law is regarded as a universal feature of human languages on the level of words [32] [2]. Hence the invalidity of the Zipf's law for Chinese characters has contributed to the ongoing debate on controversies (coming from linguistics and experimental psychology) on whether and to which extent the Chinese writing system is similar to phonological writing systems [36–38]; in particular, to which extent it is based on characters in contrast to words [3].

Results reported in this work amount to the following:

– The Zipf's law holds for sufficiently short (few thousand different characters) Chinese texts written in Classic or Modern Chinese [4]. Short texts are important, because they are building blocks for understanding long texts. For the sake of rank-frequency relations, but also more generally, one can argue that long texts are just mixtures (joining) of smaller, thematically homogeneous pieces. This premise of our approach is fully confirmed by our results.

– The validity scenario of the Zipf's law for short Chinese texts is basically the same as for short English texts [5]: the rank-frequency relation separates into three ranges. *(1)* The range of small ranks (more frequent characters) that contains mostly function characters; we call it the pre-Zipfian range. *(2)* The (Zipfian) range of middle ranks (more probable words) that contains mostly content characters. *(3)* The range of rare characters, where many characters have the same small frequency (hapax legomena).

– The essential difference between Chinese characters and English words comes in for long texts, or upon mixing (joining) different short texts. When mixing different English texts, the range of ranks where the Zipf's law is valid quickly increases, roughly combining the validity ranges of separate texts. Hence for a long text the major part of the overall frequency is carried out by the Zipfian range. When mixing different Chinese texts, the validity range of the Zipf's law increases very slowly. Instead there emerges another, exponential-like regime in the rank-frequency relation that involves a much larger range of ranks. However, the Zipfian range of ranks is still (more) important, since it carries out some 40% of the overall frequency. This overall frequency of the Zipfian range is approximately constant for all (numerous and semantically very different) Chinese texts we studied.

– We describe these two regimes via different (though closely related) theories that are based on the recent approach to rank-frequency relations [17]. This description includes a rather precise theories for rare characters (hapax legomena range) both for long and short Chinese texts.

This work is organized as follows. The next section gives a short introduction to Chinese characters and their differences and similarities with English words. Section III uncovers the Zipf's law for short Chinese texts and compares it with the English situation. Section IV studies the fate of the Zipf's law for long Chinese texts. We summarize in the last section. Appendix A contains the glossary of the used linguistic terms. Appendix B refers to the interference experiments distinguishing between Chinese characters and English words. Appendix C recollects information on the studied Chinese texts. Appendix D lists the key-characters of one studied modern Chinese text. Appendix E reminds the Kolmogorov-Smirnov test that is employed for checking the quality of our numerical fitting.

[2] Applications of the Zipf's law to automatic keyword recognition are based on this fact [33], because keywords are located mostly in the validity range of the Zipf's law. A related set of applications of this law refers to distinguishing between artificial and natural texts, fraud detection [34] *etc*; see [35] for a survey of applications in natural language processing.

[3] We stress already here that the Zipf's law holds for Chinese [25] and Japanese words [39]. This is expected and intuitively follows from the possibility of literal translation from Chinese to English, where (almost) each Chinese word is mapped to an English one (see our glossary at Appendix A for definition of various special terms). In this sense, the validity of the Zipf's law for Chinese words is consistent with the validity of this law for English texts.

[4] The Modern Chinese texts we studied are written with simplified characters, while our Classic Chinese texts are written with traditional characters. Reforms started in the mainland China since late 1940's simplified about 2235 characters. Traditional characters are still used officially in Hong-Kong and Taiwan.

[5] Here and below we refer to a typical Indo-European alphabetical based language as English, meaning that for the sake of the present discussion differences between various Indo-European and/or Uralic languages are not essential. Likewise, we expect that the basic features of the rank-frequency analysis of Chinese characters will apply for those languages (e.g. Japanese), where the Chinese characters are used.

II. CHINESE CHARACTERS VERSUS ENGLISH WORDS

Here we shortly remind the main differences and similarities between Chinese characters and English words; see Footnote 5 in this context. This subject generated several controversies (myths as it was put in [37]), even among expert sinologists [27, 28, 36–38, 40].

This section is not needed for presenting our results (hence it can be skipped upon first reading), but is necessary for a deeper understanding of our results and motivations.

The main qualitative conclusion of this section is that in contrast to English words, Chinese characters have generally more different meanings, they are more flexible, they could combine with other characters to convey different specific meanings. So there are characters, which appear many times in the text, but their concrete meanings are different in different places.

1. The unit of Chinese writing system is the character: a spatially marked pattern of strokes phonologically realized as a single syllable (please consult Appendix A for a glossary of various linguistic terms used in the paper). Generally, each character denotes a morpheme or several different morphemes.

2. The Chinese writing system evolved by emphasizing the concept of the character-morpheme, to some extent blurring the concept of the multi-syllable word. In particular, spaces in the Chinese writing system are put in between of characters and not in between of multi-syllable words [6]. Thus a given sentence can have different meanings when being separated into different sequences of words [40], and parsing a string of Chinese characters into words became a non-trivial computational problem; see [43] for a recent review.

3. Psycholinguistic research shows that the characters are important cognitive and perceptual units for Chinese writers and readers [26–28], e.g. Chinese characters are more directly related to their meanings than English words to their meanings [28] [7]; see Appendix B for additional details. The explanation of this effect would be that characters (compared to English words) are perceived holistically as a meaning-carrying objects, while English words are yet to be reconstructed from a sequence of their constituents (phonemes and syllables) [8].

4. One-character words dominate in the following specific sense. Some 54% of modern Chinese word *tokens* are single-character, two-character word tokens amount to 42%; the remaining words have three or more characters [45]. For modern Chinese word *types* the situation is different: single character words amount to some 10% against 66% of two-character words [45]. Classic Chinese texts have more single-character words (tokens), the percentage varies between some 60% and 80% for texts written in different periods.

The modern Chinese has ≈ 10440 basic (root) morphemes. 93 % of them are represented by single characters. The overall number of Chinese characters is ≈ 18000.

5. A minor part of multi-character words are multi-character morphemes, i.e. their separate characters do not normally appear alone (they are fully bound). Examples of this are the two-character Chinese words for *grape* "葡萄" *(pú táo)*, *dragonfly* "蜻蜓" *(qīng tíng)*, *olive* "橄榄" *(gǎn lǎn)*. Estimates show that some 10% of all characters are fully bound [37].

A related set of examples is provided by two-character words, where the separate characters do have an independent meaning, but this meaning is not directly related to the meaning of the word, e.g. "东西" *(dōng xī)* means *thing*, but literally it amounts to *east-west*, or "手足" *(shǒu zú)* means *close partnership*, but literally *hand-foot.*)

6. The majority of the multi-character words are semantic compounds: their separate characters can stand alone and are related to the overall meaning of the word. Importantly, in most cases, the separate meanings of the component characters are wider than the (relatively unique) meaning of the compound two-character word. An example of this situation is the two-character Chinese word for *train* "火车" *(huǒ chē)*: its first character "火" *(huǒ)* has the meaning of *fire, heat, popular, anger, etc*, while the second character "车" *(chē)* has the meaning of *vehicle, machine, wheeled, lathe, castle, etc.*

Note that in Chinese there is a certain freedom in grouping morpheme into different combinations. Hence it is not easy to distinguish the semantic compounds from lexical phrases.

7. At this point we shall argue that in general Chinese characters have a larger number of different meanings than English words. This statement will certainly appear controversial, if it is taken without proper caution, and is explained without proper usage of linguistic terms (see our glossary at Appendix A); consult Footnote 12 in this context.

[6] An immediate question is whether Chinese readers will benefit from reading a character-written text, where the words boundaries are indicated explicitly. For normal sentences the readers will not benefit, i.e. it does not matter whether the word boundaries are indicated explicitly or not [41]. But for difficult sentences the benefit is there [42].

[7] To get a fuller picture of this effect let us denote $\tau_f(E)$ and $\tau_f(C)$ for English and Chinese phonology activation times, respectively, while $\tau_m(E)$ and $\tau_m(C)$ stand for respective meaning activation times. The phonology activation time is the time passed between seeing a word in English (or character in Chinese) and pronouncing it; likewise, for the meaning activation time. Now these quantities hold [28]: $\tau_f(E) < \tau_m(C) \simeq \tau_f(C) < \tau_m(E)$.

[8] A simpler explanation would be that the characters are perceived as pictograms directly pointing to their meaning. In its literal form this explanation is not correct, since characters-pictograms are not frequent in Chinese [37, 44].

First of all note the difference between polysemes and homographs: polysemes are two related meanings of the same character (word), homographs are two characters (words) that are written in the same way, but their meanings are far from each other [9]. Now many characters are simultaneously homographs and polysemes, e.g. character "明" *(míng)* means *brilliant, light, clear, next, etc.* Here the first three meanings are related and can be viewed as polysemes. The fourth meaning *next* is clearly different from the previous three. Hence this is a homograph. Another example is the character "发" *(fā or fà)* that can mean *hair, send out, fermentation, etc.* All these three meanings are clearly different; hence we have homographs. Note the following peculiarity of the above two examples: the first example is a non-heteronym (homophonic) character, i.e. it is read in the same way irrespectively whether it means *light* or *next.* The second example is a heteronym character: it written in the same way, but is read differently depending on its meaning.

In most cases, heteronym characters—those which are written in the same way, but have different pronunciations—have at least two sufficiently different meanings. The disambiguation of their meaning is to be provided by the context of the sentence and/or the shared experience of the writer and reader [10].

Surely, also English words can be ambiguous in meaning (e.g. *get* means *obtain*, but also *understand = have knowledge*), but there is an essential difference. The major contribution of the meaning ambiguity in English is the polysemy: one word has somewhat different, but also closely related meanings. In contrast, many Chinese characters have widely different meanings, i.e. they are homographs rather than polysemes.

However, we are not aware of any quantitative comparison between homography of Chinese versus English. This may be related to the fact that it is sometimes not easy to distinguish between polysemy and homophony (see the glossary in Appendix A). Still the above statement on Chinese characters having a larger number of different meanings can be quantitatively illustrated via the relative prevalence of heteronyms in Chinese. The amount of heteronyms in English is negligible, e.g. in rather complete list of heteronyms presented in [46], we noted only 74 heteronyms [11], and only three of them had more than 2 meanings. This is a tiny amount of the overall number of English words ($> 5 \times 10^5$). To compare this with the Chinese situation, we note that at least some 14% of modern Chinese and 25% of traditional characters are heteronyms, which normally have at least two widely different meanings. Within the most frequent 5700 modern characters the number of heteronyms is even larger and amounts to 22 % [45] [12].

8. Chinese nouns are generally less abstract: whenever English creates a new word via conceptualizing the existing one, Chinese tends to explain the meaning via using certain basic characters (morphemes). Several basic examples of this scenario include: length=long+short "长短" *(cháng duǎn)*, landscape=mountains+water "山水" *(shān shuǐ)*, adult=big+person "大人" *(dà rén)*, population=person+mouth "人口" *(rén kǒu)*, astronomy=heaven+script "天文" *(tiān wén)*, universe=great+emptiness "太空" *(tài kōng)*. English tools for making abstract words include prefixes, *poly-*, *super-*, *pro-*, *etc* and suffixes, *-tion*, *-ment.* These tools either do not have Chinese analogs, or their usage can generally be suppressed.

English words have inflections to indicate the tense of verbs, the number for nouns or the degree for adjectives. Chinese characters generally do not have such linguistic attributes [13], their role is carried out by the context of the sentence(s) [14].

To summarize this section, the differences between Chinese and English writing systems can be viewed in the context of the two features: emphasizing the role of base (root) morphemes and delegating the meaning to the context of the sentence whenever this is possible [26].

The quantitative conclusion to be drawn from the

[9] Note that polysemes are defined to be related meanings of the same word, while homographs are defined to be different words. This is natural, but also to some extent conventional, e.g. one can still define homographs as far away meanings of the same word.

[10] Note that homophony in Chinese is much larger than homography: in average a syllable has around 12–13 meanings [26]. Hence, in a sense, characters help to resolve the homophony of Chinese speech. This argument is frequently presented as an advantage of the character-based writing system, though it is not clear whether this system is here not solving the problem that was invited by its usage [44].

[11] Not counting those heteronyms that arise because an English word happens to coincide with a foreign special name, e.g. *Nancy* [English name] and *Nancy* city in France.

[12] One should not conclude that in average the Chinese character has more meanings than the English word, because there is a large number of characters—between 10 and 14 % depending on the type of the dictionary employed [47]—that do not have lexical meaning, i.e. they are either function words (grammatical meaning mainly) or characters that cannot appear alone (bound characters). If now the number of meanings for each character is estimated via the number of entries in the explanatory dictionary—which is more or less traditional way of searching for the number of meanings, though it mixes up homography and polysemy—the average number of meanings per a Chinese character appears to be around 1.8–2 [47]. This is smaller than the average number of (necessarily polysemic) meanings for an English word that amounts to 2.3.

[13] Chinese expresses temporal ordering via context, e.g. adding words *tomorrow* or *yesterday*, or by aspects. The difference between tense and aspect is that the former implicitly assumes an external observer, whose reference time is compared with the time of the event described by the sentence. Aspects order events according to whether they are completed, or to which extent they are habitual. Indo-European languages tie up tense and aspect. The tie is weaker for Slavic Indo-European languages. Chinese has several tenses including perfective, imperfective and neutral.

[14] Chinese has certain affixes, but they can be and are suppressed whenever the issue is clear from the context.

above discussion is that Chinese characters have more different meanings, they are flexible, they could combine with other characters to convey different specific meanings. Anticipating our results in the sequel, we expect to see a group of characters, which appear many times in the text, but their concrete meanings are different in different places of the text.

III. THE ZIPF'S LAW FOR SHORT TEXTS

We studied several Chinese and English texts of different lengths and genres written in different epochs; see Tables I, II and III. Some Chinese texts were written using modern characters, others employ traditional Chinese characters; see Tables I and II. Chinese texts are described in Appendix C. English texts are described in Table III. The texts can be classified as short (total number of characters or words is $N = 1 - 3\times 10^4$) and long ($N > 10^5$). They generally have different rank-frequency characteristics, so discuss them separately.

For fitting empiric results we employed the linear least-square method (linear fitting), but the also checked its results with other methods (KS test, non-linear fitting and the maximum likelihood method). We start with a brief remainder of the linear fitting method.

A. Linear fitting

For each Chinese text we extract the ordered frequencies of different characters [the number of different characters is n; the overall number of characters in a text is N]:

$$\{f_r\}_{r=1}^n, \quad f_1 \geq ... \geq f_n, \quad \sum\nolimits_{r=1}^n f_r = 1. \tag{2}$$

Exactly the same method is applied to English texts for studying the rank-frequency relation of words.

We fit the data $\{f_r\}_{r=1}^n$ with a power law: $\hat{f}_r = cr^{-\gamma}$. Hence we represent the data as

$$\{y_r(x_r)\}_{r=1}^n, \quad y_r = \ln f_r, \quad x_r = \ln r, \tag{3}$$

and fit it to the linear form $\{\hat{y}_r = \ln c - \gamma x_r\}_{r=1}^n$. Two unknowns $\ln c$ and γ are obtained from minimizing the sum of squared errors [linear fitting]

$$SS_{\rm err} = \sum\nolimits_{r=1}^n (y_r - \hat{y}_r)^2. \tag{4}$$

It is known since Gauss that this minimization produces

$$-\gamma^* = \frac{\sum_{k=1}^n (x_k - \overline{x})(y_k - \overline{y})}{\sum_{k=1}^n (x_k - \overline{x})^2}, \quad \ln c^* = \overline{y} + \gamma^* \overline{x}, \tag{5}$$

where we defined

$$\overline{y} \equiv \frac{1}{n}\sum\nolimits_{k=1}^n y_k, \quad \overline{x} \equiv \frac{1}{n}\sum\nolimits_{k=1}^n x_k. \tag{6}$$

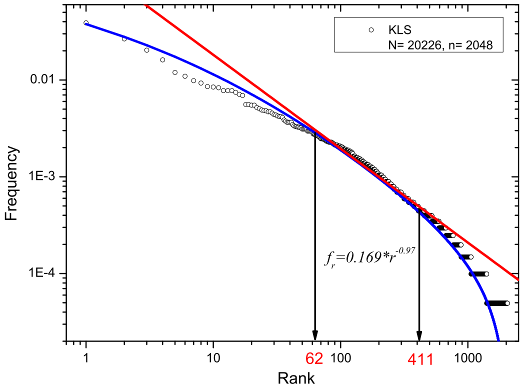

FIG. 1: (Color online) Frequency versus rank for the short modern Chinese text KLS; see Appendix C for its description. Red line: the Zipf curve $f_r = 0.169r^{-0.97}$; see Table I. Arrows and red numbers indicate on the validity range of the Zipf's law. Blue line: the numerical solution of (17, 18) for $c = 0.169$. It coincides with the generalized Zipf law (21) for $r > r_{\rm min} = 62$. The step-wise behavior of f_r for $r > r_{\rm max}$ refers to hapax legomena.

As a measure of fitting quality one can take:

$$\min\nolimits_{c,\gamma}[SS_{\rm err}(c,\gamma)] = SS_{\rm err}(c^*,\gamma^*) = SS^*_{\rm err}. \tag{7}$$

This is however not the only relevant quality measure. Another (more global) aspect of this quality is the coefficient of correlation between $\{y_r\}_{r=1}^n$ and $\{\hat{y}_r\}_{r=1}^n$ [2, 48]

$$R^2 = \frac{\left[\sum_{k=1}^n (y_k - \overline{y})(\hat{y}^*_k - \overline{\hat{y}^*})\right]^2}{\sum_{k=1}^n (y_k - \overline{y})^2 \sum_{k=1}^n (\hat{y}^*_k - \overline{\hat{y}^*})^2}, \tag{8}$$

where

$$\hat{y}^* = \{\hat{y}^*_r = \ln c^* - \gamma^* x_r\}_{r=1}^n, \quad \overline{\hat{y}^*} \equiv \frac{1}{n}\sum\nolimits_{k=1}^n \hat{y}^*_k. \tag{9}$$

For the linear fitting (5) the squared correlation coefficient is equal to the coefficient of determination,

$$R^2 = \sum\nolimits_{k=1}^n (\hat{y}^*_k - \overline{y})^2 \Big/ \sum\nolimits_{k=1}^n (y_k - \overline{y})^2, \tag{10}$$

the amount of variation in the data explained by the fitting [2, 48]. Hence $SS^*_{\rm err} \to 0$ and $R^2 \to 1$ mean good fitting. We minimize $SS_{\rm err}$ over c and γ for $r_{\rm min} \leq r \leq r_{\rm max}$ and find the maximal value of $r_{\rm max} - r_{\rm min}$ for which $SS^*_{\rm err}$ and $1 - R^2$ are smaller than, respectively, 0.05 and 0.005. This value of $r_{\rm max} - r_{\rm min}$ also determines the final fitted values c^* and γ^* of c and γ, respectively; see Tables I, II, III and Fig. 1. Thus c^* and γ^* are found simultaneously with the validity range $[r_{\rm max}, r_{\rm max}]$ of the law. Whenever there is no risk of confusion, we for simplicity refer to c^* and γ^* as c and γ, respectively.

B. Empiric results on the Zipf's law

Here are results produced via the above linear fitting.

TABLE I: Parameters of the modern Chinese texts (see Appendix C for further details). N is the total number of characters in the text. The number of different characters is n. The Zipf's law $f_r = cr^{-\gamma}$ holds for the ranks $r_{\min} \leq r \leq r_{\max}$; see section III A. Here $\sum_{k<r_{\min}} f_k$ and $\sum_{k=r_{\min}}^{r_{\max}} f_k$ are the total frequencies carried out by the pre-Zipfian and Zipfian domain, respectively. d is the difference between the total frequency of the Zipfian domain got empirically and its value according to the Zipf's law: $d = \sum_{k=r_{\min}}^{r_{\max}} (ck^{-\gamma} - f_k)$. Its absolute value d characterizes the global precision of the Zipf's law.
AQZ & KLS means joining the texts AQZ and KLS.
r_b is the conventional borderline rank between the exponential-like range and the hapax legomena; see section IV B 2 for its definition. Whenever we put "-" instead of it, we mean that either the exponential-like range is absent or it is not distinguishable from the hapax legomena.

Texts	N	n	$r_{\min}$	$r_{\max}$	c	γ	$\sum_{k<r_{\min}} f_k$	$\sum_{k=r_{\min}}^{r_{\max}} f_k$	$\lvert d\rvert$	r_b
AQZ	18153	1553	56	395	0.2239	1.03	0.42926	0.38424	0.00624	-
KLS	20226	2047	62	411	0.169	0.97	0.39971	0.379728	0.005728	-
AQZ & KLS	38379	2408	66	439	0.195	1.0	0.41684	0.369	0.0022	-
PFSJ	705130	3820	67	583	0.234	1.03	0.39544	0.425379	0.00842	1437
SHZ	704936	4376	78	590	0.225	1.02	0.39905	0.42	0.009561	1618

TABLE II: Parameters of classic Chinese texts (see Appendix C for further details). Notations have the same meaning as in Table I. Here 4 texts means joining of the texts CQF, SBZ, WJZ, and HLJ. Also, 7 (10,14) texts mean joining of the 4 with other 3 (6,11) classic texts, which we do not mention separately, because they give no new information.

Texts	N	n	$r_{\min}$	$r_{\max}$	c	γ	$\sum_{k<r_{\min}} f_k$	$\sum_{k=r_{\min}}^{r_{max}} f_k$	$\lvert d\rvert$	r_b
CQF	30017	1661	47	365	0.1778	0.985	0.43906	0.38997	0.00441	-
SBZ	24634	1959	52	357	0.1819	0.972	0.42828	0.408787	0.004353	-
WJZ	26330	1708	46	360	0.208	0.999	0.40434	0.418733	0.006923	-
HLJ	26559	1837	56	372	0.209	1.01	0.43674	0.379454	0.000832	-
CQF & SBZ	54651	2528	68	483	0.19498	0.989	0.42031	0.401661	0.00483	-
CQF & WJZ	56347	2302	66	439	0.20654	1.002	0.42815	0.383514	0.00564	-
CQF & HLJ	56576	2458	65	416	0.19498	0.998	0.43138	0.38654	0.00913	-
SBZ & WJZ	50964	2505	68	465	0.20512	0.992	0.40116	0.409017	0.00382	-
SBZ & HLJ	51193	2608	72	423	0.20893	1.000	0.41157	0.369598	0.00798	-
WJZ & HLJ	52889	2303	66	432	0.23988	1.035	0.43044	0.380801	0.002321	-
4 texts	107540	3186	75	528	0.22387	1.021	0.42526	0.391818	0.0007	681
7 texts	190803	4069	57	513	0.158	0.97	0.39381	0.4102	0.00331	789
10 texts	278557	4727	67	552	0.168	0.978	0.38058	0.4015	0.00217	1015
14 texts	348793	5018	78	625	0.176	0.98	0.39116	0.418983	0.00954	1223
SJ	572864	4932	76	535	0.236	1.025	0.40153	0.41253	0.007564	1336

1. For each Chinese text there is a specific (Zipfian) range of ranks $r \in [r_{\min}, r_{\max}]$, where the Zipf's law $f_r = cr^{-\gamma}$ holds with $\gamma \approx 1$ and $c \lesssim 0.25$; see Tables I, II and Fig. 1. Both for $r < r_{\min}$ and $r > r_{\max}$ the frequencies are below the Zipf curve; see Fig. 1. A power rank-frequency relation with exponent $\gamma \approx 1$ is the hallmark of the Zipf's law [1–4].

Note that though the validity range $|r_{\max} - r_{\min}|$ is few times smaller than the maximal rank n (see Tables I and II and Figs. 1 and 2), it is relevant, since it contains a sizable amount of the overall frequency: for Chinese texts (short or long) the Zipfian range carries 40 % of the overall frequency, i.e. $\sum_{k=r_{\min}}^{r_{\max}} f_k \simeq 0.4$.

2. In the pre-Zipfian range $1 \leq r < r_{\min}$ the overall number of function and empty characters is more than the number of content characters. Function and empty characters serve for establishing grammatical constructions (e.g. "的" *(de)*, "是" *(shì)*, "了" *(le)*, "不" *(bù)*, "在" *(zài)*). (We shall list them separately, though for our purposes they can be joined together; the main difference between them is that the empty characters are not used alone.)

But the majority of characters in the Zipfian range do have a specific meaning (content characters). A subset of those content characters has a meaning that is specific for the text and can serve as its key-characters; see Appendix D and Table IX for an example.

Let us take for an example the modern Chinese text KLS; see Table I (this text concerns military activities; see Appendix C). The pre-Zipfian range of this text con-

TABLE III: Parameters of four English texts and their mixtures: *The Age of Reason* (AR) by T. Paine, 1794 (the major source of British deism). *Time Machine* (TM) by H. G. Wells, 1895 (a science fiction classics). *Thoughts on the Funding System and its Effects* (TF) by P. Ravenstone, 1824 (economics). *Dream Lover* (DL) by J. MacIntyre, 1987 (a romance novella). TF & TM means joining the texts TF and TM.
The total number of words N, the number of different words n, the lower $r_{\rm min}$ and the upper $r_{\rm max}$ ranks of the Zipfian domain, the fitted values of c and γ, the overall frequencies of the pre-Zipfian and Zipfian range, and the difference d between the total frequency of the Zipfian domain got empirically and its value according to the Zipf's law: $d = \sum_{k=r_{\rm min}}^{r_{\rm max}}(ck^{-\gamma} - f_k)$.

Texts	N	n	$r_{\rm min}$	$r_{\rm max}$	c	γ	$\sum_{k<r_{\rm min}} f_k$	$\sum_{k=r_{\rm min}}^{r_{\rm max}} f_k$	$\vert d\vert$
TF	26624	2067	36	371	0.168	1.032	0.44439	0.35158	0.00333
TM	31567	2612	42	332	0.166	1.041	0.45311	0.33876	0.01004
AR	22641	1706	32	339	0.178	1.038	0.47254	0.33947	0.00048
DL	24990	1748	34	230	0.192	1.039	0.47955	0.33251	0.02145
TF & TM	54191	3408	30	602	0.139	1.013	0.43508	0.40876	0.02091
TF & AR	45265	2656	33	628	0.138	0.998	0.45468	0.41045	0.00239
TF & DL	47614	2877	28	527	0.162	1.014	0.42599	0.42261	0.01490
TM & AR	54208	3184	43	592	0.157	1.021	0.47582	0.39687	0.00491
TM & DL	56557	3154	45	493	0.161	1.023	0.46726	0.38456	0.01211
AR & DL	47631	2550	38	496	0.165	1.012	0.45375	0.39236	0.00947
Four texts	101822	4047	39	927	0.158	1.015	0.44245	0.44158	0.00187

tains 61 characters. Among them there are, 24 function characters, 9 empty characters, 25 content characters, and finally there are 3 key-characters [15]: horn "号" *(hào)*, army "军" *(jūn)* and soldier "兵" *(bīn)*.

The Zipfian range of the KLS contains 350 characters. Among them, 91 are function, 10 are empty, 230 are content and 19 are key-characters (see Appendix D for the full list of key-characters for this text).

3. The absolute majority of different characters with ranks in $[r_{\rm min}, r_{\rm max}]$ have different frequencies. Only for $r \simeq r_{\rm max}$ the number of different characters having the same frequency is $\simeq 10$. For $r > r_{\rm max}$ we meet the hapax legomena effect: characters occurring only few times in the text (i.e. $f_r N = 1, 2, 3...$ is a small integer), and many characters having the same frequency f_r [3]. The effect is not described by any smooth rank-frequency relation, including the Zipf's law. Hence for short texts we get that the Zipf's law holds for as high ranks as possible, in the sense that for $r > r_{\rm max}$ no smooth rank-frequency relations are possible at all.

Note that the very existence of hapax legomena is a non-trivial effect, since one can easily imagine (artificial) texts, where (say) no character appear only once. The theory reviewed below allows to explain the hapax legomena range together with the Zipf's law; see below. It also predicts a generalization of the Zipf's law to frequencies $r < r_{\rm min}$ that is more adequate (than the Zipf's law) to the empiric data; see Figs. 1 and 2.

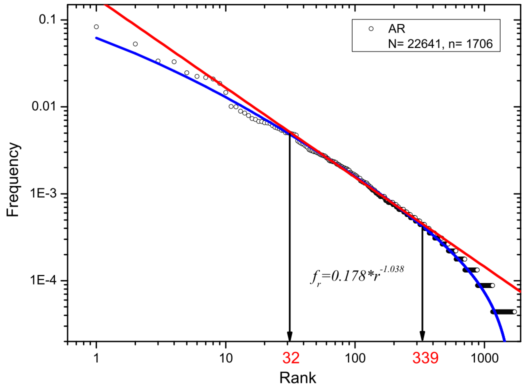

FIG. 2: (Color online) Frequency vs. rank for the English text AR; see Table III. Red line: the Zipf curve $f_r = 0.178r^{-1.038}$. Other notations have the same meanings as in Fig. 1.

4. All the above results hold for relatively short English [17]; see Table III and Fig. 2. In particular, the Zipfian range of English texts also contains mainly content words including the keywords. This is known and is routinely used in document processing [33].

We thus conclude that as far as short texts are concerned, the Zipf's law holds for Chinese characters in the same way as it does for English words.

5. To check our results on fitting the empiric data for word frequencies to the Zipf's law we carried out three alternative tests.

5.1 First we applied the Kolmogorov-Smirnov (KS) test to decide on the fitting quality of the data with the Zipf's law (in the range $[r_{\rm min}, r_{\rm max}]$). The test was carried out both with and without transforming to the logarithmic coordinates (3) and it fully confirmed our result;

[15] We present that meaning of the character which is most relevant in the context of the text.

see Table IV. For a detailed presentation of the KS test results see Appendix E and Table X therein.

5.2 It was recently shown that even when the applicability range $[r_{\min}, r_{\max}]$ of a power law is known, the linear least-square method (that we employed above) may not give accurate estimations for the exponent γ of the power law [49–51]. It was then argued that the method of Maximum Likelihood Estimation (MLE) is more reliable in this context. Hence to show that our results are robust, we calculated γ using the MLE method, as suggested in [49–51]. We got that the difference with the linear least square method is quite small (changes come only at the third decimal place); see Table IV.

5.3 We also checked whether our results on the power law exponent γ are stable with respect to non-linear fitting schemes, the ones that do not employ the logarithmic coordinates (3), but operate directly with the form (2). Again, we find that non-linear fitting schemes (that we carried out via routines of Mathematica 7) produce very similar results for γ; see Table IV.

One reason for such a good coincidence between our linear fitting results and alternative tests is that we use a rather strict criteria ($SS^{*}_{\rm err} < 0.05$ and $R^2 > 0.995$) for determining *first* the Zipfian range $[r_{\min},\, r_{\max}]$ and then the parameters of the Zipf's law. Another reason is that in the vicinity of $r_{\max}$, the number of different words having the same frequency is not large (it is smaller than 10). Hence there are no problems with lack of data points or systematic biases that can plague the applicability of the least square method for determination of the exponent γ.

C. Theoretical description of the Zipf's law and hapax legomena

1. *Assumptions of the model*

A theoretical description of the Zipf's law that is specifically applicable to short English texts was recently proposed in [17]; it is reviewed below. The theory is based on the ideas of latent semantic analysis and the concept of mental lexicon [17]. We shall now briefly remind it to demonstrate that

– The rank-frequency relation for short Chinese and English texts can be described by the same theory.

– The theory allows to extrapolate the Zipf's law to high and low frequencies (including hapax legomena).

– It allows to understand the bound $c < 0.25$ for the prefactor of the Zipf's law (since the law does not apply for all frequencies, c is not fixed from normalization).

– The theory confirms the intuitive expectation about the difference between the Zipfian and hapax legomena range: in the first case the probability of a word is equal to its frequency (frequent words). In the hapax legomena range, both the probability and frequency are small and different from each other.

– In the following section the theory is employed for describing the rank-frequency relation of Chinese characters outside of the validity range of the Zipf's law.

Our model for deriving the Zipf's law together with the description of the hapax legomena makes four assumptions (see [17] for further details). Below we shall refer to the units of the text as words; whenever this theory applies for Chinese texts we shall mean characters instead of words.

• The *bag-of-words picture* focusses on the frequency of the words that occur in a text and neglects their mutual disposition (i.e. syntactic structure) [52]. This is a natural assumption for a theory describing word frequencies, which are invariant with respect to an arbitrary permutation of the words in a text. The latter point was recently verified in [53].

Given n different words $\{w_k\}_{k=1}^n$, the joint probability for w_k to occur $\nu_k \geq 0$ times in a text T is assumed to be multinomial

$$\pi[\boldsymbol{\nu}|\boldsymbol{\theta}] = \frac{N!\,\theta_1^{\nu_1}...\theta_n^{\nu_n}}{\nu_1!...\nu_n!}, \quad \boldsymbol{\nu} = \{\nu_k\}_{k=1}^n, \quad \boldsymbol{\theta} = \{\theta_k\}_{k=1}^n (11)$$

where $N = \sum_{k=1}^n \nu_k$ is the length of the text (overall number of words), ν_k is the number of occurrences of w_k, and θ_k is the probability of w_k.

Hence according to (11) the text is regarded to be a sample of word realizations drawn independently with probabilities θ_k.

The bag-of-words picture is well-known in computational linguistics [52]. But for our purposes it is incomplete, because it implies that each word has the same probability for different texts. In contrast, it is well known (and routinely confirmed by the rank-frequency analysis) that the same words do *not* occur with same frequencies in different texts.

• To improve this point we make $\boldsymbol{\theta}$ a random vector with a text-dependent density $P(\boldsymbol{\theta}|T)$ (a similar, but stronger assumption was done in [52]). With this assumption the variation of the word frequencies from one text to another will be explained by the randomness of the word probabilities.

We now have three random objects: text T, probabilities $\boldsymbol{\theta}$ and the occurrence numbers $\boldsymbol{\nu}$. Since $\boldsymbol{\theta}$ was introduced to explain the relation of T with $\boldsymbol{\nu}$, it is natural to assume that the triple $(T, \boldsymbol{\theta}, \boldsymbol{\nu})$ form a Markov chain: the text T influences the observed $\boldsymbol{\nu}$ only via $\boldsymbol{\theta}$. Then the probability $p(\boldsymbol{\nu}|T)$ of $\boldsymbol{\nu}$ in a given text T reads

$$p(\boldsymbol{\nu}|T) = \int \mathrm{d}\boldsymbol{\theta}\, \pi[\boldsymbol{\nu}|\boldsymbol{\theta}]\, P(\boldsymbol{\theta}|T). \qquad (12)$$

This form of $p(\boldsymbol{\nu}|T)$ is basic for probabilistic latent semantic analysis [54], a successful method of computational linguistics. There the density $P(\boldsymbol{\theta}|T)$ of latent variables $\boldsymbol{\theta}$ is determined from the data fitting. We shall deduce $P(\boldsymbol{\theta}|T)$ theoretically.

• The text-conditioned density $P(\boldsymbol{\theta}|T)$ is generated from a prior density $P(\boldsymbol{\theta})$ via conditioning on the or-

TABLE IV: Comparison between different methods of estimating the exponent γ of the Zipf's law; see (1): LLS (linear least-square), NLS (nonlinear least-square), MLE (maximum likelihood estimation). We also present the p-value of the KS test when comparing the empiric word frequencies in the range $[r_{\rm min},\, r_{\rm max}]$ with the Zipf's-law within the linear lest-square method (LLS); for a more detailed presentation of the KS results see Appendix E. Recall that the p-values have to be sufficiently larger than 0.1 for fitting to be reliable from the viewpoint of KS test. This holds for the presented data; see Appendix E for details.

Texts	γ, LLS	γ, NLS	γ, MLE	p-value
TF	1.032	1.033	1.035	0.865
TM	1.041	1.036	1.039	0.682
AR	1.038	1.042	1.044	0.624
DL	1.039	1.034	1.035	0.812
AQZ	1.03	1.028	1.027	0.587
KLS	0.97	0.975	0.973	0.578
CQF	0.985	0.983	0.981	0.962
SBZ	0.972	0.967	0.973	0.796
WJZ	0.999	0.993	0.995	0.852
HLJ	1.01	1.015	1.011	0.923

dering of $\mathbf{w} = \{w_k\}_{k=1}^n$ in T:

$$P(\boldsymbol{\theta}|T) = P(\boldsymbol{\theta})\,\chi_T(\boldsymbol{\theta},\mathbf{w}) \Big/\!\!\Big/ \int {\rm d}\boldsymbol{\theta}'\,P(\boldsymbol{\theta}')\,\chi_T(\boldsymbol{\theta}',\mathbf{w})\;. \tag{13}$$

Thus if different words of T are ordered as $(w_1, ..., w_n)$ with respect to the decreasing frequency of their occurrence in T (i.e. w_1 is more frequent than w_2), then $\chi_T(\boldsymbol{\theta},\mathbf{w}) = 1$ if $\theta_1 \geq ... \geq \theta_n$, and $\chi_T(\boldsymbol{\theta},\mathbf{w}) = 0$ otherwise.

- The apriori density of the word probabilities $P(\boldsymbol{\theta})$ in (13) can be related to the mental lexicon (store of words) of the author prior to generating a concrete text. For simplicity, we assume that the probabilities θ_k are distributed identically [see [17] for a verification of this assumption] and the dependence among them is due to $\sum_{k=1}^n \theta_k = 1$ only:

$$P(\boldsymbol{\theta}) \propto u(\theta_1) \ldots u(\theta_n)\, \delta(\textstyle\sum_{k=1}^n \theta_k - 1), \tag{14}$$

where $\delta(x)$ is the delta function and the normalization ensuring $\int_0^\infty \prod_{k=1}^n {\rm d}\theta_k\, P(\boldsymbol{\theta}) = 1$ is omitted.

2. *Zipf's law*

It remains to specify the function $u(\theta)$ in (14). Ref. [17] reviews in detail the experimentally established features of the human mental lexicon (see [55] in this context) and deduces from them that the suitable function $u(\theta)$ is

$$u(f) = (n^{-1}c + f)^{-2}, \tag{15}$$

where c is to be related to the prefactor of the Zipf's law.

Above equations (11–15) together with the feature $n^3 \gg N \gg 1$ of real texts (where n is the number of different words, while N is the total number of words in the text) allow to the final outcome of the theory: the probability $p_r(\nu|T)$ of the character (or word) with the rank r to appear ν times in a text T (with N total characters and n different characters) [17]:

$$p_r(\nu|T) = \frac{N!}{\nu!(N-\nu)!}\phi_r^\nu (1-\phi_r)^{N-\nu}, \tag{16}$$

where the effective probability ϕ_r of the character is found from two equations for two unknowns μ and ϕ_r:

$$r/n = \int_{\phi_r}^\infty {\rm d}\theta\, \frac{e^{-\mu\theta}}{(c+\theta)^2} \Big/\!\!\Big/ \int_0^\infty {\rm d}\theta\, \frac{e^{-\mu\theta}}{(c+\theta)^2}\;, \tag{17}$$

$$\int_0^\infty {\rm d}\theta\, \frac{\theta\, e^{-\mu\theta}}{(c+\theta)^2} = \int_0^\infty {\rm d}\theta\, \frac{e^{-\mu\theta}}{(c+\theta)^2}, \tag{18}$$

where c is a constant that will later on shown to coincide with the prefactor of the Zipf's law.

For $c \lesssim 0.25$, $c\mu$ determined from (18) is small and is found from integration by parts:

$$\mu \simeq c^{-1}\, e^{-\gamma_{\rm E} - \frac{1+c}{c}}, \tag{19}$$

where $\gamma_{\rm E} = 0.55117$ is the Euler's constant. One solves (17) for $c\mu \to 0$:

$$\frac{r}{n} = ce^{-n\phi_r\mu}/(c + n\phi_r). \tag{20}$$

Recall that according to (16), ϕ_r is the probability for the character (or the word in the English situation) with rank r. If ϕ_r is sufficiently large, $\phi_r N \gg 1$, the character with rank r appears in the text many times and its frequency $\nu \equiv f_r N$ is close to its maximally probable value $\phi_r N$; see (16). Hence the frequency f_r can be obtained via the probability ϕ_r. This is the case in the Zipfian domain, since according to our empirical results (both for

Chinese and English) $\frac{1}{n} \lesssim f_r$ for $r \leq r_{\rm max}$, and—upon identifying $\phi_r = f_r$—the above condition $\phi_r N \gg 1$ is ensured by $N/n \gg 1$; see Tables I, II and III.

Let us return to (20). For $r > r_{\rm min}$, $\phi_r n\mu = f_r n\mu < 0.04 \ll 1$; see (19) and Figs. 1 and 2. We get from (20):

$$f_r = c(r^{-1} - n^{-1}). \tag{21}$$

This is the Zipf's law generalized by the factor n^{-1} at high ranks r. This cut-off factor ensures faster [than r^{-1}] decay of f_r for large r.

Figs. 1 and 2 shows that (21) reproduces well the empirical behavior of f_r for $r > r_{\rm min}$. Our derivation shows that c is the prefactor of the Zipf's law, and that our assumption on $c \lesssim 0.25$ above (19) agrees with observations; see Tables I, II and III.

For given prefactor c and the number of different characters n, (17) predict the Zipfian range $[r_{\rm min}, r_{\rm max}]$ in agreement with empirical results; see Figs. 1 and 2.

For $r < r_{\rm min}$, it is not anymore true that $f_r n\mu \ll 1$ (though it is still true that $f_r N = \phi_r N \gg 1$). So the fuller expression (17) is to be used instead of (20). It reproduces qualitatively the empiric behavior of f_r also for $r < r_{\rm min}$; see Figs. 1 and 2. We do not expect any better agreement theory and observations for $r < r_{\rm min}$, since the behavior of frequencies in this range is irregular and changes significantly from one text to another.

D. Hapax legomena

1. *Hapax legomena as a consequence of the generalized Zipf's law*

According to (16), the probability ϕ_r is small for $r \gg r_{\rm max}$ and hence the occurrence number $\nu \equiv f_r N$ of the character with the rank r is a small integer (e.g. 1 or 2) that cannot be approximated by a continuous function of r; see Figs. 1 and 2. In particular, the reasoning after (20) on the equality between frequency and probability does not apply, although we see in Figs. 1 and 2 that (21) roughly reproduces the trend of f_r even for $r > r_{\rm max}$.

To describe this hapax legomena range, define r_k as the rank, when $\nu \equiv f_r N$ jumps from integer k to $k+1$ (hence the number of characters that appear $k+1$ times is $r_k - r_{k+1}$). Since ϕ_r reproduces well the trend of f_r even for $r > r_{\rm max}$, see Fig. 1, r_k can be theoretically predicted from (21) by equating its left-hand-side to k/N:

$$\hat{r}_k = [\frac{k}{Nc} + \frac{1}{n}]^{-1}, \qquad k = 0, 1, 2, ... \tag{22}$$

Eq. (22) is exact for $k = 0$, and agrees with r_k for $k \geq 1$; see Table V. We see that a single formalism describes both the Zipf's law for short texts and the hapax legomena range. We stress that for describing the hapax legomena no new parameters are needed; it is based on the same parameters N, n, c that appear in the Zipf's law.

2. *Comparing with previous theories of hapax legomena*

Several theories were proposed over the years for describing the hapax legomena range; see [56] for a review. To be precise, these theories were proposed for rare words (not for rare Chinese characters), but since the Zipf's law applies to characters, we expect that these theories will be relevant. We now compare predictions of the main theories with (22). The latter turns out to be superior.

Recall that for obtaining (22) it is necessary to employ the generalized (by the factor n^{-1}) form (21) of the Zipf's law. The correction factor is not essential in the proper Zipfian domain (since it is a pure power law), but is crucial for obtaining a good agreement with empiric data in the hapax legomena range; see Figs. 1 and 2. The influence of this correcting factor can be neglected for $k \gg Nc/n$ in (22), where we get

$$\hat{r}_{k-1} - \hat{r}_k \propto \frac{1}{k(k-1)}, \tag{23}$$

for the number of characters having frequency k/N. This relation, which is a crude particular case of (22), is sometimes called the second Zipf's law, or the Lotka's law [3, 56]. The applicability of (23) is however limited, e.g. it does not apply to the data shown in Table V.

Another approach to frequencies of rare words was proposed in [57]; see [56] for a review. Its basic result (24) was recently recovered from a partial maximization of entropy (random group formation approach) [58] [16]. It makes the following prediction for the number $nP(k)$ [17] of characters that appear in the text k times (i.e. $P(k)$ is a prediction for $(r_{k-1} - r_k)/n$)

$$P(k) \propto e^{-bk}\, k^{-\gamma}, \qquad 1 \leq k \leq f_1 N, \tag{24}$$

where we omitted the normalization ensuring $\sum_{k=1}^{f_1 N} P(k) = 1$, and where the constants $b > 0$ and $\gamma > 0$ are determined from three parameters of the text: the overall number of characters N, the number of different characters n and the maximal frequency f_1 [58]. Distributions similar to (24) (i.e. exponentially modified power-laws) were derived from partial maximization of

[16] Ref. [58] presented a broad range of applications, but it did not study Chinese characters. We acknowledge one of the referees of this work who informed us that such unpublished studies do exist: Chinese characters are within the applicability range of Ref. [58], as we confirm in Table V. The predictions of (24) for the AQZ text that we reproduce in Table V were communicated to us by the referee.

[17] Please do not mix up $P(k)$ with the density of character probabilities that appear in (14, 15). Indeed, $P(k)$ is defined as empiric frequency; it has a discrete argument and applies to any collection of objects, also the one that was generated by any probabilistic mechanism. In contrast, (14, 15) amount to a density of probabities that has continuous argument(s) and assumes a specific generative model.

entropy prior to Ref. [58] (e.g. in [59, 60]), but it was Ref. [58] that emphasized their broad applicability.

Note that $P(k)$ in (24) does not apply out of the hapax legomena range, where for all k we must have $P(k) = 1/n$. However, it is expected that for $n \gg 1$ this discrepancy will not hinder the applicability of (24) to $P(k)$ with sufficiently small values of k, i.e. within the hapax legomena range.

The results predicted by (24) are compared with our data in Table V. For clarity, we transform (24) to a prediction $\widetilde{r}_k$ for quantities r_k:

$$\widetilde{r}_l = n[1 - \sum_{k=1}^{l} P(k)], \quad l \geq 1, \tag{25}$$

i.e. we go to the cumulative distribution function $\sum_{k=1}^{l} P(k)$.

While the predictions of (25) are in a certain agreement with the data, their accuracy is inferior (at least by an order of magnitude) as compared to predictions of (22); see Table V. The reason of this inferiority is that though both (24) and (22) use three input parameters, (24) is not sufficiently specific to the studied text.

Finally, let us turn to the Waring-Herdan approach which predicts for $nP(k)$ (the number of characters that appear in the text k times) a version of the Yule's distribution [56]:

$$P(k+1) = P(k)\,\frac{a+k-1}{x+k}, \quad k \geq 1, \tag{26}$$

where a and x are expressed via three (the same number as in the previous two approaches) input parameters N (the overall number of characters), n (the number of distinct characters) and $nP(1)$ (the number of characters that appear only once) [56]:

$$a = \left(\frac{1}{1-P(1)} - P(1) - 1\right)^{-1}, \quad x = \frac{a}{1-P(1)}. \tag{27}$$

Eqs. (26, 27) are turned to a prediction r'_k for r_k. As Table VI shows, these predictions [18] are also inferior as compared to those of (22), especially for $k \geq 5$.

E. Summary

It is to be concluded from this section that—as far as the applicability of the Zipf's law to short texts is concerned—the Chinese characters behave similarly to English words. In particular, both situations can be adequately described by the same theory. In particular, the hapax legomena range of short texts is described via the generalized Zipf's law.

We should like to stress again why the consideration of short texts is important. One can argue that—at least for the sake of rank-frequency relations—long texts are just mixtures (joinings) of shorter, thematically homogeneous pieces (this premise is fully confirmed below). Hence the task of studying rank-frequency relations separates into two parts: first understanding short texts, and then long ones. We now move to the second part.

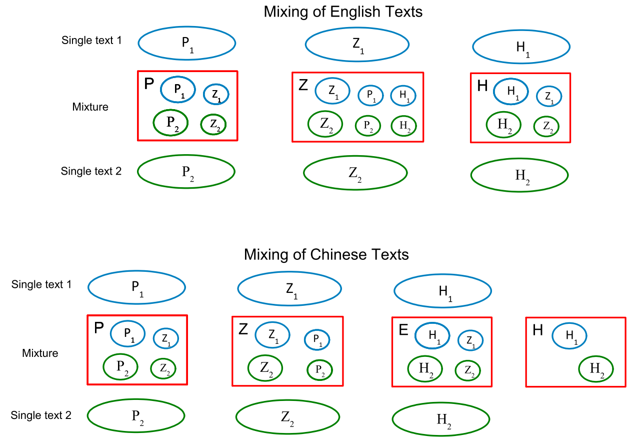

FIG. 3: Schematic representation of various ranges under mixing (joining) two English (upper figure) and two Chinese (lower figure) texts. P_k, Z_k and H_k mean, respectively, the pre-Zipfian, Zipfian and hapax legomena ranges of the text k ($k = 1, 2$). P, Z and H mean the corresponding ranges for the mixture of texts 1 and 2. E means the exponential-like range that emerges upon mixing of two Chinese texts. For each range of the mixture we show schematically contributions from various ranges of the separate texts. The relative importance of each contribution is conventionally represented by different magnitudes of the circles.

IV. RANK-FREQUENCY RELATION FOR LONG TEXTS AND MIXTURES OF TEXTS

A. Mixing English texts

When mixing (joining)[19] different English texts the validity range of the Zipf's law increases due to acquiring more higher rank words, i.e. $r_{\rm min}$ stays approximately fixed, while $r_{\rm max}$ increases; see Table III. The overall

[18] Eq. (26) can viewed as a consequence of the Simon's model of text generation. This model does not apply to real texts as was recently demonstrated in [53]. Nevertheless (26) keeps its relevance as a convenient fitting expression; see also [56] in this context.

[19] Upon joining two texts (A and B), the word frequencies get mixed: $f_k(\mathrm{A\&B}) = \frac{N_{\rm A}}{N_{\rm A}+N_{\rm B}} f_k(\mathrm{A}) + \frac{N_{\rm B}}{N_{\rm A}+N_{\rm B}} f_k(\mathrm{B})$, where $N_{\rm A}$ and $f_k(\mathrm{A})$ are, respectively, the total number of words and the frequency of word k in the text A.

TABLE V: The hapax legomena range for Chinese characters demonstrated for 4 short Chinese texts. The first and second text are in Modern Chinese, other two are in Classic Chinese; see Tables I and II. r_k is defined before (22) and is found from empirical data, while $\hat{r}_k$ is calculated from (22); see section III D. We also present the relative error for $\hat{r}_k$ approximating r_k.

Texts	k	1	2	3	4	5	6	7	8	9	10
AQZ	r_k	1097	857	702	595	522	461	414	370	339	311
	$\hat{r}_k$	1116	869	711	601	520	458	409	369	336	308
	$\frac{\lvert\hat{r}_k - r_k\rvert}{r_k}$	0.017	0.014	0.013	0.010	0.0038	0.0065	0.012	0.0027	0.0088	0.0096
KLS	r_k	1405	1060	885	767	662	582	520	455	408	377
	$\hat{r}_k$	1428	1093	884	750	656	575	515	445	404	369
	$\frac{\lvert\hat{r}_k - r_k\rvert}{r_k}$	0.016	0.031	0.0011	0.022	0.0091	0.012	0.0096	0.022	0.0098	0.021
SBZ	r_k	1460	1141	959	850	735	676	618	563	517	481
	$\hat{r}_k$	1481	1168	980	848	740	656	599	553	497	488
	$\frac{\lvert\hat{r}_k - r_k\rvert}{r_k}$	0.014	0.024	0.022	0.0024	0.0068	0.029	0.031	0.018	0.039	0.015
HLJ	r_k	1302	1045	872	756	669	604	551	501	467	430
	$\hat{r}_k$	1327	1080	900	783	684	607	545	494	462	420
	$\frac{\lvert\hat{r}_k - r_k\rvert}{r_k}$	0.019	0.033	0.032	0.035	0.022	0.0049	0.011	0.014	0.011	0.023

TABLE VI: The hapax legomena range for 2 Chinese texts; see Table I and cf. with Table V. We compare the relative errors for, respectively, $\hat{r}_k$ (given by (22)) $\widetilde{r}_k$ and r'_k in approximating the data r_k; see section III D 2. Here $\widetilde{r}_k$ is defined by (25, 24), and r'_k is the prediction made by (26, 27). For AQZ the parameters in (24) are $\gamma = 1.443$ and $b = 0.0049$. For KLS: $\gamma = 1.574$ and $b = 0.0033$. It is seen that the relative error provided by $\hat{r}_k$ is always smaller; the only exclusion is the case $k = 2$ of the KLS text. Recall that $r'_1 = r_1$ by definition.

Texts	k	1	2	3	4	5	6	7	8	9	10
AQZ	$\lvert\widetilde{r}_k - r_k\rvert/r_k$	0.141	0.161	0.153	0.138	0.129	0.112	0.097	0.069	0.057	0.039
	$\lvert r'_k - r_k\rvert/r_k$	0	0.025	0.048	0.071	0.102	0.121	0.141	0.146	0.163	0.174
	$\lvert\hat{r}_k - r_k\rvert/r_k$	0.017	0.014	0.013	0.010	0.0038	0.0065	0.012	0.0027	0.0088	0.0096
KLS	$\lvert\widetilde{r}_k - r_k\rvert/r_k$	0.194	0.221	0.245	0.267	0.259	0.250	0.240	0.206	0.183	0.179
	$\lvert r'_k - r_k\rvert/r_k$	0	0.0087	0.063	0.114	0.137	0.157	0.176	0.168	0.170	0.190
	$\lvert\hat{r}_k - r_k\rvert/r_k$	0.016	0.031	0.0011	0.022	0.0091	0.012	0.0096	0.022	0.0098	0.021

precision of the Zipf's law also increases upon mixing, as Table III shows.

The rough picture of the evolution of the rank-frequency relation under mixing two texts is summarized as follows; see Table III and Fig. 3 for a schematic illustration. The majority of the words in the Zipfian range of the mixture (e.g. AR & TM) come from the Zipfian ranges of the separate texts. In particular, all the words that appear in the Zipfian ranges of the separate words do appear as well in the Zipfian range of the mixture (e.g. the Zipfian ranges of AR and TM have 130 common words). There are also relatively smaller contributions to the Zipfian range of the mixture from the pre-Zipfian and hapax legomena range of separate texts: note from Table III that the Zipfian range of the mixture AR & TM is 82 words larger than the sum of two separate Zipfian ranges, which is (307 + 290) minus 130 common words.

Some of the words that appear only in the Zipfian range of one of separate texts will appear in the hapax legomena range of the mixture; other words move from the pre-Zipfian range of separate texts to the Zipfian range of the mixture. But these are relatively minor effects: the rough effect of mixing is visualized by saying that the Zipfian ranges of both texts combine to become a larger Zipfian range of the mixture and acquire additional words from other ranges of the separate texts; see Fig. 3. Note that the keywords of separate words stay in the Zipfian range of the mixture, e.g. after joining all four above texts, the keywords of each text are still in the Zipfian range (which now contains almost 900 words); see Table III.

The results on the behavior of the Zipf's law under mixing are new, but their overall message—the validity of the Zipf's law improves upon mixing—is expected, since it is known that the Zipf's law holds not only for short but also for long English texts and for frequency dictionaries (huge mixtures of various texts) [1–4].

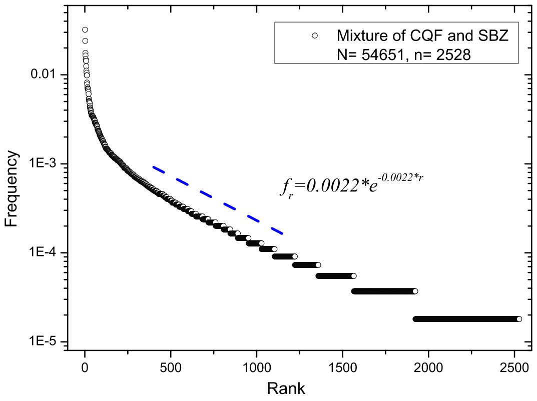

FIG. 4: (Color online) Rank frequency distribution for the mixture of CQF and SBZ.; see Tables I and II and Appendix C. The scale of the frequency is chosen such that the exponential-like range of the rank-frequency relation for $r > 500$ is made visible. For comparison, the dashed blue line shows a curve $f_r = 0.0022e^{-0.0022r}$. For the present example, the exponential-like range is essentially mixed with hapax legomena, since for frequencies f_r with $r > r_{\max}$ the number of different words having this frequency is larger than 10. Recall that the Zipf's law holds for $r_{\min} < r < r_{\max}$; see Tables I and II.

B. Mixing Chinese texts

1. Stability of the Zipfian range

The situation for Chinese texts is different. Upon mixing two Chinese texts the validity range of the Zipf's law increases, but much slower as compared to English texts; see Tables I and II. The validity ranges of the separate texts do not combine (in the above sense of English texts). Though the common words in the Zipfian ranges of separate texts do appear in the Zipfian range of the mixture, a sizable amount of those words that appeared in the Zipfian range of only one text do not show up in the Zipfian range of the mixture [20].

Importantly, the overall frequency of the Zipfian domain for very different Chinese texts (mixtures, long texts) is approximately the same and amounts to $\simeq 0.4$; see Tables I and II. In contrast, for English texts this overall frequency grows with the number of different words in the text; see Table III. This is consistent with the fact that for English texts the Zipfian range increases upon mixing.

2. Emergence of the exponential-like range

The majority of characters that appear in the Zipfian range of separate texts, but do not appear in the Zipfian range of the mixture, moves to the hapax legomena range of the mixture. Then, for larger mixtures and longer texts, a new, exponential-like range of the rank-frequency relation emerges from within the hapax legomena range.

To illustrate the emergence of the exponential-like range let us start with Fig. 4. Here there are only two short texts mixed and hence the exponential-like range cannot be reliably distinguished from the hapax legomena [21]: for all frequencies with the ranks $r > r_{\max}$ (i.e. for all frequencies beyond the Zipfian range), the number of different characters having exactly the same frequency is larger than 10. (We conventionally take this number as a borderline of the hapax legomena.) However, the trace of the exponential-like range is seen even within the hapax legomena; see Fig. 3.

For bigger mixtures or longer texts, the exponential-like range clearly differentiates from the hapax legomena. In this context, we define r_b as the borderline rank of the hapax legomena: for $r > r_b$, the number of characters having the frequency f_{r_b} is larger than 10. Then the exponential-like range

$$f_r = ae^{-br} \quad \text{with} \quad a < b, \tag{28}$$

exists for the ranks $r_{\max} < r \lesssim r_b$ (provided that $r_{\max}$ is sufficiently larger than r_b); see Table VII. Put differently, the exponential-like range exists from ranks larger than the upper rank $r_{\max}$ of the Zipfian range till the ranks, where the hapax legomenon starts. Tables I, II, VII and Fig. 5 show that the exponential-like range is not only sizable by itself, but (for sufficiently long texts or sufficiently big mixtures) it is also bigger than the Zipfian range. This, of course, does not mean that the Zipfian range becomes less important, since, as we saw above, it carries out nearly 40 % of the overall frequency; see Tables I and II. The exponential-like range also carries out non-negligible frequency, though it is few times smaller than that of the Zipfian and pre-Zipfian ranges; see Tables I, II and VII.

Finally, we would like to stress that we considered various Chinese texts written with simplified or traditional characters, with Modern Chinese or different versions of Classic Chinese; see Tables I, II and Appendix C. As far

[20] As an example, let us consider in detail the mixing of two Chinese texts SBZ and CQF; see Table II. The Zipfian ranges of CQF and SBZ contain, respectively, 306 and 319 characters. Among them 133 characters are common. The balance of the characters upon mixing is calculated as follows: 306 (from the Zipfian range of CQF) + 319 (from the Zipfian range of SBZ) - 133 (common characters) - 50 (characters from the Zipfian range of CQF that do not appear in the Zipfian range of CQF & SBZ) - 54 (characters from the Zipfian range of SBZ that do not appear in the Zipfian range of CQF+SBZ) +27 (characters that enter to the Zipfian range CQF & SBZ from the pre-Zipfian ranges of CQF or SBZ)= 415 (characters in the Zipfian range of CQF & SBZ).

[21] Recall in this context that in the hapax legomena range many characters have the same frequency, hence no smooth rank-frequency relation is reliable.

TABLE VII: Parameters of the exponential-like range (lower and upper ranks and the overall frequency) for few long Chinese texts; see also Tables I and II. Here n is the number of different characters. Recall that the lowest rank of the exponential-like range is $r_{\max}+1$, where $r_{\max}$ is the upper rank of the Zipfian range. The highest rank of the exponential-like range was denoted as r_b; see Tables I and II.

Texts	n	Rank range	Overall frequency
PFSJ	3820	584–1437	0.12816
SHZ	4376	591–1618	0.14317
SJ	4932	536–1336	0.12887
14 texts	5018	626–1223	0.12291

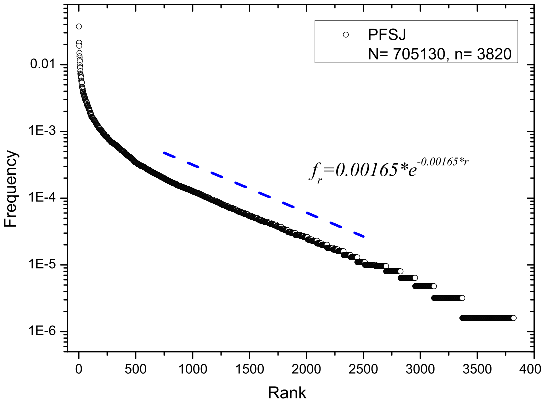

FIG. 5: (Color online) Rank frequency distribution of the long modern Chinese text PFSJ. The exponential behavior $f_r \propto e^{-0.00165r}$ of frequency f_r is visible for $r > 500$. For comparison, the dashed blue line shows a curve $f_r = 0.00165e^{-0.00165r}$. The boundary between the exponential-like range and hapax legomena can be defined as the rank r_b, where the number of words having the same frequency f_{r_b} is equal to 10. For the present example $r_b = 1437$. The Zipf's law holds for ranks $r_{\min} < r < r_{\max}$, where $r_{\max} = 583$, $r_{\min} = 67$; see Table I.

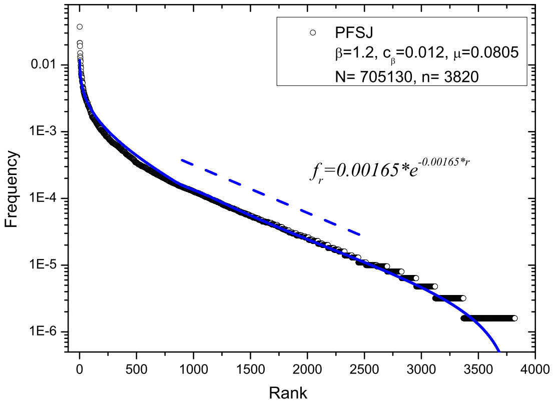

FIG. 6: (Color online) The rank-frequency relation $f(r)$ for characters from the text PFSJ; see Table I. Blue line denotes the numerical solution of (31, 32) at the indicated parameters β and c_β. The dashed blue line indicates at the exponential-like regime.

as the rank-frequency relations are concerned, all these texts demonstrate the same features showing that the peculiarities of these relations are based on certain very basic features of Chinese characters. They do not depend on specific details of texts.

C. Theoretical description of the exponential-like regime

Now we search for a theoretical description for the exponential like regime of the rank-frequency relation of Chinese characters. This description will simultaneously account for the hapax legomena range (rare words) of long Chinese texts.

We proceed with the theory outlined in section III C 1 and III C 2. There we saw that the Zipf's law results from the choice (15) of the prior density for word probabilities $\boldsymbol{\theta}$. Now we need to generalize (15). Recall that the choice of prior densities is is the main problem of the Bayesian statistics [61, 62] [22]. One way to approach this problem is to look for a natural group in the space of events (e.g. the translation group if the event space is the real line) and then define the non-informative prior density as the one which is invariant with respect to the group [61, 62]. Our event space is the simplex $\boldsymbol{\theta} \in \mathcal{S}_n$: the set of n non-negative numbers (word probabilities) that sum to one. The natural group on the simplex is the multiplicative group [61] (in a sense this is the only group that preserves probability relations [62]), and the corresponding non-informative density is the Haldane's prior [61–63] that is given by (14) under

$$u(f) = (n^{-1}c_1 + f)^{-1}, \quad c_1 \to 0. \tag{29}$$

The formal Haldane's prior is recovered from (29) under

[22] We stress that (for a continuous event space) this problem is not solved by the maximum entropy method. In contrast, this method itself does need the prior density as one of its inputs [61].

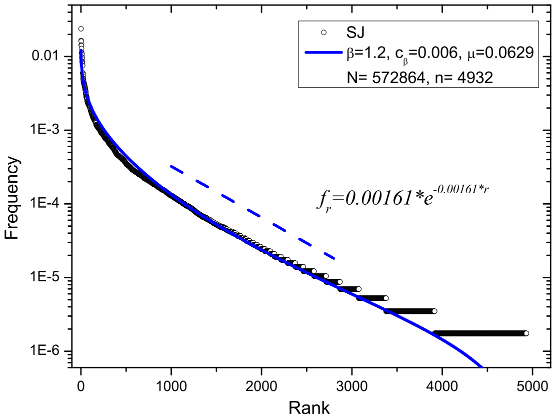

FIG. 7: (Color online) The rank-frequency relation $f(r)$ for characters from the text SJ; see Table II. For parameters and notations see Fig. 6.

$c_1 \equiv 0$; a small but finite constant c_1 is necessary for making the density normalizable.

Note that the prior density (15) which supports the Zipf's law is far from being non-informative. This is natural, because it relates to a definite organization of the mental lexicon [17].

Now the exponential-like regime of the rank-frequency relation can be deduced via a prior density that is intermediate between the Zipfian prior (15) and the non-informative, Haldane's prior (29)

$$u(f) = (n^{-1}c_\beta + f)^{-\beta}, \quad 1 < \beta < 2, \tag{30}$$

where β and $c_\beta > 0$ are to be determined from the data-fitting. Now we can still use (16) but instead of (17, 18) we get the following implicit relations for the smooth part of the rank-frequency relation f_r

$$r/n = \int_{f_r}^{\infty} \mathrm{d}\theta \, \frac{e^{-\mu\theta}}{(c_\beta+\theta)^\beta} \Bigg/ \int_0^{\infty} \mathrm{d}\theta \, \frac{e^{-\mu\theta}}{(c_\beta+\theta)^\beta} \,, \tag{31}$$

$$\int_0^{\infty} \mathrm{d}\theta \, \frac{\theta\, e^{-\mu\theta}}{(c_\beta+\theta)^\beta} = \int_0^{\infty} \mathrm{d}\theta \, \frac{e^{-\mu\theta}}{(c_\beta+\theta)^\beta}. \tag{32}$$

Figs. 6 and 7 compare these analytical predictions with data. The fit is seen to be good under parameters β and c_β that are not very far from (29). The fact that prior densities close to the non-informative (Haldane's) prior generate an exponential-like shape for the rank-frequency relations is intuitive, since such a shape means that a relatively small group of words carries out the major part of frequency.

As confirmed by Figs. 6 and 7, predictions of (31, 32) that describe the exponential-like regime are not applicable for the Zipfian range.

Importantly, (31, 32) allow to describe the hapax legomena range of long Chinese texts. Following section III D, we equate the solution f_r of (31, 32) to k/N and determine from this $r = \hat{r}_k$: the rank in the hapax legomena range, where the frequency jumps from k/N to $(k+1)/N$. Now $\hat{r}_k$ agrees well with the empiric data for the hapax legomena range of long Chinese texts, and the agreement is better than for the approach based on (24, 25); see Table VIII and cf. with Table VI.

Note that the range of rare words (hapax legomena) relates to that part of the rank-frequency relation which is closest to it, i.e. for long Chinese texts it relates to the exponential-like regime and not to the Zipfian regime.

Though suggestive, the above theoretical results are still preliminary. The full theory of the rank-frequency relations for Chinese characters should really *explain* how specifically a non-Zipfian relations result from mixing texts that separately hold the Zipf's law.

V. DISCUSSION

A. Summary of results

1. As implied by the rank-frequency relation for characters, short Chinese texts demonstrate the same Zipf's law—together with its generalization to high and low frequencies (rare words)—as short English texts; see section III. Assuming that authors write mainly relatively short texts (longer texts are obtained by mixing shorter ones), this similarity implies that Chinese characters play the same role as English words; see Footnote 5 in this context. Recall from section II that *a priori* there are several factors which prevent a direct analogy between words and characters.

2. As compared to English, there are two novelties of the rank-frequency relation of Chinese characters in long texts.

2.1 The overall frequency of the Zipfian range (the range of middle ranks, where the Zipf's law holds) stabilizes at $\simeq 0.4$. This holds for all texts we studied (written in different epochs, genres with different types of characters; see Tables I, II and Appendix C). A similar stabilization effect holds as well for the overall frequency of the pre-Zipfian range for both English and Chinese texts; see Tables I, II and III.

2.2 There is a range with an exponential-like rank-frequency relation. It emerges for relatively longer texts from within the range of rare words (hapax legomena). The range of ranks, where the exponential-like regime holds, is larger than that of the Zipf's law. But its overall frequency is few times smaller; see Tables I, II and VII.

Both these results are absent for English texts; there the overall frequency of the Zipfian range grows with the length of the text, while there is no exponential-like regime: the Zipfian range end with the hapax legomena; see Table III and Fig. 2.

The results **2.1** and **2.2** imply that long Chinese texts do have a hierarchic structure: there is a group of characters that hold the Zipf's law with nearly universal overall frequency equal to $\simeq$ 0.4, and yet another group of relatively less frequent characters that display the exponential-like range of the rank-frequency relation.

TABLE VIII: The hapax legomena range for 2 Chinese texts; see Tables I and II; cf. with Table V. We compare the relative errors for, respectively, $\hat{r}_k$ (given by (22)) and $\widetilde{r}_k$ in approximating the data r_k; see section III D 2. Here $\widetilde{r}_k$ is defined by (25, 24). For PFSJ the parameters in (24) are $\gamma = 1.302$ and $b = 0.00013$. For SJ: $\gamma = 1.299$ and $b = 0.00026$. It is seen that the relative error provided by $\hat{r}_k$ is always significantly smaller.

Texts	k	1	2	3	4	5	6	7	8	9	10
PFSJ	$\lvert\widetilde{r}_k - r_k\rvert / r_k$	0.179	0.251	0.294	0.324	0.340	0.356	0.365	0.376	0.387	0.392
	$\lvert\hat{r}_k - r_k\rvert / r_k$	0.020	0.017	0.008	0.001	0.002	0.008	0.008	0.011	0.009	0.011
SJ	$\lvert\widetilde{r}_k - r_k\rvert / r_k$	0.093	0.115	0.136	0.153	0.167	0.176	0.181	0.189	0.195	0.198
	$\lvert\hat{r}_k - r_k\rvert / r_k$	0.020	0.018	0.011	0.001	0.007	0.013	0.0011	0.012	0.008	0.004

B. Interpretation of results

Chinese characters differ from English words, since only long Chinese texts have the above hierarchic structure. The underlying reason of the hierarchic structure is to be sought via the linguistic differences between Chinese characters and English words, as we outlined in section II. In particular, the features **4, 6, 7** discussed in section II can mean that certain homographic content characters play multiple role in different parts of a long Chinese text. They are hence distinguished and appear in the Zipfian range of the long text with (approximately) stable overall frequency $\simeq 0.4$. Since this frequency is sizable, and since the range of ranks carried out by the Zipf's law is relatively small, there is a relatively large range of ranks that has to have a relatively small overall frequency; cf. Tables I, II with Table VII. It is then natural that in this range there emerges an exponential-like regime that is related with a faster (compared to a power law) decay of frequency versus rank.

Recall that the stabilization holds as well for the overall frequency of the pre-Zipfian domain both for English and Chinese texts. The explanation of this effect is similar to that given above (but to some extent is also more transparent): the pre-Zipfian range contains mostly function characters, which are not specific and used in different texts. Hence upon mixing the pre-Zipfian range has a stable overall frequency.

The above explanation for the coexistence of the Zipfian and exponential-like range suggests that there is a relation between the characters that appear in the Zipfian range of long texts and homography. As a preliminary support for this hypothesis, we considered the following construction. Assuming that a mixture is formed from separate texts $T_1, ..., T_k$, we looked at characters that appear in the Zipfian ranges of all the separate texts $T_1, ..., T_k$; see Table II for examples. This guarantees that these characters appear in the Zipfian range of the mixture. Then we estimated (via an explanatory dictionary of Chinese characters) the average number of different meanings for these characters. This average number appeared to be around 8, which is larger than the average number of meanings for an arbitrary Chinese character (i.e. when the averaging is taken over all characters in the dictionary) that is known to be not larger than 2 [47].

We should like to stress however that the above connection between the uncovered hierarchic structure and the number of meanings is preliminary, since we currently lack a reliable scheme of relating the rank-frequency relation of a given text to its semantic features; for a recent review on the (lexical) meaning and its disambiguation within machine learning algorithms see [2].

C. Conclusion

The above discussion makes clear that a theory for studying the rank-frequency relation of a long text, as it emerges from mixing of different short texts, is currently lacking. Such a theory was not urgently needed for English texts, because there the (generalized) Zipf's law (21) describes well both long and short texts. But the example of Chinese characters clearly shows that the changes of the rank-frequency relation under mixing are essential. Hence the theory of the effect is needed.

Finally, one of main open questions is whether the uncovered hierarchical structure is really specific for Chinese characters, or it will show up as well for English texts, but on the level of the rank-frequency relation for morphemes and not the words. Factorizing English words into proper morphemes is not straightforward, but still possible.

Acknowledgments

This work is supported by the Region des Pays de la Loire under the Grant 2010-11967, and by the National Natural Science Foundation of China (Grant Nos. 10975057, 10635020, 10647125 and 10975062), the Programme of Introducing Talents of Discipline to Universities under Grant No. B08033, and the PHC CAI YUAN PEI Programme (LIU JIN OU [2010] No. 6050) under Grant No. 2010008104.

Appendix A: Glossary

- Classic Chinese: (*wén yán*) written language employed in China till the early XX (20th) century. It lost its official status and was changed to Modern Chinese since the May Fourth Movement in 1919. The Modern Chinese keeps many elements of Classic Chinese. As compared to the Modern Chinese, the Classic Chinese has the following peculiarities (1) It is more lapidary: texts contain almost two times smaller amount of characters, since the Classic Chinese is dominated by one-character words. (2) It lacks punctuation signs and affixes. (3) It relies more on the context. (4) It frequently omits grammatical subjects.

- Content word (character): a word that has a meaning which can be explained independently from any sentence in which the word may occur. Content words are said to have lexical meaning, rather than indicating a syntactic (grammatical) function, as a function word does.

- Empty Chinese characters—e.g. "几" *(jǐ)* or "已" *(yǐ)* —serve for establishing numerals for nouns, aspects for verbs *etc.* In contrast, to function characters, they cannot be used alone, i.e. they are fully bound.

- Frequency dictionary: collects words used in some activity (e.g. in exact science, or daily newspapers *etc*) and orders those words according to the frequency of usage. Frequency dictionaries can be viewed as big mixtures of different texts.

- Function word (character): is a word that has little lexical meaning or have ambiguous meaning, but instead serves to express grammatical relationships with other words within a sentence, or specify the attitude or mood of the speaker. Such words are said to have a grammatical meaning mainly, e.g. *the* or *and*.

- Hapax legomena: literally means the set of words (characters) that appeared only once in a text. We employ this term in a broader sense: the set of words (characters) that appear in a text only few times. Operationally, this set is characterized by the fact that sufficiently many words (characters) have the same frequency. Texts written by human subjects contains a sizable hapax legomena. This is a non-trivial fact, since it is not difficult to imagine an artificial text (or purposefully modified natural text) that will not contain at all words that appear only once.

- Homophones: two different words that are pronounced in the same way, but may be written differently (and hence normally have different meaning), e.g. *rain* and *reign*.

- Homographs: two different words (or characters) that are written in the same way, but may be pronounced differently, e.g. *shower* [precipitation] and *shower* [the one who shows]. This example is a proper homograph, since the pronunciation is different. Another example (of both homography and homonymy) is *present* [gift] and *present* [the current moment of time]. Note that the distinction between homographs and polysemes is not sharp and sometimes difficult to make. There are various boundary situations, e.g. the verb *read* [present] and *read* [past] may qualify as homograph, but the meanings expressed are close to each other.

- Homonymes: two words (or characters) that are simultaneously homographs and homophones, e.g. *left* [past of *leave*] and *left* [opposite of *right*]. Some homonymes started out as polysemes, but then developed a substantial difference in meaning, e.g. *close* [near] and *close* [to shut (lips)].

- Heteronyms: two homographs that are not homophones, i.e. they are written in the same way, but are pronounced differently. Normally, heteronyms have at least two sufficiently different meanings, indicated by different pronunciations.

- Key-word (key-character): a content word (character) that characterizes a given text with its specific subject. The operational definition of a key-word (key-character) is that in a given text its frequency is much larger than in a frequency dictionary, which was obtained by mixing together a big mixture of different texts.

- Language family: a set of related languages that are believed (or proved) to originate from a common ancestor language.

- Latent semantic analysis: the analysis of word frequencies and word-word correlations (hence semantic relations) in a text that is based on the idea of hidden (latent) variables that control the usage of words; see [64] for reviews.

- Literal translation: word-to-word translation, with (possibly) changing the word ordering, as necessary for making more understandable the grammar of the translated text. This notion contrasts to the phrasal translation, where the meaning of each given phrase is translated. The literal translation can misconceive idioms and/or shades of meaning, but these aspects are minor for gross (statistical) features of a text, e.g. for rank-frequency relation of its words.

- Logographic writing system is based on the direct coding of morphemes.

- Mental lexicon: the store of words in the long-time memory. The words from the mental lexicon are employed on-line for expressing thoughts via phrases and sentences; see [55] for detailed theories of the mental lexicon. Ref. [55] argues that in addition to mental lexicon humans contain a mental syllabary that is activated during the phonologization of a word that was already extracted from the mental lexicon.

- Morpheme: the smallest part of the speech or writing that has a separate (not necessarily unique) meaning, e.g. *cats* has two morphemes: *cat* and *-s*. The first morpheme can stand alone. The second one expresses the grammatical meaning of plurality, but it is a bound morpheme, since it can appear only together with other morphemes.

- Phoneme: a class of speech sounds that are perceived as equivalent in a given language. An alternative definition: the smallest unit that can change the meaning. Hence normally several different sounds (frequently not distinguished by native speakers) enter in a single phoneme.

• Pictogram: a graphic symbol that represents an idea or concept through pictorial resemblance to that idea or concept.

• Polysemes: are related meanings of the same word, e.g. the English word *get* means *obtain/have*, but also *understand* (= *have knowledge*). Another example is that many English nouns are simultaneously verbs (e.g. *advocate* [person] and *advocate* [to defend]).

• Syllable: is the minimal phonetic unit characterized by acoustic integrity of its components (sounds), e.g. the word *body* is composed of two syllables: *bo-* and *-dy*, while *consider* consists of three syllables: *con- -si- -der.* In phonetic languages such as Russian the factorization of the word into syllables (syllabification) is straightforward, since the number of syllables directly relates to the number of vowels. In non-phonetic languages such as English, the correct syllabification can be complicated and not readily available to non-experts. Indo-European languages typically have many syllables, e.g. the total number of English syllables is more 10 000. However, 80 % of speech employs only 500-600 frequent syllables [55]. It was argued, based on psycholinguistic studies, that the frequent syllables are also stored in the long-term memory analogously to mental lexicon [55]. The total number of Chinese syllables is much less, around 500 (about 1200 together with tones) [47, 55]. Syllabification in Chinese is generally straightforward too, also because each character corresponds to a syllable.

• Token: particular instance of a word; a word as it appears in some text.

• Type: the general sort of word; a word as it appears in a dictionary.

• Writing system: process or result of recording spoken language using a system of visual marks on a surface. There are two major types of writing systems: logographic (Sumerian cuneiforms, Egyptian hieroglyphs, Chinese characters) and phonographic. The latter includes syllabic writing (e.g. Japanese hiragana) and alphabetic writing (English, Russian, German). The former encodes syllables, while the former encodes phonemes.

Appendix B: Interference experiments

The general scheme of interference experiments in psychology is described as follows [28, 40]. There are two tasks, the main one and the auxiliary one. Each task is defined via specific instructions. The subjects are asked to carry out the main task simultaneously trying to ignore the auxiliary task. The performance times for carrying out the main task in the presence of the auxiliary one are then compared with the performance times of the main task when the auxiliary task is absent. The interference means that the auxiliary task impedes the main one.

There is a rough qualitative regularity noted in many experiments: interference decreases upon increasing the complexity of the main task or upon decreasing the complexity of the auxiliary task.

The most known example of interference experiment is the Stroop effect, where the main task is to call the color of words. The auxiliary task is not to pay attention at the meaning of those words. The experiment is designed such that there is an incongruency between the semantic meaning of the word and its color, e.g. the word *red* is written in black. As compared to the situation when the incongruency is absent, i.e. the word *red* is written in red, the reaction time of performing the main task is sizably larger. This is the essence of the Stroop effect: the semantic meaning interferes with the color perception.

It appears that the Stroop effect is larger for Chinese characters than for English words; see [28] for a review. This is one (but not the only) way to show that getting to the meaning of a Chinese character is faster than to the meaning of an English word.

Another known interference phenomenon is the word inferiority versus word superiority effect. In English these effects amount to the following [65].

If English-speaking subjects are asked to trace out (and count) a specific letter in a text, they make less errors, when the text is meaningless, i.e. it consists of meaningless strings of letters [27, 66]. This is related to the fact that English words are recognized and stored as a whole. Hence the recognition of words—nd moving from one letter to another— interferes with the task of identifying the letter in a single word, and the English-speaking subjects make more errors when tracing out a letter in a meaningful text. This is the word-inferiority effect.

In contrast, if English subjects is presented a single word for a short amount of time, and is then asked about letters of this word, their answers are (statistically) more correct if the word is meaningful (i.e. it is a real word, not a meaningless sequence of symbols). This word-superiority effect is understood by noting that a single word is recalled and/or remembered better due to its meaning.

In contrast to this, Chinese-speaking adults display the word superiority effect, when the naive analogy with English would suggest the word inferiority. They do less errors in tracing out a given character in a string of meaningful characters, as compared to tracing it out in a list of meaningless pseudo-characters [27].

A possible interpretation of this effect is that, on one hand, the definition of Chinese words and their boundaries is somewhat fuzzy, so that the analogue of the English word-inferiority effect is not effective. On the other hand, the Chinese sentence is perceived as a whole, inviting analogies with the English word-superiority effect.

Note that when the Chinese subjects are asked to trace out a specific stroke within a character we expectedly (and in full analogy with the English situation) get that it is easier for Chinese subjects to trace out the stroke in a meaningless pseudo-character than in a meaningful character [27].

Appendix C: A list of the studied texts

1) Two short modern Chinese texts:

- 昆仑殇, *Kūn Lún Shāng* (KLS) by Shu Ming Bi, 1987, (the total number of characters $N = 20226$, the number of different characters $n = 2047$). The text is about the arduous military training in the troops of Kun Lun mountain.

- 阿Q正传, *Ah Q Zhèng Zhuàn* (AQZ) by Xun Lu, 1922, ($N = 18153$, $n = 1553$). The story traces the "adventures" of a hypocrit and conformist called Ah Q, who is famous for what he presents as "spiritual victories".

2) Two long modern Chinese texts:

- 平凡的世界, *Píng Fán de Shì Jiè* (PFSJ) by Yao Lu, 1986, ($N = 705130$, $n = 3820$). The novel depicts many ordinary people's stories which include labor and love, setbacks and pursue, pain and joy, daily life and huge social conflict.

- 水浒传, *Shuǐ Hǔ Zhuàn* (SHZ) by Nai An Shi, 14th century, ($N = 704936$, $n = 4376$). The story tells how a group of 108 outlaws gathered at Mount Liang formed a sizable army before they were eventually granted amnesty by the government and sent on campaigns to resist foreign invaders and suppress rebel forces.

3) Four short classic Chinese texts:

- 春秋繁露, *Chūn Qiū Fán Lù* (CQF), by Zhong Shu Dong, 179-104 BC, (Vol.**1**-Vol.**8**, $N = 30017$, $n = 1661$). A commentary on the Confucian thought and teachings.

- 僧宝传, *Sēng Bǎo Zhuàn* (SBZ), by Hong Hui, 1124, (Vol.**1**-Vol.**7**, $N = 24634$, $n = 1959$). A commentary on the Taoist thought and teachings. Biographies of great Taoist masters.

- 武经总要, *Wǔ Jīng Zǒng Yào* (WJZ), by Gong Liang Zeng and Du Ding, 1040-1044, (Vol.**1**-Vol.**4**, $N = 26330$, $n = 1708$). A Chinese military compendium. The text covers a wide range of subjects, from naval warships to different types of catapults.

- 虎玲经, *Hǔ Líng Jīng* (HLJ), by Dong Xu, 1004, (Vol.**1**-Vol.**7**, $N = 26559$, $n = 1837$). Reviews various military strategies and relates them to factors of geography and climate.

4) A long classic Chinese text:

- 史记, *Shǐ Jì* (SJ), by Qian Sima, 109 to 91 BC, ($N = 572864$, $n = 4932$). Reviews imperial biographies, tables, treatises, biographies of feudal houses and eminent persons.

Appendix D: Key-characters of the modern Chinese text KLS

Here is the list of the key-characters in the Pre-Zipfian and Zipfian range (Table IX) of the modern Chinese text, 昆仑殇 *Kūn Lún Shāng* (KLS) written by Shu-Ming BI in 1987. The text is about the arduous military training in the troops of Kun Lun mountain.

TABLE IX: Key-characters of the modern Chinese text 昆仑殇 *kūn lún shāng* (KLS).

No.	Rank	Character	Pinyin	English	Frequency
1	14	号	*hào*	horn	157
2	32	军	*jūn*	army	86
3	44	兵	*bīn*	soldier	67
4	113	队	*duì*	troop	38
5	118	令	*lìng*	command	37
6	123	部	*bù*	troop	36
7	152	战	*zhàn*	fight/war	28
8	156	命	*mìng*	command	28
9	180	防	*fáng*	protect	24
10	213	血	*xuè*	blood	20
11	216	立	*lì*	stand straight	20
12	224	功	*gōng*	honor	19
13	225	枪	*qiāng*	gun	19
14	252	官	*guān*	officer	16
15	295	锅	*guō*	pan	14
16	299	保	*bǎo*	protect	14
17	300	卫	*wèi*	protect	13
18	352	营	*yíng*	camp	11
19	355	谋	*móu*	strategy	11
20	360	烧	*shāo*	burn	11
21	394	烈	*liè*	martyr	10
22	407	团	*tuán*	regiment	10

Appendix E: Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test (KS test) [67, 68] is used to determine if a data sample agrees with a reference probability distribution. The basic idea of the KS test is as follows.

We need to determine whether a given set X_1, X_2, ... , X_n is generated by i.i.d sampling a random variable with cumulative probability distribution $F(x)$ (null hypothesis). To this end we calculate the the empiric cumulative distribution function (CDF) $F_n(x)$ for X_1, X_2, ... , X_n:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \le x}, \tag{33}$$

where $I_{X_i \le x}$ equals to 1 if $X_i \le x$ and 0 otherwise. Next we define:

$$D_n = \sup_x |F_n(x) - F(x)|. \tag{34}$$

The advantage of using D_n (against other measures of distance between $F_n(x)$ and $F(x)$) is that if the null hypothesis is true, the probability distribution of D_n does not depend on $F(x)$. In that case it was shown that for $n \to \infty$, the cumulative probability distribution of $\sqrt{n}D_n$

is [67, 68]:

$$P(\sqrt{n}D_n \leq x) \equiv f(x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}. \quad (35)$$

For not rejecting the null hypothesis we need that the observed value of $\sqrt{n}D_n^*$ is sufficiently small. To quantify that smallness we take a parameter (significance level) α ($0 < \alpha < 1$) and define κ_α as the unique solution of

$$f(\kappa_\alpha) = 1 - \alpha. \quad (36)$$

Now the null hypothesis is not rejected provided that

$$\sqrt{n}D_n^* < \kappa_\alpha, \quad (37)$$

where $\sqrt{n}D_n^*$ is the observed (calculated) value of D_n. Condition (37) ensures that if the null hypothesis is true, the probability to reject it is bounded from below by α. Hence in practice one takes, e.g. $\alpha = 0.05$ or $\alpha = 0.01$.

Note however that condition (37) will always hold provided that α is taken sufficiently small. Hence to quantify the goodness of the null hypothesis one should calculate the p-value p: the maximal value of α, where (37) still holds. For the hypothesis to be reliable one needs that p is not very small. As an empiric criterion of reliability people frequently take $p > 0.1$.

We applied the KS test to our data on the character (word) frequencies; see section III A. The empiric results on word frequencies f_r in the Zipfian range $[r_{\min}, r_{\max}]$ are fit to the power law, and then also to the theoretical prediction described in section III C. With null hypothesis that empiric data follows the numerical fittings and/or theoretical results, we calculated the maximum differences (test statistics) D and the corresponding p-values in the KS tests. From Table X one sees that all the test statistics D are quite small, while the p-values are *much larger* than 0.1. We conclude that from the viewpoint of the KS test the numerical fittings and theoretical results can be used to characterize the empiric data in the Zipfian range reasonably well.

[1] R.E. Wyllys, Library Trends, **30**, 53 (1981).

[2] C.D. Manning and H. Schütze, *Foundations of statistical natural language processing* (MIT Press, 1999).

[3] H. Baayen, *Word frequency distribution* (Kluwer Academic Publishers, 2001).

[4] W.T. Li, Glottometrics, **5**, 14 (2002).

[5] N. Hatzigeorgiu, G. Mikros, and G. Carayannis, Journal of Quantitative Linguistics, **8**, 175 (2001).

[6] B.D. Jayaram and M.N. Vidya, Journal of Quantitative Linguistics, **15**, 293 (2008).

[7] L. Lü, Z.K. Zhang and T. Zhou, PLoS ONE, **5**(12), e14139 (2010).

[8] J. Baixeries, B. Elvevag and R. Ferrer-i-Cancho, PLoS ONE, **8**(3), e53227 (2013).

[9] http://en.wikipedia.org/wiki/Zipf's_law
http://ccl.pku.edu.cn/doubtfire/NLP/Statistical_Approach
/Zip_law/references%20on%20zipf%27s%20law.htm

[10] J.B. Estoup, *Gammes sténographique* (Institut Sténographique de France, Paris, 1916).

[11] R. Ferrer-i-Cancho and R. Solé, PNAS, **100**, 788 (2003). M. Prokopenko *et al.*, JSTAT, P11025 (2010).

[12] B. Mandelbrot, *An information theory of the statistical structure of language*, in *Communication theory*, ed. by W. Jackson (London, Butterworths, 1953).
B. Mandelbrot, *Fractal geometry of nature* (W. H. Freeman, New York, 1983).

[13] B. Corominas-Murtra *et al.*, Phys. Rev. E, **83**, 036115 (2011).

[14] D. Manin, Cognitive Science, **32**, 1075 (2008).

[15] G.A. Miller, Am. J. Psyc. **70**, 311 (1957). W.T. Li, IEEE Inform. Theory, **38**, 1842 (1992).

[16] M.V. Arapov and Yu.A. Shrejder, in *Semiotics and Informatics*, v. 10, p. 74 (Moscow, VINITI, 1978). I. Kanter and D. A. Kessler, Phys. Rev. Lett. **74**, 4559 (1995). B.M. Hill, J. Am. Stat. Ass. **69**, 1017 (1974). G. Troll and P. beim Graben, Phys. Rev. E **57**, 1347 (1998). A. Czirok *et al.*, *ibid.* **53**, 6371 (1996). K. E. Kechedzhi *et al.*, *ibid.* **72** (2005).

[17] A.E. Allahverdyan, Weibing Deng and Q.A. Wang, Phys. Rev. E **88**, 062804 (2013).

[18] D. Howes, Am. J. Psyc. **81**, 269 (1968).

[19] R. Ferrer-i-Cancho and B. Elveva, PLoS ONE, **5**, 9411 (2010).

[20] K.H. Zhao, Am. J. Phys. **58**, 449 (1990).

[21] R. Rousseau and Q. Zhang, Scientometrics, **24**, 201 (1992).

[22] D.H. Wang *et al.*, Physica A, **358**, 545 (2005).

[23] S. Shtrikman, Journal of Information Science, **20**, 142 (1994).

[24] Le Quan Ha *et al.*, *Extension of Zipf's Law to Words and Phrases*, Proceedings of the 19th international conference on Computational linguistics, **1**, pp. 1-6, (2002).

[25] Q. Chen, J. Guo and Y. Liu, Journal of Quantitative Linguistics, **19**, 232 (2012).

[26] D. Aaronson and S. Ferres, J. Memory and Language, **25**, 136 (1986).

[27] H.C. Chen, *Reading comprehension in Chinese*, in H.C. Chen & O. J. L. Tzeng (Eds.), Language processing in Chinese (pp. 175- 205). Amsterdam, Elsevier, 1992.

[28] R. Hoosain, *Speed of getting at the phonology and meaning of Chinese words*, in *Cognitive neuroscience studies of Chinese language*, H.S.R. Kao, C.K. Leong and D.G. Gao (eds.) (Hong kong University Press, Hong kong, 2002).

[29] G.K. Zipf, *Selected studies of the principle of relative frequency in language.* (Harvard University Press, Cambridge MA, 1932).

[30] L. Lü, Z.K. Zhang and T. Zhou, Sci. Rep. **3**, 1082 (2013).

[31] C.K. Hu and W.C. Kuo, *Universality and Scaling in the Statistical Data of Literary Works*, POLA Forever, 115-139 (2005).

[32] J. Elliott *et al.*, *Language identification in unknown signals*, Proceedings of the 18th conference on Computa-

TABLE X: Kolmogorov-Smirnov test (KS test) for the fitting quality of our results (texts are defined in Tables I and II). In the KS test, D and p denote the maximum difference (test statistics) and p-value respectively. D_1 and p_1 are calculated from the KS test between empiric data and numerical fitting, D_2 and p_2 are between empiric data and theoretical result, D_3 and p_3 are between numerical fitting and theoretical result; see section III A. Note that for making the testing even more vigorous the presented results for the KS characteristics are obtained in the original coordinates (2); similar results are obtained in logarithmical coordinates (3) that are employed for the linear fitting.

Texts	D_1	p_1	D_2	p_2	D_3	p_3
TF	0.0418	0.865	0.0365	0.939	0.0381	0.912
TM	0.0529	0.682	0.0562	0.593	0.0581	0.568
AR	0.0564	0.624	0.0469	0.783	0.0443	0.825
DL	0.0451	0.812	0.0421	0.865	0.0472	0.761
AQZ	0.0586	0.587	0.0565	0.623	0.0601	0.564
KLS	0.0592	0.578	0.0641	0.496	0.0626	0.521
CQF	0.0341	0.962	0.0415	0.863	0.0421	0.857
SBZ	0.0461	0.796	0.0558	0.635	0.0616	0.538
WJZ	0.0427	0.852	0.0475	0.753	0.0524	0.691
HLJ	0.0375	0.923	0.0412	0.875	0.0425	0.862

tional linguistics, **2**, pp. 1021-1025 (2000).
J. Elliot and E. Atwell, Journal of the British Interplanetary Society **53**, 13 (2000).

[33] H.P. Luhn, IBM J. Res. Devel. **2**, 159 (1958).

[34] S.M. Huang *et al.*, Decision Support Systems, **46**, 70 (2008).

[35] D.M.W. Powers, *Applications and explanations of Zipf's law*, in D.M.W. Powers (ed.), New Methods in Language Processing and Computational Natural Language Learning (NEMLAP3/CONLL98), ACL, 1998, pp. 151-160.

[36] G. Sampson, Linguistics, **32**, 117 (1994).

[37] J. DeFrancis, *Visible Speech: the Diverse Oneness of Writing Systems* (University of Hawaii Press, Honulu, 1989).

[38] J. L. Packard, *The Morphology of Chinese: A linguistic and cognitive approach* (Cambridge University Press, Cambridge, 2000).

[39] K. Turner, *Visualizing Zipf's Law in Japanese*, available at this link:
http://classes.soe.ucsc.edu/cmps161/Winter12/projects/katurner/proj/paper/paper.pdf

[40] R. Hoosain, *Psychological reality of the word in Chinese*, in H.C. Chen and J.L. Tseng (eds.), Language processing in Chinese, pp. 111-130, (Amsterdam, Netherlands, 1992).

[41] I.M. Liu *et al.* Chinese Journal of Psychology, **16**, 25 (1974).

[42] S.H. Hsu and K.C. Huang, Perceptual and Motor Skills, **91**, 355 (2000); *ibid.* **90**, 81 (2000).

[43] X. Luo, *A maximum entropy Chinese character-based parser*. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003.

[44] Wm. C. Hannas, *Asia's Orthographic Dilemma* (University of Hawaii Press, 1997).

[45] C.Y. Chen *et al.*, *Some distributional properties of Madanrin Chinese*, Proceedings of the first Pasific Asia conference on formal and computational linguistics, p. 81 (Taipei, 1993).

[46] http://myweb.tiscali.co.uk/wordscape/wordlist/homogrph .html

[47] N.V. Obukhova, Quantitative linguistics and automatic text analysis (Proc. of Tartu university), **745**, 119 (1986).

[48] N.J.D. Nagelkerke, *A Note on a General Definition of the Coefficient of Determination*, Biometrika, **78** (3), 691 (1991).

[49] M. L. Goldstein, S. A. Morris, G. G. Yen, *Eur. Phys. J. B*, **41**, 255 (2004).

[50] H. Bauke, *Eur. Phys. J. B*, **58**, 167 (2007).

[51] A. Clauset, C. R. Shalizi and M. E. J. Newman, *SIAM Rev.*, **51**, 4 (2009).

[52] R.E. Madsen *et al.*, *Modeling word burstiness using the Dirichlet distribution*, in Proc. Intl. Conf. Machine Learning, 2005.

[53] S. Bernhardsson, L. E. Correa da Rocha, P. Minnhagen, Physica A **389**, 330 (2010); New J. Phys. **11**, 123015 (2009).

[54] T. Hofmann, *Probabilistic Latent Semantic Analysis*, in Uncertainty in Artificial Intelligence, 1999.

[55] W.J.M. Levelt *et al.*, Beh. Brain Sciences, **22**, 1 (1999).

[56] J. Tuldava, Journal of Quantitative Linguistics **3**, 38 (1996).

[57] D. Krallmann, *Statistische Methoden in der stilistischen Textanalyse* (Inaug.-Dissert. Bonn, 1966).

[58] S.K. Baek, S. Bernhardsson and P. Minnhagen, New Journal of Physics **13**, 043004 (2011).

[59] Y. Dover, Physica A **334**, 591 (2004).

[60] E.V. Vakarin and J. P. Badiali, Phys. Rev. E **74**, 036120 (2006).

[61] E.T. Jaynes, IEEE Trans. Syst. Science & Cyb. **4**, 227 (1968).

[62] M. Jaeger, Int. J. Approx. Reas. **38**, 217 (2005).

[63] J. Haldane, Proceedings of the Cambridge Philosophical Society, **28**, 55 (1932).

[64] T. Hofmann, *Probabilistic Latent Semantic Analysis*, in Uncertainty in Artificial Intelligence, 1999.

[65] A.F. Healy and A. Drewnowski, Journal of Experimental Psychology: Human Perception and Performance **9**, 413 (1983).

[66] *Reading Chinese Script: A Cognitive Analysis*, edited by J. Wang, A.W. Imhoff and H.-C. Chen (Lawrence Erl-

baum Associates, New Jersey, 1999).

[67] A.N. Kolmogorov, Giornale dell' Instituto Italiano degli Attuari, **4**, 77 (1933).

[68] P.T. Nicholls, *J. Am. Soc. Information Sci.*, **40**, 379 (1989).