Physica A 444 (2016) 828–837

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Randomness versus specifics for word-frequency distributions

Xiaoyong Yan [a,b], Petter Minnhagen [c,*]

[a] Systems Science Institute, Beijing Jiaotong University, Beijing 100044, China
[b] Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, China
[c] IceLab, Department of Physics, Umeå University, 901 87 Umeå, Sweden

HIGHLIGHTS

- Pseudo-randomness is a characteristic feature in many complex deterministic systems.
- Pseudo-randomness rather than scale-freeness determines word-frequency distributions.
- Pseudo-randomness predicts the word-frequency distributions from minute information.
- Pseudo-randomness deletes linguistic features from the shape of the distribution.
- Pseudo-randomness predictions are conceptually different from curve-fitting.

ARTICLE INFO

Article history:
Received 26 May 2015
Received in revised form 1 October 2015
Available online 3 November 2015

Keywords:
Word-frequency distributions
Zipf's law
Random Group Formation
Maximum entropy

ABSTRACT

The text-length-dependence of real word-frequency distributions can be connected to the general properties of a random book. It is pointed out that this finding has strong implications, when deciding between two conceptually different views on word-frequency distributions, *i.e.* the specific 'Zipf's-view' and the non-specific 'Randomness-view', as is discussed. It is also noticed that the text-length transformation of a random book does have an exact scaling property precisely for the power-law index $\gamma = 1$, as opposed to the Zipf's exponent $\gamma = 2$ and the implication of this exact scaling property is discussed. However a real text has $\gamma > 1$ and as a consequence γ increases when shortening a real text. The connections to the predictions from the RGF (Random Group Formation) and to the infinite length-limit of a meta-book are also discussed. The difference between 'curve-fitting' and 'predicting' word-frequency distributions is stressed. It is pointed out that the question of randomness versus specifics for the distribution of outcomes in case of sufficiently complex systems has a much wider relevance than just the word-frequency example analyzed in the present work.

1. Introduction

The question of trying to understand what *linguistic* information is hidden in the *shape* of the word-frequency distribution has a long tradition. It goes back to the first part of the twentieth century when it was discovered that the word-frequency distribution of a text typically has a broad "fat-tailed" shape, which often can be well approximated with a power law over a large range [1–4]. This led to the empirical concept of Zipf's law which states that the number of words that occur k-times

* Corresponding author.
E-mail address: Petter.Minnhagen@physics.umu.se (P. Minnhagen).

http://dx.doi.org/10.1016/j.physa.2015.10.082
0378-4371/

in a text, $N(k)$, is proportional to $1/k^2$ [2–4]. The question is then what special principle or property of a language causes this power law distribution of word-frequencies and this is still an ongoing research [5–10].

In middle of the twentieth century Simon in Ref. [11] instead suggested that, since quite a few completely different systems also seemed to follow Zipf's law in their corresponding frequency distributions, the explanation of the law must be more general and stochastic in nature and hence independent of any specific information of the language itself. Instead he proposed a random stochastic growth model for a book written one word at a time from beginning to end. This became a very influential model and has served as a starting point for much later works [12–17]. In the 'Simon-view' the shape of the word-frequency distribution does not reflect any specific property of a language but is shaped by a random stochastic element. An extreme random model was proposed in the middle of the twentieth century by Miller in Ref. [18]: the resulting text can be described as being produced by a monkey randomly typing away on a typewriter. However the properties of the monkey book are quite unrealistic and different from a real text [19]. This 'Randomness-view' was recently developed further in a series of paper in terms of concepts like Random Group Formation, Random Book Transformation and the Meta-book [19–23]. A crucial difference, compared to the 'Zips-view', is that the 'Randomness-view' is based on the notion that the shape of the word-frequency distribution is a general consequence of randomness which carries no specific information of the language.

In other words the Zipf-view is leaning more on the idea that a language is a special system and that as a consequence the functional form of the word-frequency distribution reflects some specific property of the language, whereas the Randomness-view maintains that very little specific language information can be extracted from this distribution.

The concept of randomness in a text dates back to at least 1913 and A. Markov [24,25]: Markov demonstrated that even an exquisitely crafted poem like Pushkin's Eugene Onegin, when viewed as a string of letters, contained random features like e.g. how often a randomly chosen letter is followed by a consonant or a vowel. This was at the beginning of what developed into the fundamental statistical concept of Markov chains. This begs the conceptual question of how something crafted with such an amount of intention, purpose and meaning could possibly contain something entirely random. A somewhat related question is hidden within the decimal tail of the number $\pi = 3.14159265\ldots$: The decimal tail of π has a definite cause since it is the ratio between the circumference and diameter of a circle. Thus every decimal in the expansion is solidly given. However, if you pick a decimal place randomly and read off its value and ask yourself what the value of the next decimal might be, then it is with equal probability any of the numbers $0, 1, \ldots, 9$. Thus the poem Eugene Onegin and the number π both display some randomness in spite of their perfectly deterministic cause.

From a statistical point of view the decimal tail of π is pseudo-random and equivalent to a number-series created by throwing a dice with ten fair outcomes. However, if the only thing you know is that the decimal tail of π is equivalent to a pseudo-random series, throwing the dice will not give you any information as to the ratio between the circumference and the diameter of a circle.

Words in a text are random in an analogous fashion; A specific word occurs k times in the text and $N(k)$ specific words occur the same number of times. Suppose you randomly pick a word in the text and that this word occurs k' times. What is the total number of occurrence in the text of the following word? The randomness view argues that this occurrence is random and given by a probability proportional to $N(k)$. The dice $N(k)$ itself can be estimated using the maximum entropy principle [22].

The fact that frequency distributions of possible outcomes for some sufficiently complex deterministic systems reduce to equivalent random distributions is not restricted to words [22,26–30]. Deterministic systems which display random features are termed *pseudo-random*. In the discussion section some more examples are mentioned. However, in the present paper we analyze the consequences for words in a text. The general point is that the ideas of scale-freeness ingrained into the various Zipf's law approaches are superseded by the inherent randomness, which we argue is a very basic property of a written text.

In order to be concrete we will focus on the difference between, on the one hand, a generalized scaling law for word-frequency distributions proposed by Font-Close et al. in Ref. [10] and suggesting a bona fide specific property of a language, and, on the other hand, the general predictions from the Randomness-view [19–23].

We will in the present paper use the following notation: $N_M(k)$ ($N_M(\geq k)$) is the number of distinct words which occur k-times (k-times or more) in a text which in total contains M words. The scaling law proposed in Ref. [10] can be cast into the form $N_M(\geq k) = G(k/M)$.

In Section 2, we first demonstrate directly from raw data that $N_M(\geq k)$ does indeed change shape with text-length in a very systematic manner such that the proposed scaling-form $N_M(\geq k) = G(k/M)$ cannot be conceptually valid. This means that this scaling function cannot be a true specific feature of the word-frequency distribution. In Section 3, we then compare the systematic length dependence of $N_M(\geq k)$ with the predictions from the 'Randomness-view' and indeed find consistent agreement. We elucidate just how little information you need about the language in order to *predict* the characteristic features of the data for the word-frequency. This has a crucial and more far reaching consequence: whenever you need very little information to describe a particular feature, then indeed very little specific information about the system can be extracted from this characteristic feature. In Section 4, we discuss and show that for a distribution $N_M(k) \propto 1/k$ the shape is indeed length-invariant under the randomness (more precisely under the Random Book Transformation assumption [20–23]). In Ref. [21] it was observed that the limit of a very large text by an author seems to approach the limit $N_M(k) \propto 1/k$. This suggests that this approximate scaling should work better the longer the text is. Some concluding remarks are added in Section 5, in particular on the applicability of the 'Randomness view' to a much broader spectrum of complex systems.