Understanding Zipf’s law of word frequencies through sample-space collapse in sentence formation

Stefan Thurner\textsuperscript{1,2,3}, Rudolf Hanel\textsuperscript{1}, and Bernat Corominas-Murtra\textsuperscript{1} \textsuperscript{*}
\textsuperscript{1} Section for Science of Complex Systems, CeMSIIS, Medical University of Vienna, Spitalgasse 23, A-1090, Austria.
\textsuperscript{2} Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA.
\textsuperscript{3}IIASA, Schlossplatz 1, A-2361 Laxenburg, Austria.

The formation of sentences is a highly structured and history-dependent process, in the sense that the probability of using a specific word in a sentence strongly depends on the ‘history’ of word-usage earlier in the sentence. Here we study a simple history-dependent model of text generation where it is assumed that the sample-space – where words are drawn from – reduces as sentences are formed. We are able to explain the approximate Zipf law for word frequencies as a direct consequence of sample space reduction. We empirically quantify the amount of sample-space reduction in the sentences of ten famous English books. The analysis is based on word-transition tables that capture which word follows any given word at least once in a given text. We find a highly nested structure in the transition tables and show that this ‘nestedness’ is linearly related to the observed power law exponents of the word frequency distributions. Remarkably, it is possible to relate global word frequency distributions of long texts to the local, nested structure in sentence formation. On a theoretical level we are able to show that in case of weak nesting Zipf’s law breaks down in a fast transition. Unlike previous attempts to understand Zipf’s law in language the model is not based on multiplicative, preferential, or self-organised critical mechanisms.

PACS numbers: 05.10.-a, 05.40.-a, 05.65.+b
Keywords: stochastic process, word-transition networks, ageing random walks, language formation

I. INTRODUCTION

Written texts show the remarkable feature that the rank ordered distribution of word frequencies follows an approximate power law

\begin{equation}
 f(r) \sim r^{-\alpha},
\end{equation}

where \( r \) is the rank that is assigned to every word in the text. For most texts, regardless of language, time of creation, genre of literature, its purpose, etc., one finds that \( \alpha \sim 1 \), which is referred to as Zipf’s law \cite{1}. In Fig. 1 the word frequency is shown for Darwin’s text, \textit{The origin of species}. The quest for an understanding of the origin of this statistical regularity is going on for almost a century. Zipf himself offered a qualitative explanation based on the efforts invested in communication events by a sender and a receiver \cite{1}. These ideas where later formalised within an information-theoretic framework \cite{2,3,4}. The first quantitative model based on linguistic assumptions about text generation has been proposed by H. Simon \cite{5}. The model assumes that as context emerges in the generation of a text, words that have already appeared in the text are favoured over others. By the simple assumption that words that have previously appeared are added to the text with a probability proportional to their previous appearance (preferential attachment), and assuming that words that have so far not appeared are added at a constant rate, it is possible to derive Zipf’s law, given the latter rate is low. This preferential attachment model has been refined by implementing the empirical fact that the rate of appearance of new words decreases as the length of texts increases \cite{6}. It has been shown in classical works that random typewriting models can lead to Zipf-like distributions of word frequencies \cite{7,8,9,10}. However these works are based on unrealistic assumptions on word-length distributions and lead to unstructured and

\textsuperscript{*}stefan.thurner@meduniwien.ac.at

\textbf{FIG. 1:} Rank ordered distribution of word frequencies for \textit{The origin of species} (blue) shows an approximate power law with a slope of approximately \( \alpha \sim 0.9 \). The model result (red line) explains not only the power law exponent, but also captures details of the distribution. The exponential cutoff can be explained by the randomised version of the model.
FIG. 2: Schematic view of nestedness in sentence formation. (a) Among all the potential $N$ words defining the initial sample-space we choose “wolf” (b). This choice restricts the sample-space for the next word (orange circle) that has to be grammatically and semantically compatible with “wolf”. (c) From this set we chose “howls”, which reduces the sample-space again (red circle) (d), since the next word must now be consistent both semantically and grammatically with “The wolf howls”. The sequence of words show a nested structure. The effect of sample space collapse is also present in the wider context of discourse formation, since a topic and its rhetoric development impose a successive nested constraints of sample space.

uninterpretable texts. However, as we will show, syntactic mechanisms, may play an essential a role in the origin of Zipf’s law in a realistic context. It is important to stress that the detailed statistical study of language properties does not end here; important work beyond Zipf’s law have been put forward, see e.g. [11,14]. Recent studies deal with the detailed dependence of the scaling exponents on the length of the body of text under study [13,16].

Zipf’s law is not limited to word frequencies but appears in countless, seemingly unrelated, systems and processes [17]. Just to mention a few, it has been found in the statistics of firm sizes [18], city sizes [11,19,22], the genome [23], family names [24], income [25,26], financial markets [27], internet file sizes [28], or human behaviour [29], for more examples see [30]. There has been tremendous efforts to understand the origin of Zipf’s law, and more generally the origin of scaling in complex systems. There are three main routes to scaling: multiplicative processes [3,7,31], preferential processes [32,33], and self-organised criticality [35]. Several other mechanisms that are more or less related to these basic routes to scaling have been proposed e.g. in [4,36,37].

Recently a fourth, independent route to scaling has been introduced on the basis of stochastic processes that reduce their potential outcomes (sample-space) over time [41]. These are history-dependent random processes that have been studied in different contexts in the mathematical literature [39,40], and more recently in the context of scaling laws [41,42]. An example of sample-space reducing processes is the following. Think of a set of $N$ dice where dice number 1 has 1 face, dice number 2 has two faces (coin), dice number 3 has three faces, and so on. Dice number $N$ has $N$ faces. Start by picking one of the $N$ dice at random, say dice number $i$. Throw it and record the obtained face value, which was say $k$. Next, take dice number $k−1$ throw it, get $j$, record $j$, take dice number $j−1$, throw it, etc. Keep throwing dices in this way until you throw 1 for the first time. Since there is no dice with less than 1 faces, the process ends here. The sequence of recorded face values in the above prescription $(i,k,j,\ldots,1)$, is obviously strictly ordered or nested, $i > k > j > \ldots > 1$.

In [41] it was shown rigorously that if this process is repeated many times, the distribution of outcomes (face values $1,2,\ldots,N$) is an exact Zipf law, i.e. the probability to observe a face value $m$ in the above process (sequence of throws) is exactly $P_N(m) = m^{−1}$, given we start with $N$ dices. More formally, every dice $N$ has a sample-space, denoted by $\Omega = \{1,2,\ldots,N\}$, which is the number of potential outcomes, i.e. the number of faces of dice $N$. Throwing these dice in the prescribed way gives rise to a sequence of nested sample-spaces

$$\Omega_1 \subset \Omega_2 \subset \ldots \subset \Omega_N$$

The nestedness of sample-spaces in a history-dependent sequence is at the heart of the origin of scaling laws in this type of processes. For details see [41] where it is also shown that if noise is added to the history-dependent processes, the scaling law, $P_N(m) \propto m^{−\lambda}$ is obtained, where $0 < 1 − \lambda < 1$, is the noise level.

In this paper we present a derivation of Zipf’s law of word frequencies, based on a simple model for sentence/discourse formation. The model is motivated by the observation that the process of forming a sentence – or more generally a discourse – is a history-dependent sample-space reducing process. Words are not randomly drawn from the sample-space of all possible words, but are used in strict relations to each other. The usage of specific words in a sentence highly restricts the usage for consecutive words, leading to a nesting (or sample-space reducing) process, similar to the one described above. Sample-space collapse in texts is necessary to
convey meaningful information. Otherwise, any interpretation, even in metaphoric or poetic terms, would become impossible. Let us make the point more concrete with an example for the formation of a sentence, where both grammatical and contextual constraints are at work, Fig. 2. We form the sentence: “The wolf howls in the night”. In principle the first word “The wolf” (ignoring articles and prepositions for the moment) can be drawn from all possible words. Let us assume there exist \( N \) possible words, and denote the respective sample-space by \( \Omega_N = \{1, 2, \ldots, N\} \), where each number now stands for one word. This is schematically illustrated in Fig. 2(a). Given that we chose “The wolf” from \( \Omega_N = \{1, 2, \ldots, N\} \), Fig. 2(b), the next word will now (usually) not be chosen from \( \Omega_N = \{1, 2, \ldots, N\} \), but from a subset of it, Fig. 2(c). Imagine that the subset contains \( L \) words, we have \( \Omega_L \subset \Omega_N \). Typically we expect the subset to contain words that are associated to properties of canines, biological functions, other animals, etc., but not all possible words anymore. Once we specify the second word “howls” \( \in \Omega_L \), context, intelligibility, and grammatical structure further restricts sample-space for the third word to \( \Omega_M \subset \Omega_L \), from which we finally draw “night”. Obviously, the nestedness in the formation of sentences is similar to the example of the nested dice before. Nesting is imposed through grammatical and/or contextual constraints. The role of grammar for nesting is obvious. Typically in English the first word is a noun with the grammatical role of the subject. The fact that the first word is a noun restricts the possibilities for the next word to the subset of verbal phrases. Depending on the particular verb chosen, the words that can now follow are typically playing the grammatical role of the object, and are again more restricted.

We use the terms sample-space reduction and nested hierarchical structure in sentences interchangeably. Not only grammatical structure imposes consecutive restrictions on sample-space of words as the sentence progresses, also the need for intelligibility has the same effect. Without hierarchical structure in sentence formation, their interpretation would become very hard. The sequence of words must display a nested hierarchy that only one or a few alternatives remain, so that meaning can be interpreted [46]. However, nested structures in sentences will generally not be strictly realised. Otherwise the creative use and flexibility of language would be seriously constrained. Sometimes words can act as a linguistic hinge, meaning that it allows for many more consecutive words, than were available for its preceding word. One expects that nestedness will be realised only to some extent. Imperfect nestedness allows for a degree of ambiguity in the linguistic code, and is one of the sources of its astonishing versatility [47].

In this paper we quantify the degree of nestedness (see Methods) from the word-transition matrix (network) in several texts. These networks serve as the input to a simple model for sentence formation that is based on the actual sample-space reduction. We then study the word frequency distributions of these artificially produced texts, and compare them with the distributions of the original texts. For the first time we show that it is possible to relate the topological feature of (local) nestedness in sentence formation to the global features of word frequency distributions of long texts.

![FIG. 3: Section of word-transition matrix](image)

**FIG. 3**: Section of word-transition matrix \( M \) for the 250 words that show the largest sample-space volume of consecutive words \( \Omega \). A black entry \( M_{ij} = 1 \) means that a given word \( i \) is followed by word \( j \). Non-trivial nestedness is seen by the approximate funnel like shape of the density of words. The actual value of the sample-space volume for every word \( i \), \( |\Omega_i| \) is shown in (b), which is obtained by shifting all entries of the lines \( i \) to the rightmost positions. We call (b) the sample-space profile.

**II. MODEL**

We assume a finite vocabulary of \( N \) words. From any given text we obtain an empirical word-transition matrix \( M \). Words are labeled with Latin indices. \( M_{ij} = 1 \) means that in the text we find at least one occasion where word \( j \) directly follows \( i \), if \( M_{ij} = 0 \) word \( j \) never follows \( i \) in the entire text. Figure 2(a) shows the transition matrix for The origin of species. To quantify sample-space for words, note that a line \( i \) in \( M \) contains the set of words, \( \Omega_i = \{k | M_{ik} = 1\} \), that directly follow word \( i \). By \( |\Omega_i| \) we denote the size of \( \Omega_i \), which is the number of different words that can follow \( i \). \( \Omega_i \) is an approximation for the sample-space volume that is accessible after word \( i \) has occurred. Different words have different sample-space volumes, see Fig. 3(b), where the sample-space profile is shown. We parametrize the profile as \( y_i = x^\kappa \), where \( x \) is the volume, \( |\Omega_i| \). We call a system linearly nested if \( \kappa = 1 \) (as in Eq. 2), weakly nested for \( \kappa < 1 \) (as in Fig. 3(b)), and strongly nested if \( \kappa > 1 \). Note that the profile in Fig. 3(b) is actually not well fitted with a power law, the reason for the parametrisation is for a purely theoretical argument that will become clear below. We exclude words that are followed by less than 2 different words in the entire text, i.e. we remove all lines \( i \) from \( M \) for which \( |\Omega_i| \leq 2 \).

Given \( M \) we construct random sentences of length \( L \) with the following sample-space reducing model:
• pick one of the $N$ words randomly. Say the word was $i$. Write $i$ in a wordlist $W$, so that $W = \{i\}$.

• jump to line $i$ in $M$ and randomly pick a word from the set $\Omega_i$. Say the word chosen is $k$; update the wordlist $W = \{i, k\}$.

• jump to line $k$, and pick one of the words from $\Omega_k$; say you get $j$, and update $W = \{i, k, j\}$.

• repeat the procedure $L$ times. At this stage a random sentence is formed.

• repeat the process to produce $N_{\text{sent}}$ sentences.

In this way we get a wordlist with $L \times N_{\text{sent}}$ entries, which is a random book that is generated with the word-transition matrix of an actual book. From the wordlist we obtain the word frequency distribution $f_{\text{model}}$. The present model is similar to the one in [31] but differs in three aspects: it allows for non-perfect nesting $n < 1$, it has no explicit noise component, and it has a fixed sequence (sentence) length.

III. RESULTS

We simulate the model with computer simulations, specifying $L = 10$, and $N_{\text{sent}} = 100,000$. We use 10 randomly chosen books\footnote{In particular we use An American tragedy, by Theodore Dreiser, The Origin of Species, Decent of Man, and Different Forms of Plants, by Ch. Darwin, Tale of two Cities, and David Copperfield, by Ch. Dickens, Romeo and Juliet, Henry V, and Hamlet, by W. Shakespeare, and Ulysses by J. Joyce. Vocabulary varies from $N = 3,102$ (Romeo and Juliet) to 22,000 (Ulysses) words.} from the Project Gutenberg [48]. For every book we determine its vocabulary $N$, its matrix $M$, its $\Omega_i$ for all words, its nestedness $n(M)$, and the exponent of the rank ordered word frequency distribution $\alpha$ (least square fits to $f(r)$, fit range between the $5 \leq r \leq 200$). $f(r)$ is seen for in Fig. 1 (blue), the exponent is $\alpha \sim 0.90$. We run the model for the parameters of every individual book. Using the empirical $\Omega_i$ for the model ensures that we have exactly the same sample-space profile and nestedness as the book.

The distribution obtained from the model $f_{\text{model}}$ is clearly able to reproduce the approximate power law exponent for The origin of species, $\alpha_{\text{model}} \sim 0.86$ (same fit range). Moreover it captures details of the distribution $f$. For large values of $r$ in $f_{\text{model}}(r)$ a plateau is forming before the exponential finite size cutoff is observed. Both, plateau and cutoff can be fully understood with the randomised model.

In Fig. 1(a) we compare the $\alpha$ exponents as extracted from the books with the model results $\alpha_{\text{model}}$. The model obviously explains the actual values to a large extend, slightly underestimating the actual exponents. We get a correlation coefficient of $\rho = 0.95 \ (p < 3.7 \times 10^{-5})$. In Fig. 1(b) we show that nesting $n(M)$ is related to the exponents $\alpha$ in an approximately linear way. We test the hypothesis that by destroying nestedness the exponents will vanish. Using the randomised $M_{\text{rand}}$ we find $\alpha_{\text{rand}} \sim 0.19 \pm 0.03$ (same fit range), which effectively destroys the power law. To validate our assumption that word ordering is essential, we computed

![FIG. 4: $\alpha$ exponents from rank ordered word frequency distributions of 10 books versus model results for $\alpha_{\text{model}}$ (a). Clearly the model explains the actual values to a large extend. (b) $\alpha$ exponents versus nestedness $n(M)$ of the 10 books. $\alpha_{\text{model}}$ exponents versus the sample-space profile parameter $\kappa$. For large vocabularies $N = 100,000$, at $\kappa \sim 1$ a fast transition from the weak nesting to the strong nesting regime occurs, where we find $\alpha_{\text{model}} \sim 0$ and $\alpha_{\text{model}} \sim 1$, respectively. Weak and strong nesting profiles are schematically indicated. For smaller (realistic) $N$ the transition appears at $\kappa < 1$, and $\alpha_{\text{model}}$ covers a range between $\sim 0.85$ and $1.1$ in the scaling phase, which fits the empirical range seen in (a).](image-url)
the model rank distributions by using the transposed matrix $M^T$, meaning that we reverse the time flow in the model. We find two results. First, the correlation between the exponents of the books $\alpha$ and the model $\alpha_{\text{model}}$ vanishes, reflected by an insignificant correlation coefficient $\rho = 0.47$ ($p = 0.17$). Second, the exponents (averaged over the 10 books) are consistently smaller $\alpha_{\text{model}} = 0.85 \pm 0.03$, than for the correct time flow, where we get $\alpha_{\text{model}} = 0.90 \pm 0.06$. Finally we try to understand the importance of the sample-space profile on the scaling exponents. For this we generate a series of $M$ matrices that have a profile parametrized with a power $\kappa$. In Fig. 4(c) the model exponents $\alpha_{\text{model}}$ from these artificially generated $M$ are shown as a function of $\kappa$, for various sizes of vocabulary $N$. For $\kappa < 1$ (weak nesting) we find exponents $\alpha_{\text{model}} \approx 0$, i.e. no scaling law. For large $N$ at $\kappa = 1$ a fast transition to $\alpha_{\text{model}} \approx 1$ (Zipf) occurs. For smaller $N$ we find a more complicated behaviour of the transition, building a maximum exponent at a $\kappa < 1$. The range of book exponents $\alpha$ ranges between 0.85 and 1.1, which is exactly the observed range for realistic vocabulary sizes $N \sim 1,000-10,000$. We verified that neither changes in $L$ nor $N_{\text{sent}}$ influence the described results significantly.

IV. DISCUSSION

In this paper we focused on the fundamental property of nesting in any code that conveys meaningful information, such as language. We argue that if nesting was not present one would easily end up in confusing situations as described in La Biblioteca de Babel by J. L. Borges, where a hypothetical library owns all books composed of all possible combinations of characters filling 410 pages. Nesting is not realised strictly in actual language. We defined and quantified a degree of nestedness in the linguistic code, and found that it correlates perfectly with the power law exponents of word frequencies in the overall text. As expected, texts have a well defined, but not strictly nested structure, which might arise from a compromise of specificity (to convey unambiguous messages) and flexibility (to allow a creative use of language). By use of a simple sample-space reducing model we have shown that nestedness explains the emergence of Zipf’s law. More precisely, we were able to relate the emergence of scaling laws with topological structure of the word transition matrix, or “phasespace”. The result is remarkable since the $M$ does not encode any information how often $j$ follows $i$, it just tells that $j$ followed $i$ at least once in a text. It is also remarkable that no (non-local) preferential, multiplicative, or self-organized critical assumptions are needed to understand the observed scaling. The class of sample-space reducing processes offer an independent route to scaling that might have a wide range of applications for aging processes. In statistical physics it is known that processes that successively reduce their phasespace as they unfold are characterised by power law or stretched exponential distribution functions. These distributions directly arise as a consequence of phasespace collapse.

V. METHODS

To characterise the hierarchical structure of a text with a single number, we define its nestedness $n$ as a property of $M$ by

$$n(M) = \left\langle \frac{\Omega_i \cap \Omega_j}{\min(|\Omega_i|, |\Omega_j|)} \right\rangle_{(i,j)},$$

where the average is taken over all word pairs $(i,j)$. Nestedness is a number between 0 and 1, and specifies to what extent sample-space reduction is present in the text. A strictly nested system, like the one shown in Eq. (2), has $n(M) = 1$. In linguistic terms strictly nestedness is clearly unrealistic. Don’t confuse strict nesting with strong or weak nesting defined above, which are properties of the sample-space profile. For statistical testing we construct a randomised version $M_{\text{rand}}$ that keeps the statistics of $M$ (non-zero entries in every line) but destroys its nestedness. $M_{\text{rand}}$ is generated by permuting the entries $M_{ij}$ individually for every line $j$.

Acknowledgements. We thank Bo Liu for preprocessing the books. This work was supported by the Austrian Science Fund FWF under KPP23378FW.

[1] G. K. Zipf (1949) Human Behavior and the Principle of Least Effort. Addison-Wesley.
[2] Harremoës, P, Topsøe, F (2001) Maximum Entropy Fundamentals. Entropy 3, 191-226.
[3] R. Ferrer i Cancho, R. V. Sole (2003) Least effort and the origins of scaling in human language. PNAS 100, 788-791.
[4] Corominas-Murtra, B, Fortuny, J, Solé, RV (2011) Emergence of Zipf’s law in the evolution of communication. Phys Rev E 83, 036115.
[5] H. A. Simon (1955). On a class of skew distribution functions. Biometrika, 42, 425-440.
[6] D. H. Zanette, M. A. Montemurro (2005). Dynamics of text generation with realistic Zipf’s distribution. Journal of Quantitative Linguistics 12, 29-40.
[7] B. Mandelbrot B (1953) An informational theory of the statistical structure of language. In: Jackson W, editor. Communication theory. London: Butterworths. 486 p.
[8] W. Li (1992) Random texts exhibit Zipf’s-law-like word frequency distribution. IEEE T Inform Theory 38: 1842-1845.
[9] G. A. Miller (1957) Some effects of intermittent silence.
Am J Psychol 70: 311-314.

[10] G. A. Miller, N. Chomsky (1963) Finite models of language users. In: Luce RD, Bush R, Galanter E, editors. Handbook of Mathematical Psychology. New York: Wiley, volume 2, pp. 419-491.

[11] Kosmidis, K., Kalampokis, A., and Argyrakis, P. Statistical Mechanical Approach to Human Language. Physica A 366, 495-502. (2006)

[12] Atkinson, Q.D., Meade, A., Venditti, C., Greenhill, S.J., Pagel, M. Languages Evolve in Punctuational Bursts. Science, 319, 588-588. (2008)

[13] M. A. Serrano, A. Flammini, F. Menczer Modeling Statistical Properties of Written Text, PLoS ONE 4 e5372. (2009)

[14] Decker, E-H Kerkhoff, A-J and Moses, M-E (2007) ONE 293, 1251-1258.

[15] D. H. Zanette, M. A. Montemurro, , Universal Entropy of Word Ordering Across Linguistic Families PLoS One 6(5): e19875 2011.

[16] Blank, A, Solomon, S (2000) Power laws in cities population, financial markets and internet sites (scaling in systems with a variable number of components). Physica A 287, 279-288.

[17] Kawamura, K and Hatano, N. Universality of Zipf’s law (2002) J Phys Soc Jpn 71(5) 1211-1213.

[18] R. L. Axtell, (2001) Zipf distribution of US firm sizes Science 293, 5536, 1818.

[19] Makse, H-A, Havlin, S, Stanley, H-E (1995) Modelling urban growth patterns. Nature 377, 608-612.

[20] Krugman, P (1996) Confronting the Mystery of Urban Hierarchy. J Jap Int Econ 10, 399-418.

[21] Blank, A, Solomon, S (2000) Power laws in cities population, financial markets and internet sites (scaling in systems with a variable number of components). Physica A 287, 279-288.

[22] Decker, E-H Kerkhoff, A-J and Moses, M-E (2007) PLoS ONE 2, 934.

[23] Stanley H, Buldyrev S, Goldberger A, Havlin S, Peng C, et al. (1999) Scaling features of noncoding DNA. Physica A 273, 1-18.

[24] Zanette, D-H, Manrubia, S-C (2001) Vertical transmission of culture and the distribution of family names. Physica A 295, 1-8

[25] Pareto, V. (1896) Cours d’Economie Politique, (Droz, Geneva).

[26] Okuyama, K., Takayasu, M. and Takayasu, H. (1999) Physica A 269, 125

[27] Gabaix, X, Gopikrishnan, P, Plerou, V, Stanley, E H (2003) A theory of power-law distributions in financial market fluctuations. Nature 423, 267-270.

[28] Reed, W J, Hughes, B D (2002) From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature. Phys Rev E 66, 067103-067106.

[29] S. Thurner, M. Szell, R. Sinatra, (2012) Emergence of Good Conduct, Scaling and Zipf Laws in Human Behavioral Sequences in an Online World PLoS ONE 7, e29796.

[30] Newman, M E J (2005) Power Laws, Pareto Distributions and Zipf’s Law. Contemp Phys 46 323-351.

[31] Solomon, S, Levy, M (1996) Spontaneous Scaling Emergence in Generic Stochastic Systems. Int J Mod Phys C 7, 745-751.

[32] Malcai, O, Biham, O, Solomon, S (1999) Power-law distributions and Lévy-stable intermittent fluctuations in stochastic systems of many autocatalytic elements. Phys Rev E 60, 1299-1303.

[33] Lu, E T, Hamilton, R J (1991) Avalanches of the distribution of solar flares. Astrophys J 380, 89-92.

[34] Barabasi, A-L, Albert, R (1999) Emergence of scaling in random networks. Science 286, 509-512.

[35] Bak P, Tang, C, Wiesenfeld, K (1987) Self-organized criticality: An explanation of the 1/f noise. Phys Rev Lett 59, 381-384.

[36] Saichev, A, Malevergne, Y, Sornette, D (2008) Theory of Zipf’s Law and of General Power Law Distributions with Gibrat’s Law of Proportional Growth. Lecture Notes in Economics and Mathematical Systems. Berlin, Heidelberg, New York: Springer.

[37] Pietronero, L, Tosatti, E, Tosatti, V, Vespignani, A (2001) Explaining the uneven distribution of numbers in nature: The laws of Benford and Zipf. Physica A 293, 297-304.

[38] Thurner, S, Tsallis, C (2005) Nonextensive aspects of self-organized scale-free gas-like networks. Europhys Lett 72, 197-203.

[39] Corominas-Murtra, B, Solé, R V (2010) Universality of Zipf’s law. Phys Rev E 82, 011102.

[40] Montroll, E-W and Shlesinger, M-F. On 1/f noise and other distributions with long tails. (1982) PNAS 79, 3380-3383.

[41] Corominas-Murtra, B, Hanel, R, Thurner, S (2014) Understanding Zipf’s law with playing dice: history-dependent stochastic processes with collapsing sample-space have power-law rank distributions. arXiv:1407.2775.

[42] Kac, M. (1989) A history-dependent random sequence defined by Ulam. Adv in Appl Math 10 270-277.

[43] Clifford, P, Stirzaker, D (2008) History-dependent random processes. Proc R Soc A 464, 1105-1124.

[44] R. Hanel, S. Thurner (2013) Generalized (c,d)-entropy of complex stochastic systems with Gibrat’s Law of Proportional Growth. Europhys Lett 99, 745-751.

[45] R. Hanel, S. Thurner, M. Gell-Mann (2014). How multiplicity of random processes determines entropy: derivation of the maximum entropy principle for complex systems. PNAS 111, 6905-6910.

[46] R. Hanel, S. Thurner (2013) Generalized (c,d)-entropy and aging random walks Entropy 15, 5324-5337.

[47] Partee, B H (1976) Montague Grammar. New York: Academic Press

[48] Fortuny, J, Corominas-Murtra, B (2013) On the origin of ambiguity in efficient communication. J of logic, language and information 22 249-267.

www.gutenberg.org