2021 IEEE XIIth International Conference on Electronics and Information Technologies (ELIT)

Large-Scale Studies of the Repetition Characteristic for Different Models of Symbolic Sequences

O. S. Kushnir
Optoelectronics and Information Technologies Department
Ivan Franko National University of Lviv
Lviv, Ukraine
o.s.kushnir@lnu.edu.ua

L. B. Ivanitskyi
Optoelectronics and Information Technologies Department
Ivan Franko National University of Lviv
Lviv, Ukraine
lubomyr.ivanitskiy@gmail.com

A. I. Kashuba
General Physics Department
Lviv Polytechnic National University
Lviv, Ukraine
andrii.i.kashuba@lpnu.ua

M. R. Mostova
Optoelectronics and Information Technologies Department
Ivan Franko National University of Lviv
Lviv, Ukraine
mariana.mostova@lnu.edu.ua

V. B. Mykhaylyk
Diamond Light Source
Didcot, UK
vitaliy.mykhaylyk@diamond.ac.uk

***Abstract*—Using a standard suffix-tree algorithm, we study the repetition characteristic $v(t)$, which has been introduced by F. Golcher, for different models of random symbolic sequences and compare it with the corresponding data obtained for natural-language texts and program codes. The character of $v(t)$ function, the saturated repetition parameter v_0 averaged at large enough times t and the appropriate standard deviation Δv_0 are examined for 144 natural, random and randomized texts of different types. The main peculiarities of repetitions peculiar for the Simon, Markov and Miller's monkey text-generating models are analyzed. The results obtained for these analytically tractable models can be useful for developing mathematical fundamentals of the repetition characteristic.**

***Keywords*—repetition characteristic, symbolic sequences, random texts, Simon model, Markov chains, monkey texts**

I. Introduction

Quantitative regularities for the repetitions observed in symbolic sequences or (natural or artificial) texts are important for better understanding of these sequences and elucidating the nature of their underlying generating mechanisms. Moreover, the above regularities can hopefully be used when distinguishing uncorrelated random symbolic sequences from heavily correlated natural texts. Recently, F. Golcher has offered a repetition characteristic $v(t)$ [1]. It can be expressed through dependence of the number n of internal nodes of the suffix tree, which is built upon a given symbolic sequence, on the current length t of this sequence. More specifically, we have

$$v(t) = \frac{n(T_t)}{t}, \; t = 1, \ldots, L \qquad (1)$$

In (1), T_t denotes an incomplete suffix tree built on a sub-text S_t of a complete symbolic sequence (or a text) S, of which length $|S|$ is equal to L. In quite equivalent terms, $v(t)$ is nothing but the number of completed n-grams which have been repeated in the text at least once. According to [1], the $v(t)$ dependence reveals a saturation (or a 'limit', although with no relevant mathematical implications being studied) for the natural-language texts, with the saturated value v_0 being close to ½. On the other hand, randomly generated symbolic sequences, randomized natural texts and, quite surprisingly, even the 'texts' comprising source program codes manifest no equilibrium repetition parameter or, at least, have the v_0 values incompatible with the value ½ [1].

Although there have been a number of further inquiries on the repetition characteristic (see [2]–[4]), the subject remains unclear in too many aspects. Since there is still no solid mathematical background of the repetition characteristic, we consider it relevant to study the $v(t)$ function for as many artificial symbolic sequences of different types as possible, and gather the corresponding information in order to comprehend better this characteristic and its discriminating power.

II. Method and Text Models

The $v(t)$ characteristic was calculated using a standard suffix-tree algorithm [5] (see also [4]). With a few deliberate exceptions, we measured the $v(t)$ function for single texts rather than textual corpora, in order not to deal with the boundary effects and 'inhomogeneities' arising where individual texts are merged. In strict terms, we suspect that the $v(t)$ characteristic and, in particular, its saturated value v_0 can become ill-defined for the merged texts which represent partly 'incoherent' structures.

We studied the $v(t)$ dependences for the following theoretical (or empirical) text models: (i) the Miller's monkey texts [6]–[9], (ii) the Simon model [10], (iii) the model of Markov chains (see [11], [12]), and (iv) the Chomsky texts (see, e.g., [13]). These random texts were contrasted with (v) the natural-language texts, (vi) the randomized natural texts (see also [2]–[4], [14]), and (vii) the texts of program codes. In general, we examined $\Sigma = 144$ texts of different types, of which total length amounted to $\Sigma L \approx 3.6 \cdot 10^8$ characters.

A monkey text represents a simplified text model in which each of M letters is typed at random and has the same relative frequency f in a text. Besides of such 'canonical' monkey texts, we also constructed generalized monkey texts [4], where the letters were chosen randomly, although their

frequencies could be different from each other ($b = f_{\max}/f_{\min} \neq 1$, with $f_{\max}$ and $f_{\min}$ being respectively the maximal and minimal relative frequencies of letters). We varied the b parameter in the interval $b = 1 \div 300$. We adopted three different regimes for the rank−frequency dependences $f(r)$ (with r denoting a rank): (i) a simple linear $f(r)$ function, (ii) a logarithmic $f(r)$ function, which is a good approximation for the natural language [15], and (iii) an arbitrary (random) $f(r)$ dependence, with the natural normalization condition $\sum_{i=1}^{M} f_i = 1$ satisfied. Particular cases of the alphabet sizes $M = 1$, 2, 5, 15 and 26 were dealt with.

For all of our monkey texts (see Table I), we put the frequency of a word separator (a space) to be equal to 0.2, which corresponded to the average word length typical for English.

In a classical Simon text-generation model [10]−[16], one takes at random either a 'new' word from a reservoir (with a constant probability p; typically $p \ll 1$) or an 'old' word already available in a text (with the complementary probability $(1 - p)$), and every word token present in the text have the same probability to be chosen at a current position t in the text. An alternative model can be two reservoirs of different classes of words, with different probabilities p_1 and p_2, which successfully simulates some realities of the natural language (see [17], [18]).

TABLE I. TEXT MODELS AND TEXT TYPES FOR WHICH THE REPETITION CHARACTERISTIC IS EXAMINED

Text Model or Text Type	Number of Texts	Text Length, Characters
Miller's Monkey Model	21	$10^6 \div 2 \cdot 10^7$
Simon Model	65	$3 \cdot 10^5$, $2.2 \cdot 10^6$
Markov Model	23	10^6, $5.5 \cdot 10^5$
Chomsky Model	1	$2.5 \cdot 10^6$
Natural Texts	10	$7 \cdot 10^4 \div 2.5 \cdot 10^6$
Randomized Natural Texts	4	10^6, $2.5 \cdot 10^6$
Program Codes	20	$5 \cdot 10^4 \div 7 \cdot 10^6$

It is known that the Simon model yields in a too fast (linear) increase of the vocabulary V of words with increasing t, which is not typical for the natural texts where the vocabulary grows sublinearly. A simple modification of the above algorithm has been suggested [16], [19]. Here the probability p decreases with increasing t according to a power law ($p(t) \sim p_0 t^{\theta - 1}$, with $\theta < 1$), in order to provide a well-known Heaps law for the vocabulary ($V(t) \sim t^{\theta}$). We conventionally call these Simon texts as 'sublinear'.

Another modified Simon model used by us is the texts with memory (see [20]). The latter effect is mimicked such that the probability of selecting one of the 'old' words from the text becomes higher when this word is located at the position $(t - i)$, which is closer to the current position t (i.e., at smaller i's). Three memory functions are employed: an exponential one which models short-range correlations in a text, and stretched exponential or power-law functions, which correspond to long-range correlations (see, e.g., [14], [21], [22]). Besides the time range, the memory effect is also characterized by its strength, i.e. the amplitude factor involved in the memory-governing function. Larger amplitude would imply longer series of repetitions of the same words in a text.

In total, 65 Simon texts were studied (Table I), which combine different modeling properties described above ('linear' or 'sublinear', a single word reservoir or two reservoirs, and availability of a memory of some type or absence of memory effect).

We define a Markov text as a chain of symbols (letters or words), in which each symbol is chosen at random according to its predefined conditional probability [11], [12]. The latter depends on the symbol itself and N symbols preceding it. These probabilities can be calculated in advance from some natural text or a corpus of texts. According to our conventional definition, generation of a Markov text of order N is governed by the conditional probabilities of symbols that depend on their $(N - 1)$ predecessors. For example, the text of the first order ($N = 1$) is based upon single-symbol probabilities only, with considering no preceding symbols. Therefore, this text generated on the level of letters should be similar to a generalized monkey text with the letter frequencies adopted from the same natural text.

For our studies, we have generated the Markov texts that correspond to the both levels of letters (the text length $L \approx 1.0 \cdot 10^6$ characters) and words ($L \approx 5.5 \cdot 10^5$ characters), with the orders $N = 1 \div 10$ (see Table I).

Finally, we define a Chomsky text as a natural-language text in which a word separator and some letter replace each other. In spite of this small perturbation, the structure and the number of words are notably changed, if compared with the original text. Nonetheless, we still expect only minor changes in the $v(t)$ characteristic.

Natural texts of English fiction ("Redgauntlet" by W. Scott, "The Lord of the Rings" by J. R. R. Tolkien, "Harry Potter and the Sorcerer's Stone" by J. K. Rowling, etc.) were downloaded from a free source Project Gutenberg (see Table I). We studied also randomized versions of some of these natural texts. The latter were obtained through shuffling on the linguistic levels of symbols or words. 10^9 elementary random permutation (shuffling) cycles were applied to a text, thus destroying intentional repetitions characteristic for the natural language and leaving only stochastically caused ones.

The program texts prepared by us included source codes written in C, C++ and Java. Note that any program text represents a mixture of two constituents, a program itself which is written in a computer code and comments written in a natural language. To eliminate a possible impact of this intrinsic inhomogeneity of program texts, we studied both the whole texts, as well as the appropriate subtexts associated with the code and the comments, which were separated from each other. The program texts under test involved also 3 merged texts (including a source code of Linux).

Since, by definition, the models of random texts mentioned above included no punctuation marks, we preprocessed the natural texts in the same manner to make the appropriate comparison more consistent. Moreover, lower- and upper-case letters were not distinguished. However, the program texts were not preprocessed in this way. One of the reasons is that most of the non-alphabetical symbols involved in a computer code bear a semantic load, like words in a human language.

III. Results and Discussion

A. General

Considering a wide scope of our studies and a confined volume of article, we present only a brief account of the main results. A complete report on our findings will be published elsewhere.

For each text, we have found the $\nu(t)$ function. It turns out that the latter can manifest the following typical patterns: (i) a gradual saturation occurring at $t > t_0$ [Notice that we have $t_0 \approx (2 \div 5) \cdot 10^4$ characters for most of the natural texts [4]], (ii) pronounced oscillations observed even at the largest times t (as with the canonical monkey texts with large enough alphabets [1]), and (iii) an irregular, hardly predictable behavior. To adhere to quantitative rather than this qualitative terminology, we continue with the notion of 'saturated ν_0 value' though redefine it as an average over a finite number of data points, typically over 10÷50 points taken at $t > 5 \cdot 10^4$. We omit a more strict term 'limit of $\nu(t)$ function' as too claiming in its mathematical sense.

To evaluate formally a convergence of the repetition characteristic, we have introduced the corresponding standard deviation $\Delta\nu_0$, which is calculated on the same dataset as the ν_0 value. Disregarding a quite possible, though hardly analytically tractable, situation when the $\nu(t)$ function is clearly 'divergent' or unpredictable, one can hope that the $\Delta\nu_0$ parameter behaves as if its underlying stochastic process were Gaussian. According to the central limit theorem, we then should have $\Delta\nu_0 \sim L^{-1/2}$ for the dependence on the size L of stochastic system. In practice, we indeed arrive at the inverse power-law $\Delta\nu_0(L) \sim L^{-\alpha}$, with the exponent α close to the theoretical value, $\alpha \approx 0.6$ (see Fig. 1). Note that our texts can have the sizes that differ by orders of magnitude (see Table I). Then, instead of the $\Delta\nu_0$ measure, it would be better to compare the convergent properties of $\nu(t)$ for different texts, using a normalized standard deviation with accounting for text length,

$$\Delta\nu_{0n} = \Delta\nu_0 L^{1/2}/\Sigma \qquad (2)$$

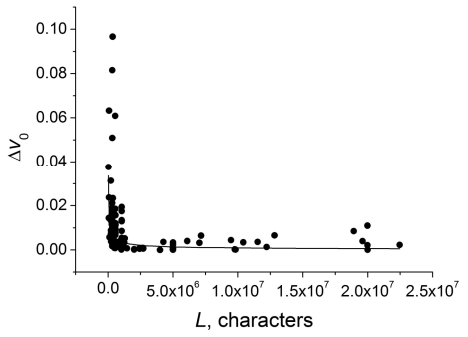

Fig. 1. Dependence of standard deviation $\Delta\nu_0$ on the text length L for all of our texts. The line corresponds to power-law fitting $\Delta\nu_0(L) \sim L^{-\alpha}$, with the exponent $\alpha \approx 0.6$

where the total number of texts Σ in the denominator represents an optional factor used only for convenience. Further on, we employ the $\Delta\nu_{0n}$ parameter as a main indicator of convergent behavior of the $\nu(t)$ dependences obtained for different texts. In particular, Table II displays the average repetition parameters ν_0 and the normalized standard deviations $\Delta\nu_{0n}$, which have been averaged over all the texts belonging to a given text type.

TABLE II. Average ν_0 Repetition Parameters and Normalized Standard Deviations $\Delta\nu_{0N}$ Averaged for Different Text Types

Text Type	Average Value ν_0	Normalized Standard Deviation $\Delta\nu_{0n}$ Averaged Over a Text Type
Miller's Monkey Texts	0.30÷1.00	0.009
Simon Texts without Memory	0.46÷0.60	0.028
Simon Texts with Memory	0.55÷0.86	0.047
Markov Texts with Letter Chains	0.40÷0.52	0.050
Markov Texts with Word Chains	0.48÷0.53	0.039
Chomsky Text	0.51	0.004
Natural Texts	0.49÷0.55	0.016
Natural Texts Randomized by Letters	0.35	0.013
Natural Texts Randomized by Words	0.47	0.006
Program Texts with Comments	0.55÷0.64	0.048
Pure Program Codes	0.62÷0.64	0.047
Pure Comments	0.55÷0.57	0.013

B. Natural, Randomized Natural and Program Texts

In accordance with the data [1]–[4], the natural texts are characterized by a clearly convergent $\nu(t)$ behavior, with the ν_0 value compatible with ½. Since there are no indications in the literature that this type of symbolic sequences can reveal deviations from convergence, while the highest $\Delta\nu_{0n}$ value found for the natural texts is equal to 0.035, one can roughly adopt that the values $\Delta\nu_{0n} \geq 0.04$ can be tentatively treated as those signalizing of possibly divergent behavior of the repetition characteristic (see Table II).

As can be easily predicted, the receipt of a Chomsky text does not imply any serious perturbation of an initial natural text. Therefore their saturated repetition parameters almost coincide.

Randomization of natural texts makes the saturation property of the $\nu(t)$ function more evident, since it drops the $\Delta\nu_{0n}$ value down (see Table II). Randomization performed on the level of letters destroys both the word order and the internal structure of words. In other words, all the 'regular' repetitions, which are peculiar for the human language, are eliminated. Only 'accidental' repetitions survive, which are determined by combinatorial rules acting at a given alphabet size and a given frequency distribution for the letters. Therefore, this randomization type is much more efficient in lowering the ν_0 value, when compared with the word-based randomization that preserves the word structure. Of course, the randomization does not change the distribution of letters' frequencies but eliminates all of those extra repetitions which are associated with the long-range correlations in texts.

Repetitions are always more characteristics for the program texts than the natural ones, irrespective of whether we consider pure program codes or whole programs with inclusions of comments. This is especially obvious in the case of pure program codes (see Table II). The latter conclusion becomes even more striking if one reminds that the computer code includes more characters than the human language and, in a hypothetical case of merely stochastic repetitions, this would have resulted in less rather than more repetitions (see also the results for the monkey texts

discussed below). Finally, even the pure comments written in human language are more iterative than the natural texts. This is evident since the language of comments is poorer and more strict and explicit than the language of literary fiction.

Following from the Δv_{0n} value for the program texts (see Table II), there are not unfounded doubts concerning their saturation property. Of course, one can speculate that the typical time regions t_0 of saturation are, for some reasons, significantly larger than those for the natural texts, although this assumption could hardly be agreed with the earlier empirical findings [1], [3]. It would indeed be natural to explain this situation by the fact that any program code represents mingled natural and computer languages (see Introduction). However, the data of Table II denies this simple explanation because the Δv_{0n} values are almost the same for the whole program texts and the codes depleted of comments.

Notice also that, among the three different programming languages considered in this study, C and C++ can be characterized as 'more regular' (with the Δv_{0n} averages being equal to 0.031 and 0.024, respectively). For these texts, one can observe particular natural language-like examples ($\Delta v_{0n} \approx 0.020$) – or (more or less evident) oscillatory $v(t)$ behavior. In this respect, the behavior of $v(t)$ for Java is less predictable and we have the average deviation $\Delta v_{0n} \approx 0.085$, which is surely out of the saturation range.

C. Monkey Texts

A general pattern observed for the canonical monkey texts (i.e., the texts in which all the letters have the same frequency) is as follows: $v(t)$ manifests a saturation for small alphabet sizes M and increasingly intense oscillations at larger M. Since the oscillation period is nearly constant on a logarithmic time scale, it is difficult to make conclusions concerning the evolution of oscillation amplitude even with so relatively long texts as those studied in the present work.

When degeneracy of the frequencies is broken ($b \neq 1$), the oscillation amplitude and the deviation Δv_{0n} decrease, so that the system approaches steadily a saturated-repetition state. The larger alphabet, the larger difference of letter frequencies is needed in order to reach the saturation. For instance, in the case of $M = 26$ corresponding to the natural-language alphabet, we have $\Delta v_{0n} \approx 0.046$ at $b = 1$, $\Delta v_{0n} \approx 0.033$ at $b = 3$, and $\Delta v_{0n} \approx 0.003$ at $b = 150$. Moreover, it seems that the exact shape of the rank−frequency dependence (e.g., logarithmic or linear $f(r)$ functions) does not matter and the oscillations are eliminated simply at larger b's.

The saturated v_0 value (or the average level of $v(t)$ oscillations) decreases with increasing M. This is evident since larger M's imply smaller-scale random repetitions of combinations of letters. The most important effect found for the random texts is influence of the b parameter on the saturated repetition value v_0. In particular, as seen from Fig. 2, in case of $M = 2$ we have $v_0 \approx 0.78$ at $b = 1$ and $v_0 \approx 0.96$ at $b = 150$ (cf. with the data [4]). [Notice that, as with all the other $v(t)$ illustrations given in the present work, Fig. 2 displays only a small initial part of the $v(t)$ plot on a larger linear scale.] In other words, in what the repetition rate is concerned, a radical increase in the frequency gradient $b = f_{\max} / f_{\min}$ makes the monkey text at $M = 2$ similar to the text with the alphabet $M = 1$ ($v_0 \approx 0.999$, thus implying a unit value).

Quantitatively, the effect weakens with increasing M. For instance, v_0 increases from 0.31 to 0.34 only in case of $M = 26$ at the same change in the b parameter. Finally, the $v_0(M)$ dependence obtained using our data and the results [4] (see the insert in Fig. 2) is satisfactorily described by the power-law function $v_0(M) = AM^{-B}$, with $A \approx 0.98$ and $B \approx 0.36$ (the Pearson correlation coefficient 9.996). These parameters agree very well with the data reported in [4] for a smaller dataset.

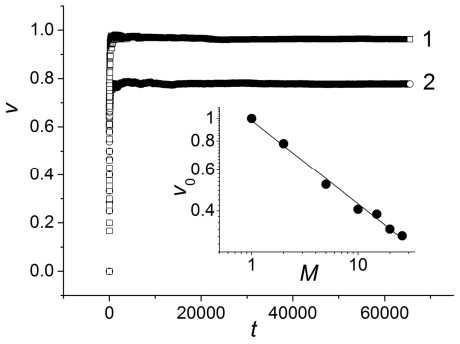

Fig. 2. Dependences of repetition characteristic $v(t)$ for the monkey texts with the alphabet size $M = 2$ and the gradient parameters $b = 1$ (curve 2) and $b = 150$ (curve 1). Insert shows dependence of average saturated value v_0 on the alphabet size M at $b = 1$, with the line corresponding to linear fit in the log-log scale

D. Simon and Markov Texts

Relatively high values of the repetition parameter typical for the Simon texts (see Table II) can be associated with the very principle of this generation model, while the memory still enhances this effect. Another feature of these texts is pronounced irregularities and breaks of their $v(t)$ functions (see Fig. 3), which are more typical for the Simon model with memory. In general, such attributes as the number of word reservoirs or linear (sublinear) vocabulary growth affect weakly the $v(t)$ characteristic. Nonetheless, the Simon texts with two reservoirs manifest the repetition properties closer to those of the natural texts.

Finally, availability of memory increases the normalized standard deviation Δv_{0n}, so that the corresponding Simon texts can hardly reveal a true $v(t)$ saturation. Maybe, the only exception is the Simon texts with (relatively weak) short-range exponential memory, for which the saturation property is still questionable. As a conclusion, the Simon texts with memory are substantially different from the natural texts in their repetition-related characteristics.

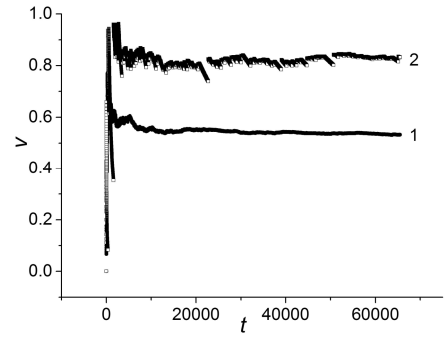

Fig. 3. Dependences of repetition characteristic $v(t)$ for the Simon texts: linear vocabulary growth and two word reservoirs (curve 1), and linear vocabulary growth, a single word reservoir and stretched exponential memory function (curve 2)

Fig. 4 summarizes the main results obtained for the Markov texts, i.e. the average repetition parameter v_0 and the appropriate standard deviation Δv_0 as functions of the chain order N. Note that, instead of Δv_{0n}, we use its non-normalized counterpart Δv_0, since the lengths of all the texts of a given type are the same.

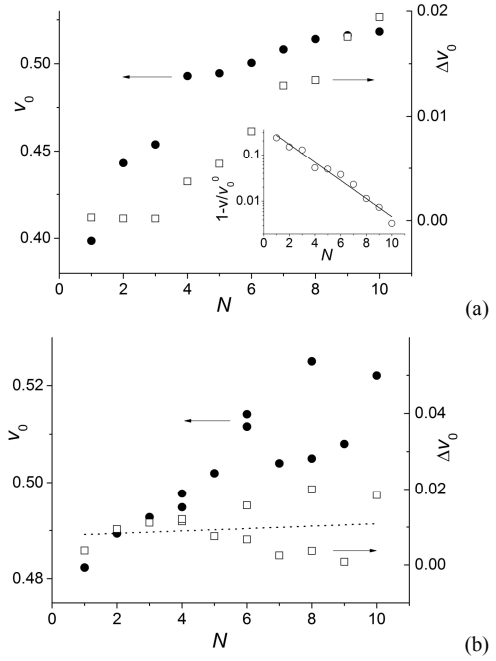

Fig. 4. Dependences of average parameter v_0 (left vertical scale) and standard deviation Δv_0 (right vertical scale) on the chain order N for the Markov texts generated on the linguistic levels of letters (panel a) and words (panel b). A straight line in panel (b) corresponds to linear fit of $\Delta v_0(N)$. Insert in panel (a) shows the dependence $[1 - (v_0/ v_0^0)]$ vs. N on a semi-logarithmic scale, with the line corresponding to the fit with exponential function $v_0(N) \sim v_0^0[1 - \exp(-N/N_0)]$ (see the text)

We begin with discussing the saturated value v_0 for the Markov texts. The v_0 parameter obtained for the letter-based Markov chains of the lowest order $N = 1$ ($v_0 \approx 0.40$) is compatible with the corresponding values typical for the randomized natural texts ($v_0 \approx 0.35$) and the monkey texts generated with a natural language-based rank dependence $f(r)$ at $M = 26$ ($v_0 \approx 0.34 \div 0.35$ – see Table II). A similar situation happens for the word-based Markov texts: we have $v_0 \approx 0.48$ at $N = 1$, which is close to that observed for the natural texts randomized by words ($v_0 \approx 0.47$). Notice also that the latter two figures are closer to each other simply because the both processes of randomization and generation of Markov texts correspond to the same initial natural text. In other words, the randomized natural texts, the generalized monkey texts and the Markov texts generated at $N = 1$ have the following properties in common: (i) they lack any correlations or semantics-driven repetitions and, moreover, (ii) they are generated issuing from the same underlying letter-frequency distribution. It comes at no surprise that their v_0's are very similar.

The empirical data illustrated in Fig 4a testifies that, with increasing N, the v_0 parameter tends to a limit determined by the v_0 value peculiar for the natural texts (~ 0.52). [By the way, the same should take place for the case of word-based Markov texts, although the appropriate results are less explicit against the background of experimental inaccuracies, so that one cannot arrive at decisive conclusion about the exact $v_0(N)$ function (see Fig. 4b).] Then it would be natural to assume that the $v_0(N)$ dependence for the letter-based Markov texts represents a canonical exponential 'transient process', which is known, e.g., from radio engineering:

$$v_0(N) \sim v_0^0[1 - \exp(-N/N_0)] \quad (2)$$

Here v_0^0 denotes a limiting 'equilibrium' v_0 value (i.e., the value for the appropriate natural text) and N_0 implies a characteristic N-'time' needed to partially reach this equilibrium. If we take $v_0^0 = 0.52$, a linear fit on the semi-logarithmic $v_0(N)$ scale results in $N_0 \approx 5.1$, with the Pearson correlation 0.988 (see Fig. 4a, insert). Hence, the empirical data confirms our assumption.

The main tendencies found for the $v_0(N)$ dependence and the appropriate quantitative differences existing between the Markov texts generated on the levels of letters and words can easily be explained from the most general reasoning:

1. While the letter-based Markov texts of the lowest orders N simulate only underlying letter frequency distribution of the natural texts, the higher-order texts mimic also some portion of (at least short-range) correlations, which are tracked through the conditional probabilities of n-grams with ever larger n's.

2. At the lowest orders N, the word-based Markov texts reveal larger v_0's than the letter-based texts, since the former texts preserve the internal structure of words. With increasing N, the word-based texts would simulate not only the word structure but the word order. Finally, one can hope that a hypothetical limit of natural texts can be formally reached at $N \to \infty$.

Summing up, the approach of Markov chains works such that it mimics formally some essential properties of the human language and, in particular, its repetition patterns (see also the discussion [11]). [Of course, this does not imply that, at large enough N, the Markov texts can acquire a true semantic load which is peculiar for the natural texts.] In general, the Markov-chain process can be considered as a reciprocal of randomization of natural texts, i.e. as a kind of 'anti-randomization' process.

Finally, one can see that any type of 'memory' introduced into random text models gives rise to increasing repetition parameter, as seen from the properties of both Markov and Simon generating models.

Regarding the convergence of the repetition characteristic, the letter-based and word-based Markov chains reveal somewhat different behaviors: while the standard deviation Δv_0 for the latter texts remains nearly constant and almost does not depend on N, Δv_0 increases with increasing N for the former texts (cf. Fig. 4a with Fig. 4b). Notice that the maximal Δv_0's (at $N = 10$) for the levels of letters and words correspond respectively to the normalized values $\Delta v_{0n} \approx 0.120$ and 0.083. This is why one can cast doubt upon the saturation property of the Markov texts of high orders. The reasons of this feature are still unclear and need further investigations.

IV. Conclusions

Let us summarize the main results derived in the present work. Following the definition of repetition characteristic $v(t)$ given by F. Golcher in 2007 and implementing the

corresponding calculations through the suffix-tree algorithm, we have examined the $v(t)$ dependences for 144 texts of different types, which comprise natural-language texts, program codes and randomized natural texts, as well as a number of well-known random-text models such as Miller's monkey, Simon and Markov schemes for generating symbolic sequences.

In order to grasp the equilibrium repetition rate and the saturation property of the $v(t)$ function, we work in terms of the parameter v_0 averaged over large enough times t and the appropriate normalized standard deviation Δv_{0n}. Three main types of $v(t)$ characteristics can be distinguished: a regular converging behavior, a regular (though oscillatory) behavior (which is typical for the 'canonical' monkey texts with large alphabets), and an irregular behavior with vague converging properties.

Among principled results, one can mention a dependence of saturated value v_0 on the gradient of letter frequencies in the generalized monkey texts. This implies that the repetition characteristic for the random texts can still be associated with the information entropy of underlying coding system (cf. with the conclusions [4]). Moreover, we have demonstrated that introduction of any memory effect in the text-generating model increases the repetition parameter v_0. This fact is confirmed by the data derived for the Simon and Markov models.

An important quantitative result has been obtained for the Markov model of random texts. We have found a saturated exponential increase in the v_0 parameter, which occurs with increasing order N of the Markov texts with letter-based chains. Then the $v_0(N)$ function is described by a transient process which should finally develop into the repetition pattern peculiar for a natural text. According to both the working algorithm and the consequences for the repetition characteristic, the Markov receipt for generating symbolic sequences can be termed as a kind of 'anti-randomization', i.e. a process reciprocal to the process of randomization of natural texts.

REFERENCES

[1] F. Golcher, "A stable statistical constant specific for human language texts," pp. 1–6, 2007. https://www.academia.edu/5986557/A_Stable_Statistical_Constant_Specific_for_Human_Language_Texts

[2] D. Kimura and K. Tanaka-Ishii, "Study on constants of natural language texts, J. Language Processing, vol. 21, pp. 877–895, 2014.

[3] K. Tanaka-Ishii and S. Aihara, "Computational constancy measures of texts – Yule's K and Renyi's entropy," Computational Linguistics, vol. 41, pp. 481–502, 2015.

[4] O. S. Kushnir, L. B. Ivanitskyi, A. I. Kashuba, M. R. Mostova, and V. B. Mykhaylyk, "Repetition characteristic for single texts," in 5th International Conference on Computational Linguistics and Intelligent Systems (Kharkiv, Ukraine). CEUR Workshop Proceedings, 2021, 13 pp. (at press).

[5] E. Ukkonen, "On-line construction of suffix-trees," Algorithmica, vol. 14, pp. 249–260, 1995.

[6] W. Li, "Random texts exhibit Zipf's-law-like word frequency distribution," IEEE Trans. Inform Theory, vol. 38, pp. 1842–1845, 1992.

[7] R. Ferrer i Cancho and R. V. Solé, "Zipf's law and random texts," Adv. Complex Syst., vol. 5 pp. 1–6, 2002.

[8] R. Ferrer-i-Cancho and B. Elvevåg, "Random texts do not exhibit the real Zipf's law-like rank distribution," PLoS ONE, vol. 5, e9411 (10 pp.), 2010.

[9] S. Bernhardsson, S. K. Baek, and P. Minnhagen, "A paradoxical property of the monkey book," J. Statist. Mechanics: Theory and Experiment, vol. 2011, P07013 (13 pp.), 2011.

[10] H. A. Simon, "On a class of skew distribution functions," Biometrika, vol. 42, pp. 425–440, 1955.

[11] I. Kanter and D.A. Kessler, "Markov processes: linguistics and Zipf's law," Phys. Rev. Lett., vol. 74, pp. 4559–4562, 1995.

[12] A. Cohen, R. N. Mantegna, and S. Havlin, "Numerical analysis of word frequencies in artificial and natural language texts," Fractals, vol. 5, pp. 95–104, 1997.

[13] O. S. Kushnir, V. O. Buryi, S. V. Grydzhan, L. B. Ivanitskyi, and S. V. Rykhlyuk, "Zipf's and Heaps' laws for the natural and some related random texts," Electronics and Information Technologies, vol. 9, pp. 94–105, 2018.

[14] M. A. Montemurro and P. A. Pury, "Long-range fractal correlations in literary corpora," Fractals, vol. 10, pp. 451–461, 2002.

[15] S. M. Gusein-Zade, "Frequency distribution of letters in the Russian language," Problemy Peredachi Informatsii, vol. 24, pp. 102–107, 1988.

[16] D. H. Zanette, "Statistical patterns in written language," Centro Atomico Bariloche, pp. 1–87, 2012, http://fisica.cab.cnea.gov.ar/estadistica/2te/

[17] R. Ferrer i Cancho and R. V. Solé, "Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited," J. Quant. Linguist., vol. 8, pp. 165–173, 2001.

[18] M. Gerlach and E. G. Altmann, "Stochastic model for the vocabulary growth in natural languages," Phys. Rev. X, vol. 3, 021006 (10 pp.), 2013.

[19] D. H. Zanette and M. A. Montemurro, "Dynamics of text generation with realistic Zipf's distribution," J. Quant. Linguist., vol. 12, pp. 29–40, 2005.

[20] C. Cattuto, V. Loreto, and V. D. P. Servedio, "A Yule-Simon process with memory," Europhys. Lett., vol. 76, pp. 208–214, 2006.

[21] W. Ebeling and A. Neiman, "Long-range correlations between letters and sentences in texts," Physica A, vol. 215, pp. 233–241, 1995.

[22] J. W. Kantelhardt, "Fractal and multifractal time series," in Mathematics of Complexity and Dynamical Systems, R. A. Meyers, Ed. New York: Springer, 2012, pp. 463–487.