# Repetition Characteristic for Single Texts

Oleh Kushnir[a], Lyubomyr Ivanitskyi[a], Andriy Kashuba[b], Mariana Mostova[a] and Vitaliy Mykhaylyk[c]

[a] *Department of Optoelectronics and Information Technologies, I. Franko National University of Lviv, 107 Tarnavskyi Street, Lviv, 79017, Ukraine*
[b] *Department of General Physics, Lviv Polytechnic National University, 12 Bandera Street, Lviv, 79046, Ukraine*
[c] *Diamond Light Source, Harwell Campus, Didcot, OX11 0DE, UK*

**Abstract**

The repetition characteristic $v(t)$ introduced by F. Golcher is calculated for single natural texts in different languages and random Miller's monkey texts. It is shown that the saturated $v(t)$ value $v_0$ obtained at the largest times $t$ is not governed by single-character information entropy and parameter of semantic load of a text. The parameter $v_0$ manifests intra-language variations comparable with inter-language ones. In a slightly modified calculation regime, it provides a powerful tool for detecting even small repeated textual fragments.

**Keywords 1**

Golcher's repetition characteristic, textual constants, information entropy, semantic load

## 1. Introduction

Nowadays statistical linguistic methods offer useful and practical solutions for many important problems of natural language processing. The examples are Zipf and Heaps laws for the word statistics of texts [1–8], intermittance of words [7, 9, 10], correlation properties and fluctuation effects [9, 11–13], word networks [14], and the methods for extracting keywords in natural texts based upon different statistical characteristics [15–20].

It is well known that 'static' statistical regularities like the Zipf rank-frequency dependence do not embrace many key properties of real natural language. Moreover, it turns out that these regularities of natural texts can be similar to those of the simplest stochastic models like a Miller's monkey text [21–23]. This should imply that a true theoretical model of a human language must involve not only a specific character of frequencies of linguistic elements but also their order in a text. In this relation, repetitions in texts represent an important matter (see, e.g., Ref. [8]).

In 2007, F. Golcher has introduced an interesting textual characteristic associated with repetitions of symbols in a text (or a corpus of texts), which is considered as a formal symbolic sequence in (discrete) time $t$ [24]. Golcher's $v(t)$ characteristics, as a function of current position $t$ of symbol in a text, represents in fact the number $V$ of completed repetitions occurring for the first time, divided by $t$:

$$v(t) = V(t)/t. \quad (1)$$

In other words, the $V$ parameter concerns the 'types' rather than 'tokens' of the repeated n-grams, i.e. it counts a size of 'vocabulary' of the completed repetitions of n-grams having arbitrary lengths. It has been empirically demonstrated [24] that, at moderately small $t$'s (in practice, at $t > 10^4$ characters or so), the $v(t)$ function begins to 'saturate', and there is an equilibrium limiting $v_0$ value (of the order of ½) for the combined natural-text corpora written in a number of Indo-European languages. The exact $v_0$ value has been found to depend on both the language and the writing system. On the contrary, the

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: oleh.kushnir@lnu.edu.ua (O. Kushnir); lubomyr.ivanitskiy@gmail.com (L. Ivanitskyi); andriykashuba07@gmail.com (A. Kashuba); mariana.mostova@lnu.edu.ua (M. Mostova); vitaliy.mykhaylyk@diamond.ac.uk (V. Mykhaylyk)
ORCID: 0000-0002-1545-7666 (O. Kushnir); 0000-0003-3650-3892 (L. Ivanitskyi); 0000-0002-5445-1064 (A. Kashuba); 0000-0002-8330-0046 (M. Mostova); 0000-0003-0106-2724 (V. Mykhaylyk)

$v_0$ limit can hardly be observed for artificial or random texts of several types (e.g., for computer program codes or Miller's monkey texts with the sizes of alphabet compatible with that of human languages). As a consequence, the behavior of $v(t)$ and the $v_0$ value itself can be used as an indirect criterion for distinguishing semantically loaded natural texts from artificial texts and semantically empty random symbolic sequences.

The next, larger-scale studies on the subject [25, 26] have extended the scope of languages (including Chinese and Japanese) and the corpus lengths (up to $10^9$ characters). They have been mainly involved in searching for a so-called 'constancy measure', which is invariant for a given text and does not depend on its length, at least at the lengths $t$ larger than a certain threshold, with possible applications in author identification and stylometry. The results [25, 26] testify that the $v_0$ value remains invariant for the corpus lengths as large as $10^7$ characters, although there are doubts that $v_0$ represents a constancy measure for still larger corpora ($t \sim 10^7$–$10^9$ characters). Besides, a modified version of the approach [24] has been applied to quantize stylistic similarity of texts [27].

We think that the behavior of repetition characteristic in natural and random texts deserve its further theoretical and empirical investigations. Our arguments are as follows:

- It would be worthwhile to concentrate on single texts only rather than corpora. Even in the first study [24] is has been admitted that merging of different texts can produce 'bumps' in the $v(t)$ function. The influence of this effect on the $v_0$ has not been studied and we cannot exclude that, in general, it would be better to consider $v(t)$ as a characteristic of a single text but not a corpus or a language as a whole.
- It has been stated in the work [25] that the $v_0$ value can be somehow linked with the redundancy of language or writing system. This raises the question: Can the repetition parameter $v_0$ be dependent on the information entropy governing distribution of character frequencies?
- It would be tempting to find another quantitative characteristics, e.g. the characteristics associated with semantic load of a text, which could somehow predict the $v(t)$ behavior or, at least, correlate with the equilibrium $v_0$ value.
- It is interesting whether modified definitions of repetition characteristic can be of some use.
- Finally, we wish to study the resources of $v(t)$ functions in detecting artificial repetitions in a text, i.e., a kind of self-plagiarism in it (see also the discussion [24]).

## 2. Materials and Methods

### 2.1. Texts

We studied three types of texts: 50 natural texts in English and 34 natural texts in the human languages belonging to different families (taken from the source [28]), as well as random Miller's monkey texts. All of the natural texts were literary fiction, except for 'The Origin of Species' by C. Darwin and 'Relativity: The Special and General Theory' by A. Einstein. The languages included Germanic, Romanic, Slavic and Ugro-Finnic languages, as well as Arabic, Chinese and Japanese.

All the texts were in *.txt format, with UTF-8 coding. The sizes of English texts varied from 70 kB to 2.4 MB, with the mean size $t_{max} \approx 630$ kB and the standard deviation 450 kB (in UTF-8, the same figures describe the numbers of characters, including spaces). The texts in different languages were characterized by the mean size 330 kB and the standard deviation 170 kB. The natural texts were preprocessed such that the difference between lowercase and uppercase letters was disregarded, while the numerals, special characters and punctuation were eliminated. This was done for correct comparison of repetitions in the natural texts with those occurring in the random monkey texts since, by construction, the latter texts included no numerals, special characters and punctuation.

By definition, monkey texts are random sequences of letters and spaces, in which all the letters have the same frequency. Besides of these texts, we also constructed 'generalized' Miller's monkey texts, in which different letters are chosen at random though they could have different frequencies. Here we considered the simplest case when the rank-frequency dependence had a linear character. This case can be exhaustively described by a single parameter, a gradient $b = f_{max}/f_{min}$, where $f_{max}$ and $f_{min}$ are respectively the maximal and minimal frequencies of letters. Then the common monkey texts are recovered in the limiting case $f_{max} = f_{min}$, i.e. $b = 1$. We studied different alphabet sizes $M = 2$, 5, 10 and

20. For each of these $M$'s, we chose particular cases of $b$ = 1, 10, 50, 100 and 300. Note that the gradient $b$ typical for the natural texts is not too far from the value $b = 300$, although the rank-frequency dependence is logarithmic rather than linear. In order to reduce potentially enormous size of monkey texts measured in words, we used a relatively high frequency of space as a word separator, $f_{\_} = 0.2$. This frequency is close to that typical for the natural texts in English. In other terms, our texts were not 'true' monkey texts, for which all the characters should have the same probability, including the separator of words. All of our 24 monkey texts had a fixed size, $2.5 \cdot 10^6$ characters.

## 2.2. Calculation of repetition characteristic

To illustrate better the essence of the repetition characteristic $v(t)$, we consider a simple 'text' ISN'T_IT_FUNNY taken from the work 'Winnie-the-Pooh' by A. A. Milne (see Ref. [24]). The ordered list of completed n-grams, which are repeated at least once in the text, is as follows: I, T_, _ and N. Here the term 'completed' implies that the algorithm counts the repeated n-gram only if there is no longer continuing repeated n-gram ahead. The n-grams I, T_, _ and N are counted respectively at the positions $t$ = 8, 10, 10 and 13 in the text, and the $v(t)$ values at these positions are equal to 1/8, 3/10 and 4/13 (see formula (1)). Figure 1a shows a more detailed $v(t)$ plot. Here the descending regions in the $v(t)$ curve appear since the $V$ parameter remains the same (i.e., no new completed repetitions occur), while the current time $t$ in denominator of formula (1) increases.

Since the number of n-grams of arbitrary lengths increases according to a power law with increasing length of symbolic string, calculation of the number of repeated n-grams requires huge computational and storage facilities. This problem can be solved and the $v(t)$ dependences can be calculated, using a standard Ukkonen's algorithm for finding suffix trees [29]. It enables one to compute the repetition characteristic linear in time. The relevant procedures are illustrated by the suffix tree in Figure 1b. It has been built with an online visualization facility [30]. Since we work with relatively short single texts rather than large corpora, there is no need in recoursing to smarter algorithms like the construction of suffix arrays (cf. with Ref. [25]).

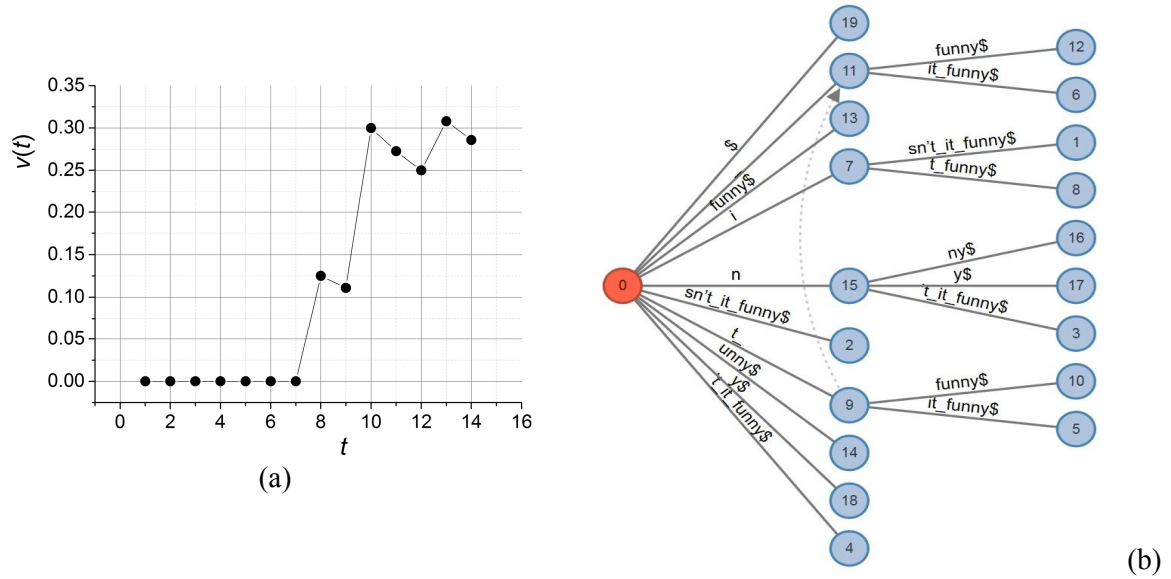


**Figure 1**: (a) $v(t)$ dependence for the text ISN'T_IT_FUNNY and (b) visualization of the Ukkonen's algorithm for finding the corresponding suffix tree [30]: the completed repeated n-grams correspond to internal nodes of the trie, while the symbol "$" indicates end of the text

To probe possible practical resources of the repetition characteristic $v(t)$, we generalized it according to three different ways:

- A text can be analyzed on the linguistic levels of characters, as originally meant by Golcher [24], or on the level of words (see also Ref. [26]). This results in the two alternative calculation regimes.

- In the original algorithm, each repetition in a text is scored as a single point, irrespective of the length n of repeated n-gram. However, one may be interested in progressive scoring of longer repetitions, e.g. with the aim of finding easier longer repetitions. As a result, we compare the two alternative regimes in which a repeated n-gram is scored as either 1 point or n points.
- Originally, the Patricia suffix tree algorithm implies searching for the internal nodes of the trie, which corresponds to scoring only the first repetitions or, quite equivalently, scoring n-gram 'types'. However, it would be interesting to estimate the overall scope of repetitions in a text, which requires a regime of scoring each repetition, i.e. counting each of repeated 'tokens' of a given n-gram.

Since the three regimes described above can be combined independently, we arrive at eight different regimes for the calculations of $v(t)$ parameter. These extended calculations have been performed for 10 natural texts written in English.

## 2.3. Information entropy of character distribution

It is known (see, e.g., Ref. [31]) that a message coded in some language can be viewed as an information flow and its expected rate measured in bits per character is given by the Shannon entropy $H$:

$$H = -\sum_i p_i \log_2 p_i \ . \qquad (2)$$

Here $i = 1 \div N$, $N$ denotes the size of alphabet, and the probability $p_i$ of a character $i$ is estimated by its relative frequency $f_i$. In other words, we have $p_i = f_i$, with $f_i = F_i / t_{\max}$, $F_i$ being the absolute frequency of character and $t_{\max}$ the total text length. Note that we leave aside a more complex definition of the entropy based upon conditional entropy and n-gram representation [32], because this definition demands much more computational efforts. Then the measure (2) is in fact a unigram-based entropy estimated from a finite-size probability mass function $\{ p_i \}$ (see also the work [33]).

The main idea underlying the entropy calculations is finding its possible correlation with the repetition parameters (see also Section 1).

## 2.4. Parameters evaluating a semantic load of text

It is known that semantically poor stopwords are 'stochastically uniformly' distributed in a text, while semantically richer content words and, especially, keywords manifest a so-called effect of intermittance (or clusterization) [9, 10]. This can easily be quantized through introducing a parameter $R_i = \Delta\tau_i / \langle \tau_i \rangle$, where $\langle \tau_i \rangle$ and $\Delta\tau_i$ imply respectively the average waiting time and its standard deviation for a given word $i$ in text. Here the waiting times of a word represent discrete time intervals (i.e. numbers of another words happening in text) between two neighboring occurrences of this word (see, e.g., Ref. [10]). As a matter of fact, we have $R_i \sim 1$ for most of the words in texts and $R_i >> 1$ for only some of them which are keywords, while the situation $R_i < 1$ is typical maybe for the words with the lowest absolute frequencies $F_i$, which lack satisfactory statistics. Let us introduce the parameter $R = \langle R_i \rangle$ averaged over all the words satisfying the condition $F_i \geq F_{\min}$ (see also the work [34]). Considering typical sizes of our texts, we adopt $F_{\min} = 10$ in the present work. It can be easily proved that $R$ as a parameter of whole text is somewhat larger than unity, $R > 1$, while the appropriate standard deviation $\Delta R$ remains large enough ($\Delta R < 1$ or $\Delta R \sim 1$). This refers only to meaningful natural texts, whereas the relations $\Delta R \approx 1$ and $\Delta R << 1$ are typical for the meaningless random character sequences [35], because those 'texts' have no 'keywords'. Put another way, one can consider $R$ and $\Delta R$ as characteristics of a given text, which represent cumulative measures of its semantic load.

Note that we have also tried to employ more refined techniques for extracting keywords, which reveal some advantage over the $R_i$ parameter (see, e.g., Ref. [36] and references therein). However, the data obtained by us on this basis are qualitatively similar. Ascribing a weight to the $R_i$ parameter of each word proportional to its frequency in text, when finding the average $R$ value, is another possible improvement of the method. Still, it has not been able to provide radically different results.

## 3. Results and Discussion

### 3.1. Monkey texts

The examples of $v(t)$ dependences for the generalized Miller's monkey texts are illustrated in Figure 2a–e. The $v(t)$ functions for larger $M$'s reveal more and more intense oscillations at the largest times, so that it becomes somewhat problematic to find the exact saturated value $v_0$, which has to be treated rather as a mean value (see also [24, 25]). These oscillations are the best seen when the abscissa scale is logarithmic. Note also that increase in the gradient parameter $b$ decreases the oscillation amplitude (not shown in Figure 2). This finding can help in developing consistent theoretical models that explain the reasons for a stable (or oscillatory-like) asymptotic behavior of the $v(t)$ function.

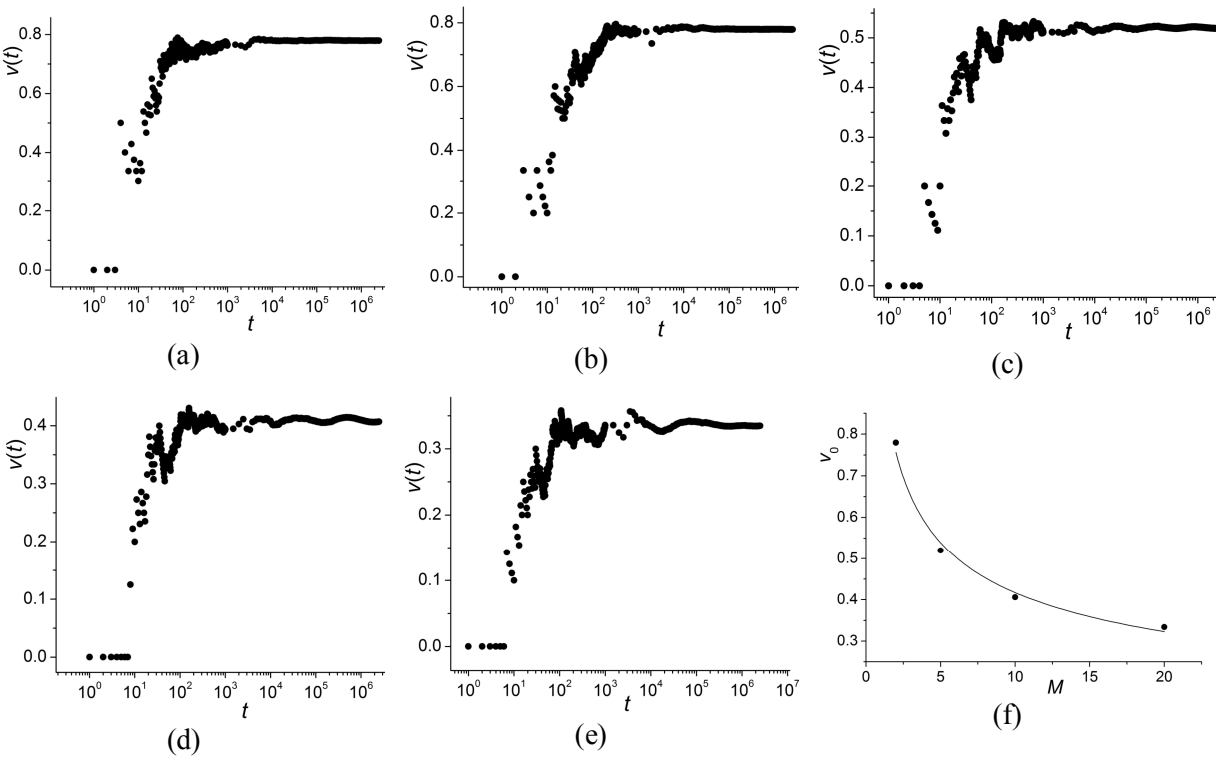


**Figure 2**: Dependences of repetition characteristics $v(t)$ for generalized monkey texts: (a) $M = 2$ and $b = 1$, (b) $M = 2$ and $b = 300$, (c) $M = 5$ and $b = 10$, (d) $M = 10$ and $b = 50$, and (e) $M = 20$ and $b = 100$. Panel (f) shows a dependence of saturated $v_0$ value on the alphabet size $M$: circles correspond to empirical data and line to power-law fitting

The second important fact seen from Figure 2 is a decrease in the (approximate) $v_0$ value occurring with increasing alphabet size $M$. This fact is readily understood from the combinatoric argumentation, because combinations of larger number of elements would imply a less number of random repetitions. It is interesting that even such a drastic change in the ratio $f_{max}/f_{min}$ as 300:1 does not affect the $v_0$ value (cf. Figure 2a and Figure 2b). Indeed, all the differences among the saturated $v_0$ parameters calculated for different gradients $b$ at the same alphabet sizes $M$ are typically less than 0.001. This testifies that the $b$ parameter is of no importance in what the $v_0$ value is concerned. This is a striking fact since, e.g., the monkey text at $M = 2$ and $b = 300$ is very 'close' to the text at $M = 1$ which reveals essentially larger $v_0$ than 0.78 (not shown in Figure 2). However, we still observe the same saturated repetition parameter as for the case $b = 1$.

As a consequence, our calculations have also demonstrated that the $v_0$ value does not depend at all on the single-character entropy $H$ of the random texts. In particular, the $H$ parameter calculated according to formula (2) changes by almost 70% (from 0.47 to 0.96) in the case of monkey texts with $M = 2$ and different $b$'s, although the repetition parameter remains essentially the same. Eventually, this fact is also true for the natural texts analyzed in Subsections 3.2 and 3.3. Moreover, it agrees indirectly with the

other known fact that randomization of natural texts changes the $v_0$ parameter substantially [24, 26], although the initial and randomized texts have the same single-character entropy $H$.

A lack of relations between $v_0$ and $H$ hardly conforms to assumption that the $v_0$ parameter can be linked with the redundancy of language associated with the number of 'units' of the writing system. It would be natural to suppose that a similar situation occurs for the 'true' entropy defined through conditional probabilities (i.e., through n-grams), which governs the 'true' redundancy of language. This conclusion has far-reaching consequences. So, it is known that different concisenesses of languages can be consistently interpreted in terms of their different information entropies or, in a slightly simplified manner, in terms of different slopes of their rank-frequency dependences for the letters [33]. However, it is not the case for the repetition parameter: it would have been hasty to suggest that different languages differ by their $v_0$ values (see the data [24]) due to different Shannon entropies of the corresponding codings. Something subtler must be at work, which awaits its further investigations.

Finally, the dependence of $v_0$ value on the alphabet size $M$ is displayed in Figure 2f. Among the simplest functions, the best fit is provided by the inverse power law:

$$v_0 = AM^{-B}, \tag{3}$$

where $A = 0.977$ and $B = 0.370$. The quality of this fit is quite satisfactory: the Pearson coefficient of the linear fit performed in the log-log scale amounts to −0.995. Nonetheless, larger-scale studies on the subject are necessary.

## 3.2. Natural texts in English

Examination of different natural texts written in the same language aims at fixing a variable associated with the writing system and studying instead whether (and why) these texts manifest different repetition rates. Here all of the finer points linked with local $v(t)$ bumps for individual texts, especially in the region of initial $t$'s, are of no importance, hence the linear abscissa scale in Figure 3a−d.

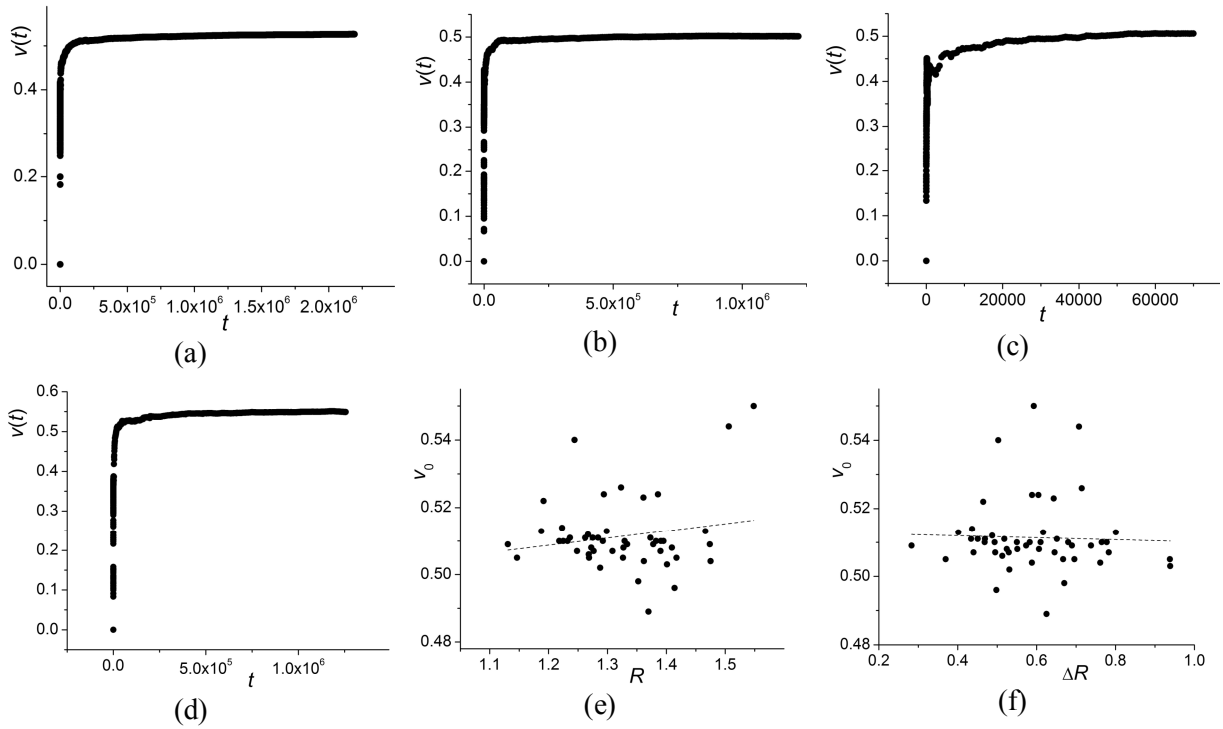


**Figure 3**: Examples of dependences of the repetition characteristic $v(t)$ for the natural texts in English: (a) 'Don Quixote of la Mancha' by M. Cervantes, (b) 'Moby Dick' by H. Melville, (c) 'The Jungle Book' by R. Kipling, and (d) 'The Origin of Species' by C. Darwin. Panels (e) and (f) show correlations of saturated $v_0$ values with the semantic load parameters $R$ and $\Delta R$ for 50 English texts

The analysis of empirical data, including the data presented in Figure 3, demonstrates that the $v(t)$ dependence begins to saturate roughly at $10^4$ characters. Introducing a conditional time $t_0$ after which the total $v(t)$ changes do not exceed 4%, one obtains $t_0 \approx (2-5)\cdot 10^4$ characters. However, the curve $v(t)$ 'converges' only after $t_0 \sim 10^5$ for a couple of texts. The $v_0$ data for different English texts are scattered in the region from 0.49 to 0.55, with the relative changes amounting approximately to 12%, so that the notion of 'converging repetition parameter' for a given language, which is implicitly exploited in the works [24, 25], is rather conventional (see also the discussion [26]).

Since it is known that randomization (i.e., shuffling) of natural texts decreases the saturated repetition parameter $v_0$ [24, 26] and, at the same time, decreases the $R$ and $\Delta R$ parameters [35], it seems promising to study a possible link of $v_0$ with $R$ and $\Delta R$. However, the results gathered in Figure 3e, f and Table 1 imply that these parameters are not correlated. It is even more so because $R$ and $\Delta R$ themselves are highly correlated ($r = 0.75$ for the English texts and $r = 0.83$ for the texts in different languages – see Subsection 3.3), although the correlation coefficients for the dependences $v_0(R)$ and $v_0(\Delta R)$ have the opposite signs for the English texts (see Table 1).

Trying to be cautious in discussing a possible correlation between $v_0$ and $R$ (or $\Delta R$), one can suppose that the two latter parameters are influenced by some additional factors, which do not affect $v_0$. In particular, it is our experience that $R$ and $\Delta R$ can differ for the texts of different sizes. However, the effect is the most pronounced for small text lengths only ($t_{\max} \sim 10^3 - 10^4$ characters), which is not our case. The Pearson coefficients for the dependences $R(t_{\max})$ and $\Delta R(t_{\max})$ are relatively low for our texts (0.24 and 0.21, respectively). Therefore, we are not in a position to find some factors which can overshadow a possible relationship of the semantic load and the saturated repetition parameter. We therefore are forced to admit that this link hardly exists. This also agrees with the fact that, in spite of notably different $v_0$'s, the monkey texts have almost the same $R$ parameter, $R \approx 1$.

**Table 1**

Pearson coefficients for the correlations between saturated $v_0$ value and parameters of semantic load R and ΔR for the natural texts under study

| Texts under study | Dependence | Pearson coefficient *r* |
|---|---|---|
| 50 English texts | $v_0$ vs. $R$ | 0.181 |
| 50 English texts | $v_0$ vs. $\Delta R$ | −0.04 |
| 34 texts in different languages | $v_0$ vs. $R$ | 0.132 |
| 34 texts in different languages | $v_0$ vs. $\Delta R$ | 0.185 |

## 3.3. Natural texts in different languages

To study inter-language differences in the repetition characteristic, we have examined single natural texts in 34 languages. Examples for two languages are depicted in Figure 4a, b. Note that the saturated values $v_0$ found for Ukrainian and its predecessor language of Old Rus' are only 1% different.

The main features of the $v(t)$ functions for the languages under test are similar to those found for the English texts. The only notable difference is Chinese and Japanese, for which the $v_0$ values are significantly less than those for the rest of languages. These results agree perfectly with the data derived in the work [25].

Let us put aside the data for Chinese and Japanese, which correspond to a completely different script. Then one can state that the interval of changes $\Delta v_0$ is about 25% for the languages that share what can be termed as, more or less, a common script. Moreover, the smallest value $v_0 \approx 0.44$ obtained for the Arabic text is not reliable enough due to its insufficient size ($t_{\max} \sim 1.5\cdot 10^4$ characters). Dropping this data point, we obtain a notably less value, $\Delta v_0 \approx 15\%$. It is not too larger than that found for different texts in English (see Subsection 3.2). In other terms, the intra-language $v_0$ variation makes up a significant fraction of the corresponding inter-language variation. This fact must be properly taken into account when building a consistent model of the repetition characteristic.

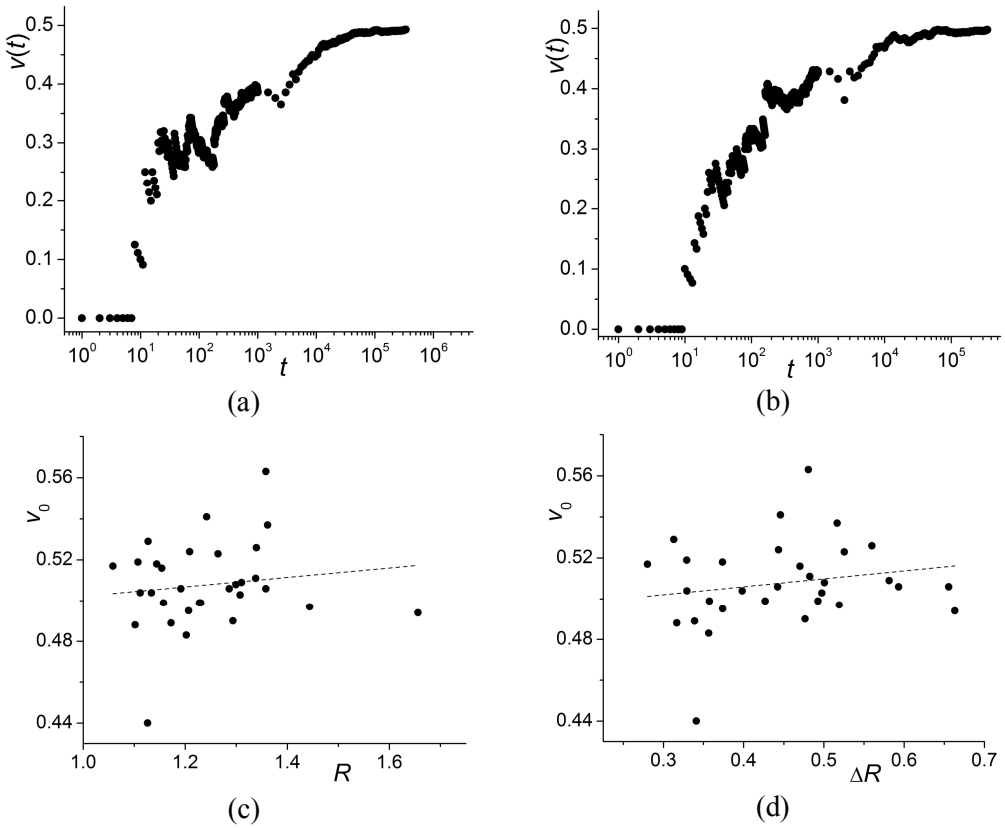


**Figure 4**: Examples of dependences of the repetition characteristic $v(t)$ for the natural texts in different languages: (a) 'Zakhar Berkut' by I. Franko (in Ukrainian) and (b) 'The Tale of Bygone Years' (in the language of Old Rus'). Panels (c) and (d) show correlations of saturated $v_0$ values with the semantic-load parameters $R$ and $\Delta R$ for the texts written in 34 languages

Similar to the data for the English texts, the idea of explaining variations of the $v_0$ parameter by varying semantic load of different texts seems to be not fruitful. As seen from Table 1, the Pearson coefficients of the corresponding correlations (Figure 4c, d) are low enough. Summing up the results of Subsections 3.2 and 3.3, one cannot state that differences in the semantic load can trigger in some way the appropriate differences in the repetition parameter $v_0$.

## 3.4. Different $v(t)$ regimes

To analyze the main features of different calculation regimes for the repetition characteristic, we have chosen 10 natural texts in English. Figure 5 displays exemplifying plots for one of these texts. The $v(t)$ function seems to be bounded in the repetition-counting regime "n-gram type" (Figure 5a, c, e, g). It is clearly unbounded in the alternative "n-gram token" regime (Figure 5b, d, f, h). These conclusions are valid irrespective of the mode accepted for scoring each repetition (1 point or n points per repeated n-gram). It looks like the combined regime "n points and n-gram types" (Figure 5c, g) is somewhat trivial. Namely, the $v(t)$ function tends to the unit limit, i.e. the total number of repetitions approaches the current time: $V = t$. The only difference of the regimes "characters" and "words" is a slower rate of this process in the latter regime.

One can see that a change in the linguistic unit (character or word) yields in quantitative rather than qualitative differences. Passing to "words" from "characters" affects both the saturated $v_0$ value and the corresponding characteristic time $t_0$ in the calculation regime "1 point and n-gram type". For all of the texts, we have $v_0 \sim 0.20{-}0.25$ and $t_0 > 10^5$ words. As a result, the saturation region is reached only for half of the texts, while the rest of them turn out to be too short to see this region (including the text 'Don Quixote of la Mancha' by M. Cervantes in Figure 5e). In general, a less $v_0$

value for the regime "words" is quite natural, since the repetition rate on this linguistic level is lower. In this respect, the situation with 'delayed' $v(t)$ saturation is less obvious.

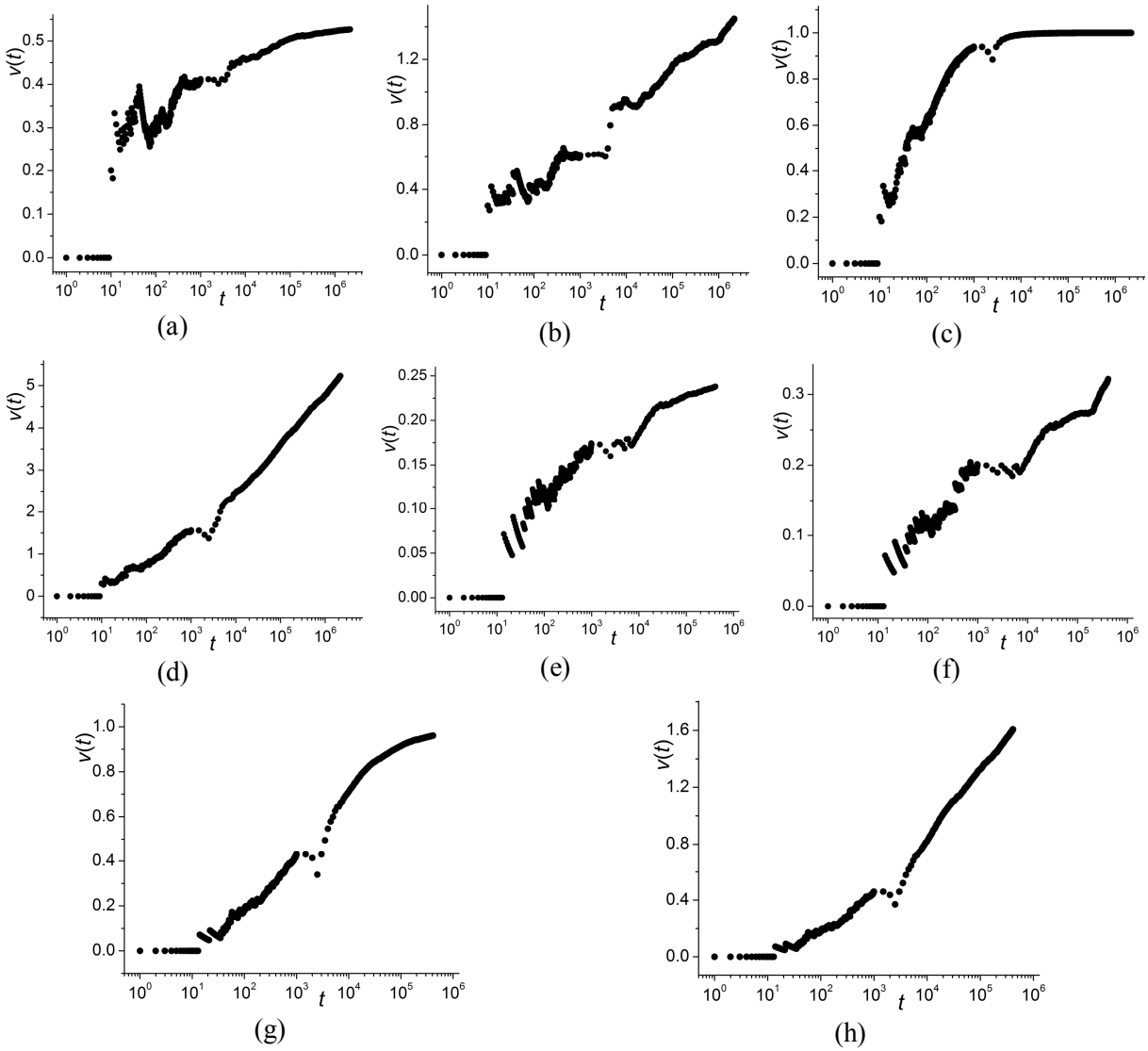


**Figure 5**: Dependences of repetition characteristic $v(t)$ for the text 'Don Quixote of la Mancha' by M. Cervantes calculated in different regimes: (a) characters, 1 point, n-gram types, (b) characters, 1 point, n-gram tokens, (c) characters, n points, n-gram types, (d) characters, n points, n-gram tokens, (e) words, 1 point, n-gram types, (f) words, 1 point, n-gram tokens, (g) words, n points, n-gram types, and (h) words, n points, n-gram tokens

## 3.5. Probing 'self-plagiarism' in texts

Of course, repetitions are a necessary aspect of a text. However, it can happen that a text manifests 'too many' repetitions, e.g. when the author deliberately repeats some textual fragments, especially long enough ones. Although this cannot be necessarily qualified as a negative (in some sense) fact, we would term, rather conveniently, this situation as a 'self-plagiarism' phenomenon. This ambiguous expression is used only because we found it difficult to pick up a more specific and relevant term.

Irrespective of the exact terminology, it seems important to develop a method for detecting this phenomenon. For instance, 'self-plagiarism' can be caught when checking dynamics of word vocabulary growth, since the vocabulary does not increase inside the region where a repeated textual fragment is available. However, the accuracy of this approach becomes sufficient only when the

above fragment constitutes a considerable fraction of the text itself. On the other hand, one can hope that calculating the repetition characteristics $v(t)$ in different regimes can offer a ready solution.

To examine the potential of the $v(t)$ function, we have taken the text 'The Jungle Book' by R. Kipling. Then two textual fragments with the lengths 695 words (or 3758 characters, including spaces) and 70 words (or 367 characters) have been copied from this text. These fragments amount respectively 5% and 0.5% of the total text size ($t_{max} \approx 7 \cdot 10^4$ characters). We have inserted them into the text at the position $t_{sp}$ = 7003 words (or $t_{sp}$ = 35750 characters – see Figure 6).

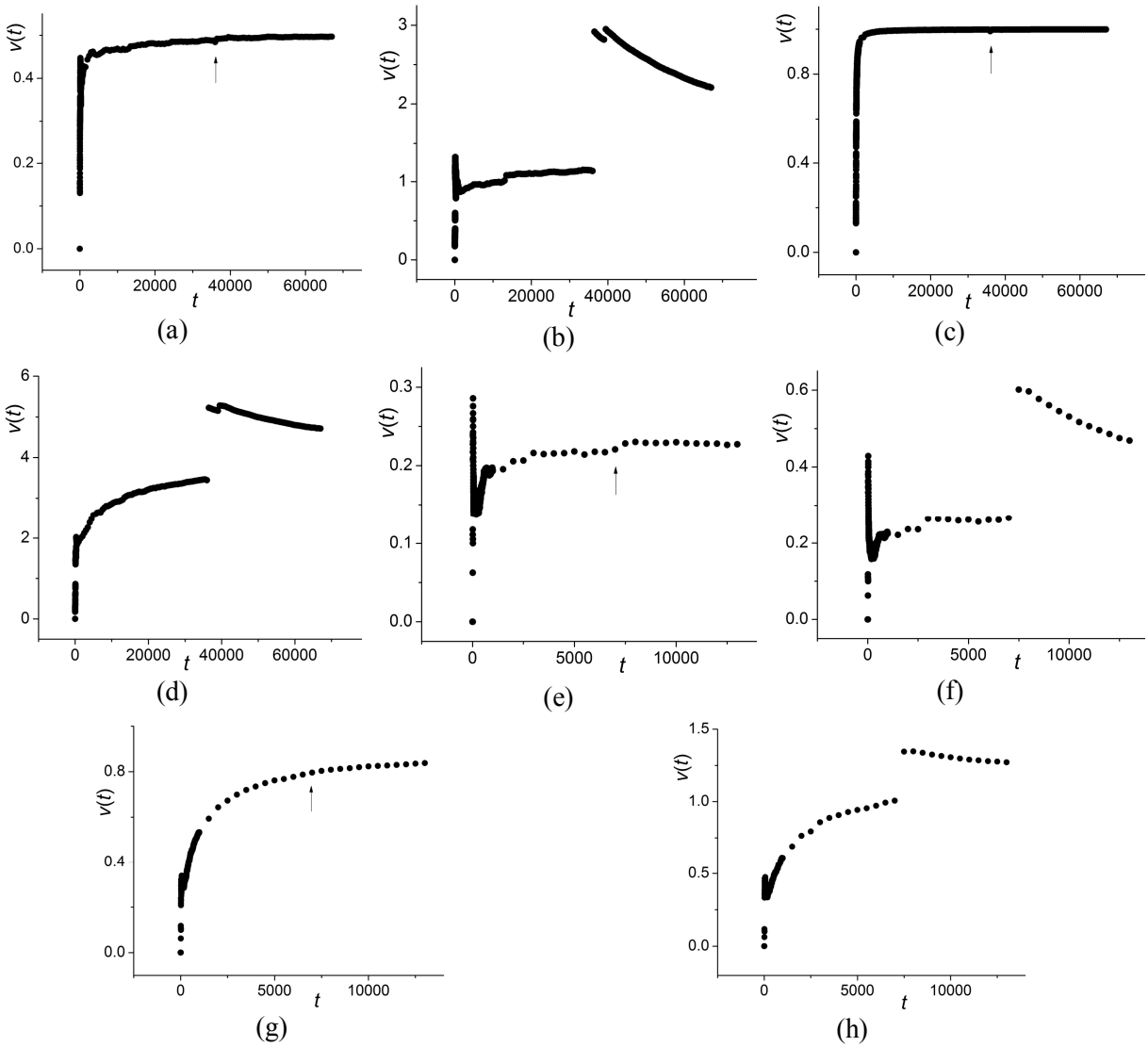


**Figure 6**: Dependences of repetition characteristic $v(t)$ for the text 'The Jungle Book' by R. Kipling with 'self-plagiarism' fragment (0.5% from the total text length), as calculated in different regimes: (a) characters, 1 point, n-gram types, (b) characters, 1 point, n-gram tokens, (c) characters, n points, n-gram types, (d) characters, n points, n-gram tokens, (e) words, 1 point, n-gram types, (f) words, 1 point, n-gram tokens, (g) words, n points, n-gram types, and (h) words, n points, n-gram tokens. When it is not evident, arrows indicate the position $t_{sp}$ where the fragment is inserted into the text

Figure 6 shows the dependences of repetition characteristics $v(t)$ calculated in different regimes in the case of second (shorter) 'self-plagiary' fragment. First of all, we stress that detection of this fragment relies upon local $v(t)$ behavior rather than its global trend or the $v_0$ value. Then, it is evident that the regime "n-gram tokens" can hardly detect the phenomenon (Figure 6a, c, e, g). This does not depends on the other modes used ("characters" or "words" and "1 point" or "n points"). The above

conclusion would be particularly relevant if the fragment were located somewhere in the beginning of text, where relatively large local $v(t)$ irregularities dominate.

**Table 2**
Sensitivity of $v(t)$ characteristic to 'self-plagiarism' in different calculation regimes

| $v(t)$ regime | Relative jump $\delta v$, % |
|---|---|
| Characters, 1 point, n-gram types | 1.1 |
| Characters, 1 point, n-gram tokens | 91.0 |
| Characters, n points, n-gram types | 0.6 |
| Characters, n point, n-gram tokens | 42.2 |
| Words, 1 point, n-gram types | 3.0 |
| Words, 1 point, n-gram tokens | 78.5 |
| Words s, n points, n-gram types | 1.0 |
| Words, n point, n-gram tokens | 28.8 |

On the contrary, the calculation regime "n-gram tokens" provides a very high sensitivity to the 'self-plagiarism'. This is testified by Figure 6b, d, f, h and Table 2, where the relative jumps $\delta v$ in the $v$ parameter occurring at the insertion position $t_{sp}$ are gathered. The repeated fragment can be easily detected by any of the alternative regimes "characters" or "words" and "1 point" or "n points". The mode combining the options "1 point" and "n-gram tokens" reveals better resources that the mode "n points" and "n-gram tokens". This finding is surprising enough, since one might hope that scoring more points for every repetition must have 'amplified' the effect of numerous repetitions and provided their better account. Another nontrivial result is that the sensitivity of the calculation regime "characters" is slightly higher that that of the regime "words". This refers to the both alternative combined modes "1 point and n-gram tokens" and "n points and n-gram tokens" (see Table 2).

In general, one can state that the sensitivity of the best versions of our detection technique is huge. As a further example, we ascertain that the maximal absolute jumps $\Delta v$ of the repetition parameter in case of the larger-scale (5%) 'self-plagiarism' are equal to 185 (in the regime "characters") and 32 (in the regime "words"). In other words, all of the other details in the $v(t)$ plot are simply lost, except for the 'self-plagiarism'. As a matter of fact, the sensitivity of this method is high enough to enable detecting reliably so small-scale repetitions which can hardly be qualified as manifestations of a 'self-plagiarism'.

## 4. Concluding Remarks

Let us sum up the main results of the present work. We have studied the statistical characteristic $v(t)$ of textual repetitions known from the earlier literature, focusing on natural and artificial single texts rather than large corpora. The reason is that we treat the latter linguistic objects as 'inhomogeneous' mixtures of single texts, with pronounced boundary effects present where the texts are joined together. We deem in this relation that the $v(t)$ characteristic, which still lacks a solid theoretical background, should be examined for as simple objects as possible.

It has been demonstrated that the saturated $v(t)$ value for a symbolic sequence, $v_0$, achieved at moderately large times $t$ is not correlated with its single-character entropy. This casts doubt upon a possible link of $v_0$ with the 'true' information entropy and redundancy of text as an information message. Similarly, the $v_0$ parameter is not linked with the total semantic load of text, which is quantized conventionally by the clusterization parameter averaged over all the words in a given text. Finally, comparison of $v(t)$ curves obtained for different languages shows that intra-language $v_0$ variations comprise a notable fraction of the corresponding inter-language variations. All the above facts have to be taken into consideration when developing a proper model of the repetition characteristic.

A number of modified regimes have been advised for calculating the repetition characteristic, which result in essentially different $v(t)$ functions. One of these modes, a regime of scoring each repetition instead of only the first one, proves to be ideal for detecting repeated fragments in texts. It

manifests a huge sensitivity, which is enough to detect not only the 'excessive' repetitions that include relatively long duplicated fragments ('self-plagiarism') but the repetitions of much shorter lengths.

Emphasizing the important remaining problems on the subject, we state first of all that the repetition characteristic needs a solid mathematical ground, at least for the simplest case of null hypotheses or stochastic language models, e.g. the random monkey texts. One can hope that these models can be examined in the frame of probability theory and yield analytical results. For proper comparison of these results with empirical data, it would be necessary to smooth numerous irregularities present in the region of initial times for any real texts. This can be done while applying a standard algorithm of sliding window to $v(t)$ calculations. In particular, the studies mentioned above would aid in understanding the nature of converging property of the $v(t)$ function at relatively small alphabet sizes $M$ – or lack of this property at larger $M$. It is interesting in this respect that the natural languages with clearly pronounced $v(t)$ converging property correspond just to the latter case of larger $M$'s. The other effect in need of study is a mysterious randomization-induced generation of $v(t)$ oscillations, which imply transition to a non-stationary process.

## 5. References


[1] L. Lü, Z.-K. Zhang, T. Zhou, Deviation of Zipf's and Heaps' laws in human languages with limited dictionary sizes, Sci. Rep. 3 (2012) 1082 (7 p.). doi:10.1038/srep01082.

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

[3] C. Bentz, R. Ferrer-i-Cancho, Zipf's law of abbreviation as a language universal, in: C. Bentz, G. Jäger, I. Yanovich (Eds.), Proc. Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics, University of Tübingen, 2016, online publication system. URL: https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558. doi:10.15496/publikation-10057.

[4] I. Moreno-Sánchez, F. Font-Clos, Á. Corral, Large-scale analysis of Zipf's law in English texts, PLoS ONE 11 (2016) e0147073 (19 p.). doi:10.1371/journal.pone.0147073.

[5] G. Cocho, R. F. Rodríguez, S. Sánchez, J. Flores, C. Pineda, C. Gershenson, Rank-frequency distribution of natural languages: a difference of probabilities approach, Physica A 532 (2019) 121795 (8 p.). doi:10.1016/j.physa.2019.121795.

[6] R. Ferrer-i-Cancho, C. Bentz, C. Seguin, Optimal coding and the origins of Zipfian laws, J. Quant. Lingvistics (2019) 31 pp. doi:10.1080/09296174.2020.1778387.

[7] M. Gerlach, E. G. Altmann, Testing statistical laws in complex systems, Phys. Rev. Lett. 122 (2019) 168301 (5 p.). doi:10.1103/PhysRevLett.122.168301.

[8] C. Casalnuovo, K. Sagae, P. Devanbu, Studying the difference between natural and programming language corpora, Empirical Software Engineering 24 (2019) 1823–1868 (46 p.). doi:10.1007/s10664-018-9669-7.

[9] K.-I. Goh, A.-L. Barabáshi, Burstiness and memory in complex systems, Europhys. Lett. 81 (2008) 48002 (5 p.). doi:10.1209/0295-5075/81/48002.

[10] E. G. Altmann, J. B. Pierrehumbert, A. E. Motter, Beyond word frequency: bursts, lulls, and scaling in the temporal distributions of words, PLoS ONE 4 (2009) e7678 (7 p.). doi:10.1371/journal.pone.0007678.

[11] A. Schenkel, J. Zhang, Y.-C. Zhang, Long range correlations in human writings, Fractals 1 (1993) 47–57. doi:10.1142/S0218348X93000083.

[12] E. G. Altmann, G. Cristadoro, M. D. Esposti, On the origin of long-range correlations in texts, Proc. Natl. Acad. Sci. (USA) 109 (2012) 11582–11587. doi:10.1073/pnas.1117723109.

[13] M. Gerlach, E. G. Altmann, Scaling laws and fluctuations in the statistics of word frequencies, New J. Phys. 16 (2014) 113010 (19 p.). doi:10.1088/1367-2630/16/11/113010.

[14] R. Ferrer i Cancho, R. V. Sole, The small world of human language, Proc. Roy. Soc. Lond. B 268 (2001) 2261–2265. doi:10.1098/rspb.2001.1800.

[15] S. Brin, L. Page, The anatomy of a large-scale hyper-textual Web search engine, Computer Networks and ISDN Systems, 30 (1998) 1–7. doi:10.1016/S0169-7552(98)00110-X.

[16] M. A. Montemurro, Entropic analysis of the role of words in literary texts, Adv. Complex Syst. 5 (2002) 7–17. doi:10.1142/S0219525902000493.

[17] R. Mihalcea, P. Tarau, TextRank: bringing order into texts, in: Proc. 2004 Conf. on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Barcelona, 2004, pp. 404–411. URL: https://www.aclweb.org/anthology/W04-3252.
[18] G. K. Palshikar, Keyword extraction from a single document using centrality measures, in: A. Ghosh, R. K. De, S. K. Pal (Eds.), Pattern Recognition and Machine Intelligence. PReMI 2007. Lecture Notes in Computer Science, volume 4815, Heidelberg Springer-Verlag, Berlin Heidelberg, 2007, pp. 503–510. doi:10.1007/978-3-540-77046-6_62.
[19] C. Wartena, R. Brussee, W. Slakhorst, Keyword extraction using word co-occurrence, 2010 Workshops on Database and Expert Systems Applications, volume 1, 2010, pp. 54–58. URL: https://info.computer.org/csdl/proceedings-article/dexa/2010/05592000/12OmNyxFKfX. doi:10.1109/DEXA.2010.32.
[20] R. G. Rossi, R. M. Marcacini, S. O. Rezende, Analysis of domain independent statistical keyword extraction methods for incremental clustering, Learning and Nonlinear Models 12 (2014) 17–37. doi:10.21528/lmln-vol12-no1-art2.
[21] W. Li, Random texts exhibit Zipf's-law-like word frequency distribution, IEEE Trans. Inform. Theory 38 (1992) 1842–1845. doi:10.1109/18.165464.
[22] R. Ferrer-i-Cancho, B. Elvevåg, Random texts do not exhibit the real Zipf's law-like rank distribution, PLoS ONE 5 (2010) e9411 (10 p.). doi:10.1371/journal.pone.0009411.
[23] F. Tria, V. Loreto, V. D. P. Servedio, Zipf's, Heaps' and Taylor's laws are determined by the expansion into the adjacent possible, Entropy 20 (2018) 752 (19 p.). doi:10.3390/e20100752.
[24] F. Golcher, A stable statistical constant specific for human language texts, 2007, pp. 1–6. URL: https://www.academia.edu/5986557/A_Stable_Statistical_Constant_Specific_for_Human_Language_Texts.
[25] D. Kimura, K. Tanaka-Ishii, Study on constants of natural language texts, J. Language Processing 21 (2014) 877–895. doi:10.5715/jnlp.21.877.
[26] K. Tanaka-Ishii, S. Aihara, Computational constancy measures of texts – Yule's K and Renyi's entropy, Computational Linguistics 41 (2015) 481–502. doi:10.1162/COLI_a_00228.
[27] F. Golcher, A new text statistical measure and its application to stylometry, in: Proceedings of Corpus Linguistics, 2007, 26 p. URL: https://www.semanticscholar.org/paper/A-new-text-statistical-measure-and-its-application-Golcher/f47a60c4d2e5a700443694b2a48990c3f a038fbd
[28] Project Gutenberg. URL: https://www.gutenberg.org/.
[29] E. Ukkonen, On-line construction of suffix-trees, Algorithmica 14 (1995) 249–260. doi:10.1007/BF01206331.
[30] B. Kokoszka, Visualization of Ukkonen's algorithm, 2016. URL: http://brenden.github.io/ukkonen-animation/.
[31] C. D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, Institute of Technology, Massachusetts, 1999.
[32] C. E. Shannon, Prediction and entropy of printed English, Bell Syst. Techn. J. 30 (1951) 50–64. doi:10.1002/j.1538-7305.1951.tb01366.x.
[33] O. S. Kushnir, O. V. Dzera, L. O. Kushnir, Conciseness of Ukrainian, Russian and English: application to Translation Studies, in: Proc. XI Int. Sci. and Pract. Conf. on Electron. and Inform. Technol. (ELIT-2019), Lviv, Ukraine, 2019, pp. 44–50. doi:10.1109/ELIT.2019.8892290.
[34] D. R. Amancio, E. G. Altmann, D. Rybski, O. N. Oliveira Jr., L. da F. Costa, Probing the statistical properties of unknown texts: application to the Voynich manuscript, PLoS ONE 8 (2013) e67310 (10 p.). doi:10.1371/journal.pone.0067310.
[35] O. S. Kushnir, A. I. Kashuba, V. V. Yaremkiv, Distinguishing between natural and random texts: a statistical measure linked to word clustering, in: Proc. 2nd Int. Conf. on Computational Linguistics and Intelligent Systems, volume II: Workshop, Lviv, Ukraine, 2018, pp. 112–113. URL: http://ena.lp.edu.ua:8080/handle/ntb/42556.
[36] P. Carpena, P. A. Bernaola-Galván, C. Carretero-Campos, A. V. Coronado, Probability distribution of intersymbol distances in random symbolic sequences: applications to improving detection of keywords in texts and of amino acid clustering in proteins, Phys. Rev. E 94 (2016) 052302 (13 p.). doi:10.1103/PhysRevE.94.052302.