Electronics and Information Technologies, 2026, 35, 31–44
http://publications.lnu.edu.ua/collections/index.php/electronics/index

UDC: 004.6, 004.89, 004.932, 81`32

COMPRESSION COEFFICIENT IN THE ANALYSIS OF NOISY TEXTS AND IMAGES

Oleksii Kravchuk* & ***Oleh Kushnir***
Department of Optoelectronics and Information Technologies,
Ivan Franko National University of Lviv,
107 Gen. Tarnavskoho Street, Lviv, 79017, UKRAINE

Kravchuk, O. M., & Kushnir, O. S. (2026). Compression Coefficient in the Analysis of Noisy Texts and Images. *Electronics and Information Technologies*, 35, 31–44. https://doi.org/10.30970/eli.35.3

ABSTRACT

Background. We analyze the compression coefficient *S* associated with the two versions of a Lempel–Ziv algorithm (LZA) as a quantitative characteristic of structural order in noisy or corrupted data, using a natural-language text and a digital image. To model a stochastically perturbed text or image, we randomize the linguistic elements of the former and add uncorrelated Gaussian noise to the latter or randomize its pixels.

Materials and Methods. A fiction text "Moby Dick" by H. Melville was randomized at the linguistic levels of sentences, words, and characters. The number of randomization cycles *N* varied in the range from 0 to 10^9. We also analyzed the red channel of an 8-bit image "Peppers" with the size $M = 512\times512$. The Gaussian noise with the standard deviations σ changing in the range from 0 to 400 was added to the image. On the whole, 50 randomized and 50 noisy image versions were generated. To apply the LZA compression to the image, we unfolded it into a one-dimensional sequence of pixel intensities and processed the appropriate textual representation of the image as usual text.

Results and Discussion. The sentence-level randomization almost does not change the *S* value, because it remains intact the internal structures of words and sentences, while the LZA has a local character. The word-level randomization decreases *S*, whereas the character-level randomization leads to the largest *S* decrease and produces almost incompressible text files at the largest *N*. The results obtained with the two LZA-type versions differ slightly in the effect of sentence randomization on the compressibility. For the randomized images, the *S* value sharply decreases when *N* approaches the total pixel number *M*, and then reaches saturation. The strongest changes in *S* for the noisy images occur in the range of $\sigma \in [10, 100]$.

Conclusion. Our results demonstrate that the LZA compression coefficient is sensitive to the loss of local order in the noisy data of different types, either randomized texts or randomized and noisy images. Since the LZA is based on repetitions and local regularities in a symbolic sequence, a decrease in *S* indicates partial destruction of order caused by randomization or noise. Hence, the compression coefficient can, in principle, serve as a measure of disorder in both symbolic sequences and pixel grids and so can help in detecting noisy data.

Keywords: data compression, noisy data analysis, noise detection, randomization, natural language processing, image processing.

1. INTRODUCTION

There are many practical situations when the data of interested for information retrieval is noisy or partly corrupted. Among diverse examples, one can recall, in this

© 2026 Oleksii Kravchuk & Oleh Kushnir. Published by the Ivan Franko National University of Lviv on behalf of Електроніка та інформаційні технології / Electronics and Information Technologies. This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 License which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

respect, mangled or damaged textual documents, visual or compression artifacts of digital images, grainy images, corrupted footage, or damaged frames. As a limiting case, the data can become so corrupted that its information content cannot be retrieved against the background of dominating noise. Then the problem is to recognize this noisy data from the neat information-carrying one, using no human intervention or visual inspection.

The main point of interest in the present work is the possibilities for detection of 'noise' in textual documents. This can be done, e.g., using different complexity measures and some other approaches [1–6]. Technically, a text represents a sequence – or time series – of linguistic symbols. Our auxiliary point is therefore to find such statistical properties or characteristics that can help in differentiating between the informative and (partly) noisy symbolic time series. One of the possible characteristics is the compressibility of a text file [7–11]. Here the motivation is simple: ordered and information-bearing sequences are usually more compressible than noisy or randomized ones [1, 3, 12–14]. Indeed, any lossless compression algorithm finds repeated fragments and regularities in the input data. If the noise destroys these regularities, the size of the compressed file increases and the compression coefficient decreases.

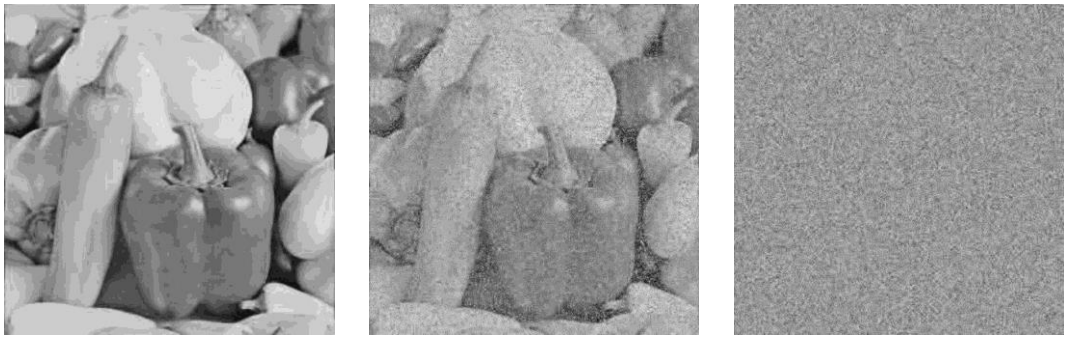
In other words, it is important that the compression coefficient reacts not only to the frequencies of linguistic elements given by Zipf's or Heaps' laws (see Ref. [15]), but also to their order. The other point of our interest is automatic detection of noisy images [16–20]. It is important that the compression-based approach can, in principle, be applied to images, since the latter can be viewed as grids of pixel values [20, 21].

The aim of this study is to analyze the compression coefficient for disordered or corrupted natural-language texts and noisy images. As a model of stochastically perturbed text, we choose the simplest mechanism of partial randomization (or shuffling) of its linguistic elements such as sentences, words, and characters. To model a partially corrupted image, we add uncorrelated Gaussian noise to it and, moreover, randomize its pixels. To analyze the changes occurring with a text or image under noising, we examine a gradual evolution of its compression coefficient during the noising process.

2. MATERIALS AND METHODS

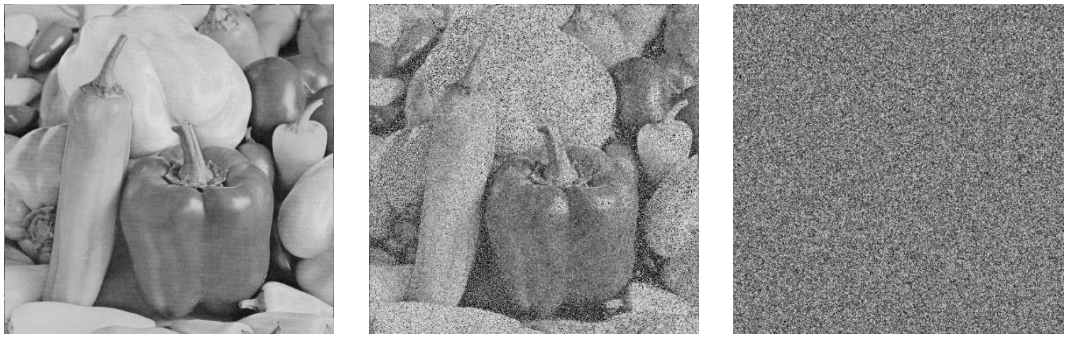
2.1. Data under study and noising procedures

The objects of our study are a natural-language text "Moby Dick" by H. Melville and a digital image "Peppers" which is often analyzed in the field of image processing (see **Fig. 1** and **Fig. 2**). Notice that the text under test includes approximately 1.24×10^6 characters, 2.12×10^5 words and 1.01×10^4 sentences. The 8-bit color image «Peppers» stored in the JPEG format has the size $M = 512\times512$ pixels.



(a) (b) (c)

Fig. 1. Examples of randomization results obtained for the red channel of the image "Peppers" at the randomization cycle numbers $N = 0$ (original image) (a), $N = 73114$ (b), and $N = 10^8$ (c)



(a) (b) (c)

Fig. 2. Examples of Gaussian noise results obtained for the red channel of the image “Peppers” at the noise parameters $\sigma = 0$ (original image) (a), $\sigma = 35.6$ (b), and $\sigma = 402.4$ (c)

As already mentioned above, we modeled corruption of texts in the simplest manner, as partial randomization of their linguistic elements. The starting point of our experimental procedures was the assumption that the compressibility of a text as a symbolic sequence should reflect the amount of order preserved in it. Randomization was treated as a multiply repeated pairwise exchange of randomly selected linguistic elements of a text. The degree of randomization was measured by the number N of elementary randomization cycles, where one cycle corresponded to a single mutual exchange of two elements. The number of cycles was varied in the range from $N = 0$ (the original text) up to $N = 109$. This was performed in 50 discrete steps. Note that the maximal cycle number $N_{\max}$ was far larger than the total number L of its linguistic elements (either characters or words, or sentences) in our text (see the figures mentioned above).

The text was randomized separately at the linguistic levels of sentences, words, and characters. As a result, three randomized datasets were obtained for the initial text. The type of element being exchanged depended on the selected linguistic level of randomization. During the sentence-, word-, and character-level randomizations, whole sentences, individual words, and separate characters were permuted, respectively.

In the same way, the pixels of our image were randomized such that, during each of the elementary randomization cycles, the pixel values were left unchanged, but the spatial locations of the pixels were swapped. The quantitative parameter governing image corruption was again the number N of randomization cycles, which was changed from 0 to 10^8. Here the relation $N_{\max} >> M$ held true. Besides, Gaussian noise with a zero mean was added to the image. It was dosed by the appropriate standard deviation σ, which was increased exponentially from 0 to 400 in order to cover the regions of both weak and strong image perturbations. Notice that the standard deviation characterizing the pixel-value variation in the unnoised image was equal to $\sigma_0 \approx 66.1$, so that the maximal σ applied was surely much larger than the σ_0 value. As a consequence, we obtained 50 noisy and 50 randomized image versions. Some examples of the randomized and noisy images generated as explained above are shown respectively in **Fig. 1** and **Fig. 2**.

2.2. Lempel–Ziv compressibility

The text file was compressed using a freely available Lempel–Ziv algorithm (LZA). The LZA family comprises lossless compression methods that exploit repetitions in the input data and require no predefined statistical model for the data [7]. Instead of assigning short codes only to frequent individual symbols, these methods identify recurring sequences in the data and replace later occurrences with the references that a decoder can interpret. This principle allows the compressor to adapt to the structure of the processed data and capture regularities at different scales, including recurring symbol

sequences or appropriate bit or byte patterns which correspond to characters, words, or still longer textual elements.

The present work employs two specific realizations of the LZA family. The first uses a dictionary-coding strategy intended primarily for symbolic strings, while the second operates with a byte-oriented format. Although both methods are lossless, they should not be treated as interchangeable implementations of a single algorithm, because they discover and encode redundancy using different internal structures and are designed for different operational purposes.

In the frame of the first compression method, the input data is viewed as a sequence of symbols. During a left-to-right scan, the encoder searches for the longest sequence already represented in its current dictionary. The code of that sequence is emitted, and a longer sequence formed by appending the next input symbol is inserted into the dictionary whenever it has not occurred before. On the contrary, the second method operates on byte sequences. Its smallest input unit is a complete byte rather than an individual bit or a linguistic token. At each position, the encoder searches for previously processed bytes for a sufficiently long sequence matching the bytes beginning at the current position. When a match is found, the repeated bytes are replaced by a back-reference that consists of an offset and a match length. The offset specifies how many bytes backward the previous occurrence begins, while the length specifies how many bytes the decoder must copy. Notice that still different LZA-type compression techniques are known from the literature, e.g. that operating on the bit level of symbolic sequences (see, e.g., Ref. [14]).

For every original and randomized text version, the compression coefficient S was calculated as the relative decrease in the file size occurring after compression:

$$S = 1 - s_{out}/s_{in},$$

where s_{in} and s_{out} are the sizes of the original and compressed text files, respectively. Notice that the results obtained with both algorithms depend on the repeatability of text fragments and local regularities in the input symbolic sequence. Then large S values correspond to well-compressible ordered sequences, whereas the S values close to zero imply that the compressed file is almost as large as the input one.

Notice that the LZA is directly fit for the one-dimensional symbolic sequences. Of course, there is a great number of compression techniques aimed at two-dimensional images [16]. However, in order to adhere to the same experimental compressibility measure in the case of two-dimensional digital images and compare the appropriate results with those obtained for the texts, one can conventionally reduce the images to linear chains of pixel values. Simple methods for this aim have been suggested in the earlier works [20, 21]. The main idea of those works was to transform a two-dimensional image into a one-dimensional sequence of pixel values, using a horizontal or vertical scan, and then analyze this sequence by the methods of computational linguistics applied to texts.

In practice, an image can be converted to a linear chain of pixel values by a row-by-row or column-by-column scanning. In this representation, the figures corresponding to pixel intensities play the role of ‘words’ in a text. Although failing to reflect all the two-dimensional neighborhoods in an input image correctly, the image-to-text transformation makes it possible to consider the image properties within the same framework based upon the analysis of one-dimensional symbolic sequences. To be specific, our image was unfolded into a text along the horizontal direction. For this purpose, we used the data of the red channel of the color image.

Finally, an additional check-up compression-related analysis was performed immediately with the JPEG files of the input image and its noised and randomized versions. Here we directly compared the sizes of the files, with no intermediate conversion of image files into text files and with no LZA analysis. The compression parameter K was defined as

$$K = s_0/s,$$

where s_0 and s denote the JPEG-file sizes of respectively the input and noised (perturbed) images. The K coefficient was calculated for both noisy and randomized image datasets. This analysis is not identical to the calculation of S, because a lossy JPEG format includes its own compression procedures. Nevertheless, it provides an additional check of how noising and randomization affect the practical compressibility of the image file.

3. RESULTS AND DISCUSSION

3.1. Randomized texts

Fig. 3 shows the dependence of the compression coefficient S on the randomization parameter N for the three linguistic levels of randomization. The panels (a) and (b) correspond respectively to the first and second LZA-type compression methods (see Subsection 2.1). Here the randomization versions that correspond to the sentence, word, and character linguistic levels are denoted by the curves of different colors (see the legend in **Fig. 3**).

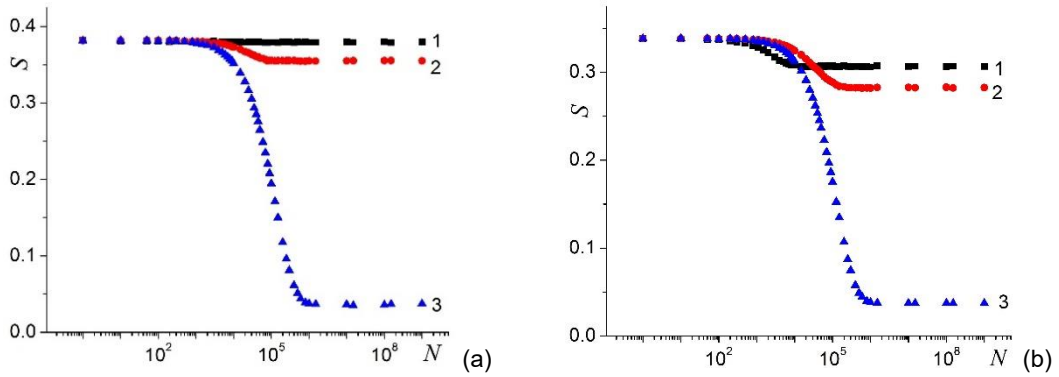


Fig. 3. Dependences of compression coefficient S on the randomization parameter N obtained for the text "Moby Dick". Panels (a) and (b) correspond to different versions of LZA-type libraries (respectively, character-level and byte-level compressions – see the text). Data points denoted by black squares (curves 1), red circles (curves 2), and blue triangles correspond, respectively, to randomization performed at the levels of sentences, words, and characters. For visualization on the logarithmic scale, the initial data points corresponding to $N = 0$ are conventionally represented by $N = 1$

The first important result is the almost complete absence of changes in the compressibility observed under the sentence-level randomization (see curve 1 in **Fig. 3** and data in **Table 1**). Note that the notation L in **Table 1** implies the text length, S_{in} is the compressibility S measured at $N = 0$, S_{avg} is the S value averaged over the last 10 data points for strongly randomized texts, and ΔS is defined as $\Delta S = S_{in} - S_{avg}$. To be exact, the changes mentioned above are negligible in the case of character-level compression and amount to only 10% of the changes observed under character-level randomization in the case of byte-level compression. This means that changing the order of sentences does not have a significant effect on the file compressibility. The reason is that the LZA algorithm applied by us has an essentially short-range character: its result depends on the order of linguistic elements (or their binary representations) located close to each other. In other words, the sentence-level permutations change the global structure of the text radically, but most of the local structures inside separate sentences remain invariable. Notice that the reasons for different results of the two alternative LZA realizations associated with the

Table 1. **Some data related to the compression coefficient obtained for the original and randomized texts, using the character-level LZA compression**

Quantitative characteristic	Sentence-level randomization	Word-level randomization	Character-level randomization
Text length L	10,076	212,288	1,239,464
Compressibility values S_{in} and S_{avg} before and after randomization	0.381 → 0.379	0.381 → 0.355	0.381 → 0.039
Compressibility change ΔS under randomization	0.002	0.026	0.342

Note: The data in the third row of the table imply $S_{in} \to S_{avg}$.

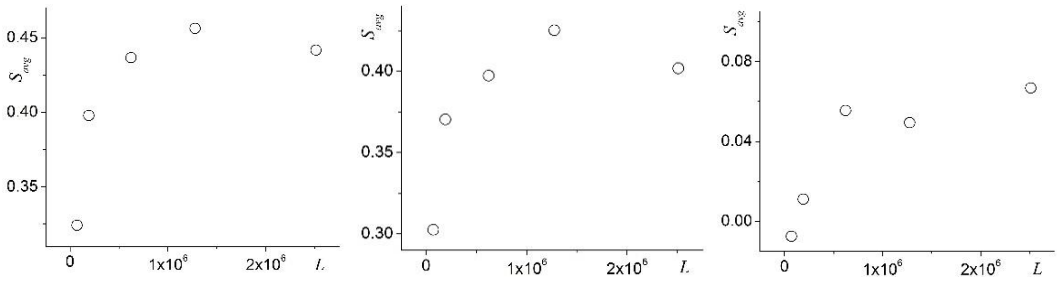
character and byte levels deserve further study. Since the second LZA realization manifested notably higher data point scattering in case of image-mapping compression, we will consider only the first LZA realization for images.

A notably different behavior is observed for the word-level and, especially, character-level randomization (see curves 2 and 3 in **Fig. 3a** and **Fig. 3b**). In both cases, the compression coefficient decreases notably with increasing N. This decrease reflects the gradual destruction of local regularities in the symbolic sequences. In particular, the word-level randomization destroys the order of words, though the internal structure of each word remains intact. Therefore, the limiting S value at the largest N remains relatively high. Finally, the character-level randomization produces the strongest effect because it destroys not only the word order but also the internal structure of words.

The other important fact is that an expressed drop in the S parameter is observed whenever the cycle number N becomes $N \approx L$ or somewhat larger, with L denoting the total number of sentences, words, or characters in the text. This can be seen both in **Fig. 3a** and, especially, in **Fig. 3b** where the effect of the sentence-level randomization is not negligible. Indeed, the S values referred to the sentence-, word-, and character-level randomization are stabilized after dropping at approximately $N \approx 10^4$, $2{\times}10^5$, and $1.2{\times}10^6$, respectively. A comparison with the L data indicated in Table 1 confirms that $N \approx L$ really represents the condition for stabilization of the S parameter. At $N \gg L$, the symbolic sequence is practically fully randomized, and any further increase in N produces no noticeable changes. In this limiting case, the S parameter approaches zero in the most radical regime of character-level randomization, so that we have $S_{in} \approx S_{avg}$ (see the S_{avg} value in the last column of **Table 1**), which means that the text file becomes almost incompressible.

Thus, the total difference ΔS between the S values found for the original text and its strongly randomized versions depends strongly on the randomization level. It is maximal for the character-level randomization, smaller for the word-level randomization, and almost absent for the sentence-level randomization. This confirms that the compression coefficient is much more sensitive to the local order than to the global one.

It would be convenient if the difference $\Delta S = S_{in} - S_{avg}$ between the compression coefficients of unnoised and strongly noised texts were a function of only randomization type. Then a large ΔS parameter would have served as a criterion of strong text-data corruption. However, this hypothesis needs examining if ΔS does not depend on the text size. To clarify whether the limiting compressibility S_{avg} of strongly randomized texts depends on the text size L, we have measured the $S(N)$ dependences for the four additional natural texts of English-language fiction with essentially different lengths L and then calculated the averaged parameters S_{avg}. The results are shown in **Fig. 4** for the three



(a) (b) (c)

Fig. 4. Dependence of the limiting compression coefficient value S_{avg} of strongly randomized texts, which is averaged over the last 10 points of the appropriate $S(N)$ curves, on the text size L expressed in terms of character number. Panels (a), (b) and (c) correspond respectively to the S data obtained at the sentence, word and character randomization levels. The data refers to the character-level version of LZA

linguistic levels of randomization. Note that the first data point $S_{avg} < 0$ in **Fig. 4c** obtained for the shortest text is merely an experimental artifact that reflects nothing but extremely small S values typical for the short random symbol sequences.

For all the linguistic levels of randomization, we observe an almost monotonous, increasing dependence $S_{avg}(L)$ that tends to become saturated at the largest L. In other words, the noised texts of larger sizes usually have larger compression coefficients. This indicates that the limiting compressibility of strongly randomized texts is determined, first of all, by the number of their linguistic elements. The appropriate reason is natural: the gain in the file size due to compression becomes larger for larger files. Besides, the data point scattering observed in **Fig. 4** testifies that some other properties of texts also influence the S_{avg} value. Since no internal local order or the corresponding regularities of repetitions are peculiar to the totally randomized texts, the properties mentioned above can be the alphabet and vocabulary sizes of the original text, as well as the frequency distributions of its linguistic elements (characters or words).

It is seen from **Fig. 4** that, unfortunately, the S_{avg} parameter is indeed affected by the text size, so that the absolute ΔS value cannot be directly used to judge if an unknown text is strongly noised. Nevertheless, one witnesses a gradual saturation of the $S_{avg}(L)$ functions with increasing L around some limiting values. The remaining shortcoming of this situation is that the corresponding lengths L are relatively large, thus preventing one from any categorical ΔS-based conclusions about the noising degree for short textual documents.

3.2. Randomized and noisy images

The dependences of the compression coefficient of the image “Peppers” on the randomization and noise parameters are shown in **Fig. 5**. For the randomized images, the compression coefficient sharply decreases when the number of randomization cycles approaches the value $M = 512\times512 \approx 2.6\times10^5$ (see **Fig. 5a**). Further increase in N has almost no influence on S. This behavior is analogous to the case of texts: the characteristic number of random permutations needed to destroy the order in the symbolic sequence is of the same order of magnitude as the number of elements in this sequence.

On the other hand, the dependence for the noisy images (see **Fig. 5b**) is controlled by the standard deviation σ of the Gaussian noise. The strongest S change occurs approximately in the interval $\sigma \in [10; 100]$. This interval is close to the standard deviation σ_0 of the pixel values in the original image ($\sigma_0 \approx 66$ – see Subsection 2.1). Therefore, the main loss of compressibility occurs whenever the external noise becomes comparable with the ‘internal’ pixel-level variations in the original image.

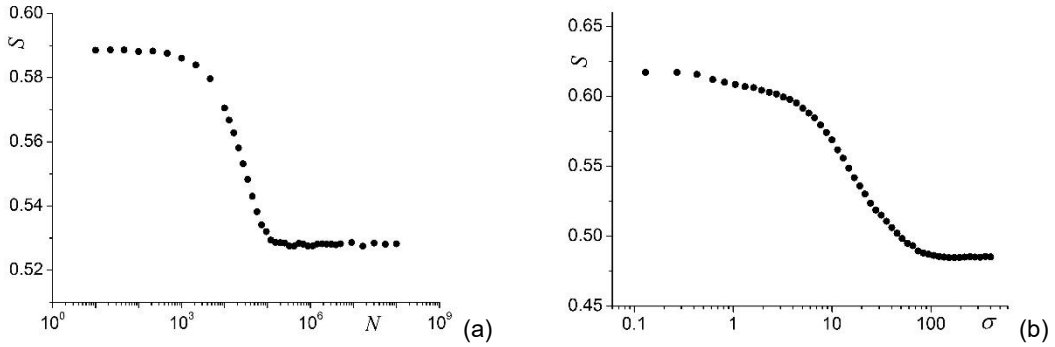


Fig. 5. Dependence of compression coefficient S on the randomization parameter N (a) and the Gaussian noise parameter σ (b), as obtained for the image "Peppers". The data refers to the character-level version of LZA. For visualization on the logarithmic scale, the initial data point in panel (a) corresponding to $N = 0$ is conventionally represented by $N = 1$

The minimal compression coefficients for our image are far from zero and notably higher than the S_{avg} values obtained for the word-randomized, not to say character-randomized texts. This is seen from **Table 2**. Here S_{in} denotes the compressibility S measured at $N = 0$ or $\sigma = 0$, S_{avg} the S value averaged over the last 10 data points, and ΔS is defined as $\Delta S = S_{in} - S_{avg}$. This is explained by a specific structure of textual representations of images. Even a strongly randomized or noised image preserves significant local regularities associated with the 'internal structure' of the numbers corresponding to pixel levels, the limited 'vocabulary' of 256 possible pixel values (0 to 255), frequent random repetitions of these values, and additional restrictions imposed by the intensity histograms.

Table 2. **Details of the compression coefficient obtained for the original, randomized, and noised images, using the character-level LZA compression**

Quantitative characteristic	Pixel randomization	Gaussian noising
Compressibility values S_{in} and S_{avg} before and after randomization or noising	0.589 → 0.528	0.589 → 0.485
Compressibility change ΔS under randomization or noising	0.061	0.104

Note: The data in the second row of the table imply $S_{in} \to S_{avg}$.

The image "Peppers" has been additionally analyzed without intermediate conversion into a one-dimensional text file. Then the characteristic parameter K is calculated simply from the sizes of JPEG files (see Subsection 2.2). The corresponding dependences are shown in **Fig. 6** (see also **Table 3**). Note that the notation K_{in} in **Table 3** refers to $N = 0$ or $\sigma = 0$, K_{avg} is the K value averaged over the last 10 data points, and ΔK is defined as $\Delta K = K_{in} - K_{avg}$. The dependence $K(N)$ in **Fig. 6a** for the randomized images looks as expected: as the randomization degree increases, the spatial structure of the image is destroyed and the JPEG representation becomes less efficient. Consequently, the file sizes of the randomized images increase and the corresponding size ratio K decreases. In general, the characters of the $K(N)$ and $K(\sigma)$ functions are similar: after a transition region, both functions approach the limiting values that correspond to strongly disordered images.

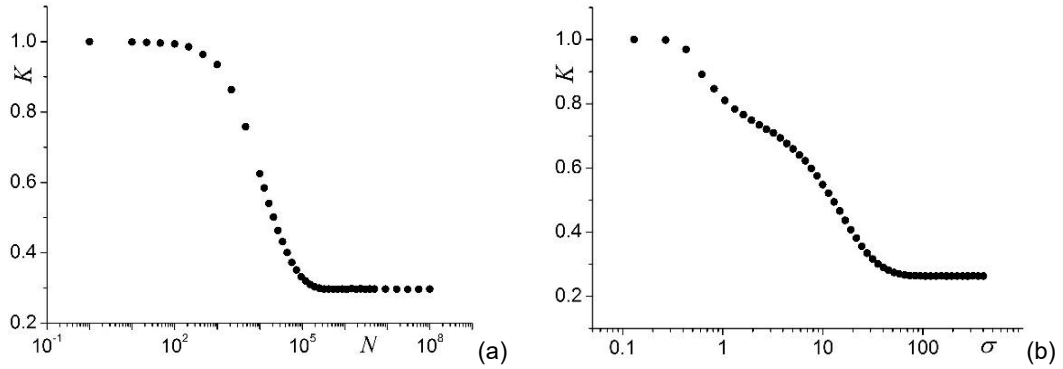


Fig. 6. Dependence of the compression parameter *K* associated with JPEG file sizes of the image "Peppers" on the randomization parameter N (a) and the Gaussian noise parameter σ (b). For visualization on the logarithmic scale, the initial data point in panel (a) corresponding to $N = 0$ is conventionally represented by $N = 1$

Table 3. **Details of the JPEG size-based compression parameter *K* obtained for the randomized and noised images**

Quantitative characteristic	Pixel randomization	Gaussian noising
Compressibility parameters K_{in} and K_{avg} before and after randomization	1.00 → 0.30	1.00 → 0.26
Change ΔK in the compressibility parameter K under randomization or noising	0.70	0.74

Note: The data in the second row of the table imply $K_{in} \rightarrow K_{avg}$.

3.3. General data interpretation

The results obtained by us for both a sample text and a sample image are consistent with each other. In particular, the compression coefficient decreases when the structural order in the symbolic time series is destroyed. The clearest analogy is observed between the word-level text randomization and the pixel-intensity randomization in an image. In both cases, the elementary 'semantics-carrying' elements (respectively words and pixel values) are permuted, and a sharp change appears when the number of permutations becomes comparable with the total number of the elements.

A weak influence of the sentence-level randomization on the text compressibility should not be interpreted as evidencing that the sentence order is unimportant. Rather, this demonstrates that the compression-based method is mainly sensitive to the local regularities. In this respect, our method is close to other short-range characteristics, like the network ones [21], and should be used for quantifying local order but not global semantic coherence.

In order to better understand the behavior of compressibility depending on N and σ, it would be convenient to study some alternative characteristics of the structural disorder in digital images, statistical or other measures of complexity, and compare them with those described above. For this aim, we have calculated the information Shannon entropy *SE* and the peak signal-to-noise ratio *PSNR* for the image "Peppers" (see **Fig. 7**) [16].

As already noticed above, the *SE* parameter $H = -\Sigma_k\, p_k \log_2 p_k$ calculated from the pixel-value probability distribution p_k (i.e., the image histogram) is an invariant of the pixel randomization procedure and reveals no changes (**Fig. 7a**). In other words, the LZA compression for images in this respect is superior if compared to the histogram-based

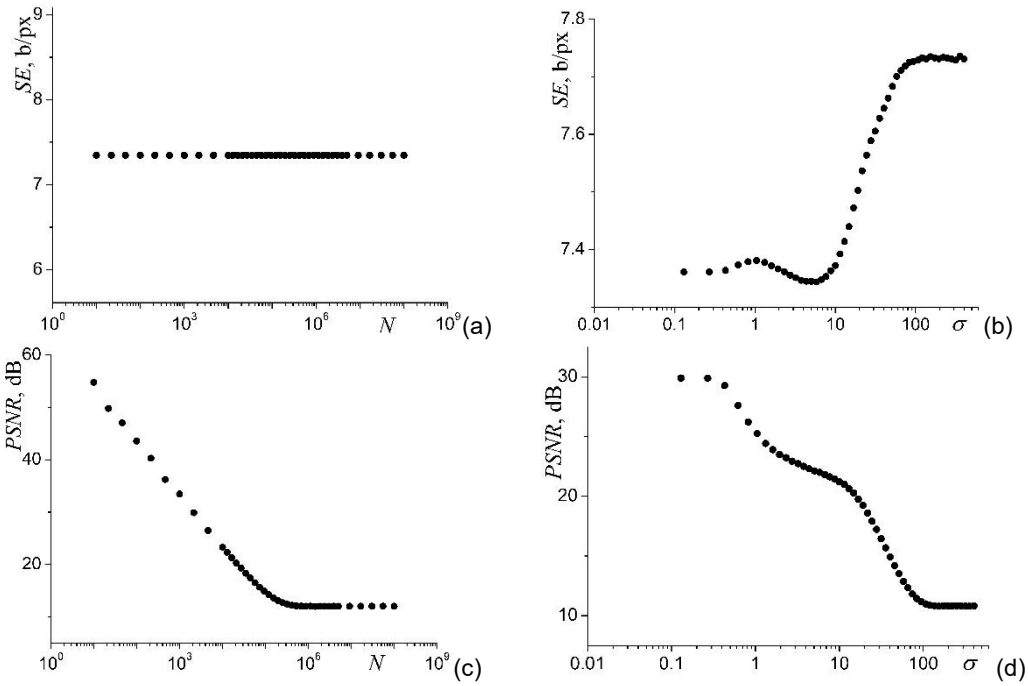


Fig. 7. Dependence of Shannon entropy *SE* (a, b) and *PSNR* parameter (c, d) of the image "Peppers" on the randomization parameter *N* (a, c) and the Gaussian noise parameter *σ* (b, d)

parameters, which are insensitive to the disorder induced by pixel shuffling. The dependence *SE*(σ) shown in **Fig. 7b** can easily be explained: the uncertainty of the p_k distribution increases with increasing noise parameter σ.

Finally, the dependences *PSNR*(N) and *PSNR*(σ) in **Fig. 7c, d** are much smoother in comparison with *SE*(σ). They are qualitatively similar to the dependences $S(\sigma)$ (Fig. 5b) and, especially, $K(\sigma)$ (Fig. 6b). In particular, the *PSNR*(N) function reveals no initial plateau at the smallest N, unlike the dependences $S(N)$ and $K(N)$. The corresponding reasons for so different behaviors need further analysis.

Overall, the main patterns observed by us agree well with the general considerations about the compressibility of nonrandom data [1, 3, 12, 13]. They also indicate a possible practical direction: the compression-based parameters can be used as simple indicators of disorder – due to either randomization or noise – which is present in the symbolic sequences of different natures.

4. CONCLUSIONS

We have modeled stochastically perturbed or corrupted textual documents and images from a unified position, by means of their partial randomization and metered adding of uncorrelated Gaussian noise. The changes occurring in a text or image under progressive disordering have been studied for the first time via analyzing a gradual evolution of their LZA-based compression coefficients S during the randomization or noising processes. The S parameter has been measured for the randomized natural-language texts and for the noisy or randomized images. We have demonstrated for the first time that the dependences $S(N)$ for both sample text and image behave in the same manner: first we observe the upper plateau $S \approx const$ at relatively small randomization parameters N, where the structural order and the information contents of a text or image persist (see also **Fig. 1**); then the $S(N)$ function changes abruptly in the intermediate region

given approximately by the equality $N \approx L$ (or $N \approx M$) and, finally it attains the lower plateau $S \approx const$ at $N \gg L$ ($N \gg M$), which corresponds to a completely disordered and so uninformative text or image.

For the text, the sentence-level randomization almost does not change the compressibility, whereas the word-level and, especially, the character-level types of randomization lead to a pronounced decrease in the S parameter. In all cases, the strongest changes occur in the vicinity of $N \approx L$. This shows that the typical number N of randomization cycles needed to destroy the order completely in a symbolic sequence is comparable with the total number L of its elements. For the image represented as a one-dimensional sequence of red-channel pixel values, the compression coefficient decreases significantly when the number N of pixel-randomization cycles reaches the total number M of pixels ($N \approx M$), quite similar to the case of our randomized text ($N \approx L$). Under Gaussian noise, the main decrease in S occurs at the σ values that correspond approximately to the standard deviation σ_0 of pixel intensities in the original image.

Unlike the case of texts, the limiting compression values for the strongly noised or randomized images remain essentially nonzero. This is explained mainly by a limited set of possible intensity values and the additional restrictions imposed on the pixel-value distribution by the image histogram. The additional JPEG-size analysis has led to qualitatively similar conclusions about the influence of randomization and noise on file compressibility. Finally, the dependences $S(N)$ and $K(N)$ for the image are qualitatively different from $S(\sigma)$ and $K(\sigma)$: the former reveal a clear initial plateau, contrary to the latter. This emphasizes different effects of randomization and noising on the structural order in the image. However, the dependences *PSNR*(N) and *PSNR*(σ) are much more similar to each other. Finally, it is worthwhile that the *SE*(σ) dependence also reveals the initial plateau at the smallest σ, unlike the *PSNR*(σ) dependence. In a sense, the S and K parameters are closer to *SE* than to *PSNR* in this respect.

The compression measure applied to images should be distinguished from purely histogram-based characteristics. In principle, a corrupted image can have exactly the same intensity histogram as the original image, while its spatial organization can be at least partly lost. This holds true, e.g., for an original image and its randomized versions. One can see that their compressibilities are radically different (see **Fig. 5a**), i.e. the compression parameter can detect the loss of spatial organization which is invisible in the histogram alone (see Fig. 7a). In other words, our LZA-based compression parameter for images has an important advantage over histogram-based parameters. Note in this respect that the computational complexity of compressibility is higher than that of *SE* but lower than that of *PSNR*.

Of course, modeling of stochastically perturbed textual and graphical data by randomization and Gaussian noising is restricted enough. The other factors hindering us from too categorical conclusions are a limited set of experimental samples and prematurity of any S-based detection-accuracy estimates. Nonetheless, our experimental dependences $S(N)$ and $S(\sigma)$ testify that the compression parameters can be, in principle, used for quantifying the loss of structural order in the data and detecting noised natural-language texts or images.

We perceive that further promising work can be related to broadening of the experimental objects under test, comparing different compression algorithms, as well as different methods of image unfolding and combinations of compression parameters with other descriptors of order and noise. This would aid in formulating the practical criteria of noise detection based upon compressibility.

ACKNOWLEDGMENTS AND FUNDING SOURCES

The authors received no financial support for the research, writing and publication of this article.

COMPLIANCE WITH ETHICAL STANDARDS

The authors declare that the research was conducted in the absence of any conflict of interest.

AUTHOR CONTRIBUTIONS

Conceptualization, [O.Ku.]; methodology, [O.Kr.]; validation, [O.Kr.]; formal analysis, [O.Kr., O.Ku.]; investigation, [O.Kr.]; resources, [O.Kr., O.Ku.]; data curation, [O.Kr.]; writing – original draft preparation, [O.Kr.]; writing – review and editing, [O.Ku.]; visualization, [O.Kr., O.Ku.]; supervision, [O.Ku.]; project administration, [O.Ku.].

All the authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] Dębowski, Ł. (2011). Excess entropy in natural language: present state and perspectives. *Chaos*, *21*, 037105. https://doi.org/10.1063/1.3630929

[2] Amancio, D. R., Altmann, E. G., Rybski, D., Oliveira Jr, O. N., & da F Costa, L. (2013). Probing the statistical properties: application to the Voynich manuscript. *PLoS ONE*, *8*, e67310. https://doi.org/10.1371/journal.pone.0067310

[3] Dębowski, Ł. (2015). The relaxed Hilberg conjecture: a review and new experimental support. *J. Quant. Linguistics*, *22*(4), 311–337. https://doi.org/10.1080/09296174.2015.1106268

[4] Timm, T., & Schinner, A. (2020). A possible generating algorithm of the Voynich manuscript. *Cryptologia*, *44*, 1–19. https://doi.org/10.1080/01611194.2019.1596999

[5] Feng, J., & Yu S. (2024). A method for measuring word sequence complexity of text. *J. Quant. Linguistics*, *32*(1), 82–102. https://doi.org/10.1080/09296174.2024.2417448

[6] Wieczyński, P., & Dębowski, Ł. (2025). Long-range dependence in word time series: the cosine correlation of embeddings. *Entropy*, *27*, 613. https://doi.org/10.3390/e27060613

[7] Cover, T., & Thomas, J. (1991). *Elements of Information Theory*. Wiley. 542 p.

[8] Akagi, T., Funakoshi, M., & Inenaga, S. (2023). Sensitivity of string compressors and repetitiveness measures. *Information and Computation*, 291, 104999. https://doi.org/10.1016/j.ic.2022.104999

[9] Jiang, Z., Yang, M., Tsirlin, M., Tang, R., Dai, Y., & Lin, J. (2023). “Low-resource” text classification: a parameter-free classification method with compressors. *Findings of the Association for Computational Linguistics: ACL 2023*, 6810–6828. https://doi.org/10.18653/v1/2023.findings-acl.426

[10] Karjus, A., Canet Solà, M., Ohm, T., Ahnert, S. E., & Schich, M. (2023). Compression ensembles quantify aesthetic complexity and the evolution of visual art. *EPJ Data Science*, *12*, 21. https://doi.org/10.1140/epjds/s13688-023-00397-3

[11] Merhav, N. (2026). Lempel–Ziv complexity, empirical entropies, and chain rules. *IEEE Trans. Inform. Theory*, *72*(5), 3486–3496. https://doi.org/10.1109/TIT.2026.3668013

[12] Lande, D. V., & Snarskii, A. A. (2007). On the role of autocorrelations in texts. *arXiv:0710.0225*. https://doi.org/10.48550/arXiv.0710.0225
[13] Gurzadyan, A. V., & Allahverdyan, A. E. (2016). Non-random structures in universal compression and the Fermi paradox. *Eur. Phys. J. Plus*, *131*, 26. https://doi.org/10.1140/epjp/i2016-16026-6
[14] Deng, W., Xie, R., Deng, S., & Allahverdyan, A. E. (2021). Two halves of a meaningful text are statistically different. *J. Statist. Mechanics: Theory and Experiment*, *3*, 033413. https://doi.org/10.1088/1742-5468/abe947
[15] Serrano, M. Á., Flammini, A., & Menczer, F. (2009). Modeling statistical properties of written text. *PLoS ONE*, 4(4), e5372. https://doi.org/10.1371/journal.pone.0005372
[16] Gonzalez, R., & Woods, R. (2018). *Digital Image Processing*. Pearson. 1020 p.
[17] Tang, C., Yang, X., & Zhai, G. (2015). Noise estimation of natural images via statistical analysis and noise injection. *IEEE Trans. on Circuits and Systems for Video Technol.*, *25*(8), 1283–1294 https://doi.org/10.1109/TCSVT.2014.2380196
[18] Zhang, Z. (2015). Image noise: detection, measurement, and removal techniques. In *Knoxville: Department of Electr. Eng. & Comp. Sci.* (10 pp.). https://zzutk.github.io/docs/reports/2015.07%20-%20Image%20Noise%20Detection,%20Measurement,%20and%20Removal%20Techniques.pdf
[19] Wu, Z., Lu, X., & Deng, Y. (2015). Image edge detection based on local dimension: a complex networks approach. *Physica A*, *440*, 9–18. https://doi.org/10.1016/j.physa.2015.07.020
[20] Dovhan, I. Ya., Kushnir, O. S., Furhala, Yu. M., & Panochko, H. I. (2022). Analogies between images and texts: burstiness phenomenon in texts and digital images. *Electronics and Information Technologies*, *17*, 3–15. https://doi.org/10.30970/eli.17.1
[21] Kushnir, O., Kravchuk, O., Drebot, A., & Franiv, V. (2025). Properties of networks built on digital images. *Electronics and Information Technologies*, *28*, 1–15. https://doi.org/10.30970/eli.29.2

КОЕФІЦІЄНТ СТИСНЕННЯ В АНАЛІЗІ ТЕКСТІВ І ЗОБРАЖЕНЬ ІЗ ШУМАМИ

Олексій Кравчук, Олег Кушнір
Кафедра оптоелектроніки та інформаційних технологій,
Львівський національний університет імені Івана Франка,
вул. Ген. Тарнавського, 107, м. Львів, 79017, УКРАЇНА

АНОТАЦІЯ

Вступ. Проаналізовано коефіцієнт стиснення S, пов'язаний з двома версіями алгоритму Лемпеля–Зіва (АЛЗ, англ. LZA), як кількісну характеристику структурного порядку в зашумлених або пошкоджених даних на прикладі тексту природною мовою та цифрового зображення. Для моделювання частково пошкоджених тексту або зображення ми рандомізували лінгвістичні елементи першого та додавали некорельований ґаусівський шум до другого або рандомізували його пікселі.

Матеріали та методи. Художній текст «Мобі Дік» Г. Мелвілла був рандомізований на лінгвістичних рівнях речень, слів і символів. Кількість циклів рандомізації N змінювали в діапазоні від 0 до 10^9. Проаналізовано також червоний канал 8-бітного зображення «Peppers» із розмірами $M = 512\times512$, до якого було додано ґаусівський шум зі стандартними відхиленнями σ у діапазоні від 0 до 400. Загалом було згенеровано 50 рандомізованих і 50 зашумлених версій зображення. Для застосування стиснення АЛЗ зображення розгортали в одновимірну послідовність інтенсивностей пікселів і обробляли відповідне текстове представлення зображення як звичайний текст.

Результати та обговорення. Рандомізація на рівні речень майже не змінює значення S, оскільки вона не змінює внутрішньої будови слів і речень, а АЛЗ має локальний характер. Рандомізація на рівні слів помірно зменшує S, а рандомізація на рівні символів приводить до найбільших змін S і майже нестисливих текстових файлів за найбільших N. Результати, одержані для двох версій АЛЗ, слабко відрізняються за впливом рандомізації речень на стисливість. Для рандомізованих зображень значення S різко зменшується, коли N наближається до загальної кількості пікселів M, а потім сягає насичення. Найбільші зміни S для зображень із шумами маємо в області $\sigma \in [10; 100]$.

Висновки. Метрики на основі стиснення АЛЗ чутливі до втрати локального порядку в будь-яких даних із шумами – і в рандомізованих текстах, і в рандомізованих та зашумлених зображеннях. Оскільки АЛЗ базується на повтореннях і локальних закономірностях у послідовності символів, зменшення S вказує на часткове руйнування порядку, спричинене рандомізацією або шумом. Отже, коефіцієнт стиснення може бути мірилом невпорядкованості в послідовностях лінгвістичних символів і в мережах пікселів, і, отже, може допомогти у виявленні зашумлених даних.

Ключові слова: стиснення даних, аналіз даних із шумами, розпізнавання шумів, рандомізація, опрацювання природної мови, опрацювання зображень.

Received / Одержано	Revised / Доопрацьовано	Accepted / Прийнято	Published / Опубліковано
06 August, 2026	23 September, 2026	25 September, 2026	28 September , 2026