Information-Entropic Laws of Language as Complex Self-Organizing System

Bikesh Ospanova

Karaganda State Technical University, Mira blvd, 56, 100027, Karaganda, Kazakhstan

Abstract: In the article there are reflected the results of the study which purpose was using a measure of information amount permitting to analyze the general mechanisms of information-entropic characteristics of the texts laying in the basis of all spontaneously running in the surrounding world processes of information accumulating that lead to the system structure self-organization. There are presented some aspects of the experimental approach to calculating the text entropy, experimental data demonstrating the results. There is suggested a linguistic-mathematical model for analyzing the text structure based on the fundamental law of preserving the sum of information and entropy using Shannon's formula, as well as a comparative analysis of the text entropic characteristics in Kazakh and Russian.

[Ospanova B. **Information-Entropic Laws of Language as Complex Self-Organizing System**. *Life Sci J* 2013;10(12s):460-463] (ISSN:1097-8135). http://www.lifesciencesite.com. 78

Keywords: Entropy, information, text, language, self-organizing system, study, dynamics, information theory, hierarchy.

1. Introduction

Informatization of the society is inseparably connected with informatization of science that in turn uses information in complex managerial, social, information systems. Today information technologies play the key role in all the spheres of social life, therefore the main kind of human activity become more often methods of obtaining, storing and widening of knowledge and information.

Studying a language by methods of information theory became a prospective scientific line studying complex systems from the point of view of running in them processes of self-organization. Within this line there takes place the language modeling as a complex, dynamic, self-organizing system from the disordered state to the ordered one.

In the article there is used the fundamental law of preserving the sum of information and entropy using Shannon's formula.

In the present day world information presents one of the most important resources, one of the factors of the human society development. Information processes taking place in the material world, life and human society are studied by almost all scientific disciplines.

"Among fundamental laws of the nature there is especially important the value of the laws of preserving the amount of motion, energy, etc., as well as the growth of **entropy** which essence consists in that any system, left to its own resources, disorganizes, there increases disorder, chaos in it, there take place energy losses, and at last there comes disintegration. In the nature and in the society, however, there exist reverse processes: of ordering, of forming, system-forming, accumulating, hierarchic energy levels increasing as a result of reasonable activity of a man striving for the world inspiration and transformation" [1].

Today entropy is a concept widely used in various fields of science: in mathematical theory of metric spaces, in theory of management, in biological ecology, in linguistics, medicine, for example, in statistic physics, in information theory, etc.

Information and entropy characterize a complex system from the point of view of order and chaos, at this information is a measure of order, then entropy is a measure of disorder. This measure stretches from the maximum entropy, i.e. chaos, complete uncertainty, to the highest level of order. So, the level of organization is defined by the level of information at which the system is.

Information entropy is a measure of information chaotic character, an uncertainty of some symbol of the primary alphabet occurring. If there are no information losses, it is numerically equal to the amount of information per a symbol of the message transmitted [2].

The basis of studies devoted to studying the methodology of information and entropy can be taken, in our opinion, the works of such scientists, as S. Angrist and L. Hepler [3], C. Shannon [4], R. Arnheim [5], L. Whyte [6], R. Narashimha[7], R. Carnap [8] and many others.

There can be presented a lot of different works devoted to this subject. The concept of entropy was introduced by Clausius in the XIX century as a characteristic of the disorder degree. With the help of entropy it became possible to evaluate such important concepts, as order and disorder. For example, S. Angrist and L. Hepler give the following definition of entropy: "…entropy is defined as a quantitative measure of disorder in a system…" [3]. In the doctor of philology M.Yu. Oleshkov's opinion, "…entropy is understood as a measure of uncertainty (unpredictability) of the text development in the discourse process characterized by the possibility of selecting as the following some stage of a number of variants. The entropy indicator characterizes quantitatively the level of information order of the text as a system – the larger is it, the less is ordered the system (=the text), the more is its deflection from the "ideal" development. Thus, entropy is a function of state; any state of a system can be given quite certain value of entropy" [9].

When defining the amount of information there is considered a language text that consists of letters, words, word combinations, sentences, etc. each letter occurrence is described as a consequent realization of a certain system. The amount of information expressed by the indicated letter is equal in its absolute value to the entropy (uncertainty) that characterized the system of possible choices and that was removed as a result of selecting a certain letter.

It is known that to calculate entropy it is necessary to have a complete distribution of possible combinations probability. To calculate this or that letter entropy it is needed to know each possible letter occurrence probability. From this point of view it is interesting to consider the language information interpretation. In this aspect we analyzed the existing methods of presenting complex systems from the point of view of entropic-information laws.

The basis of the information-entropic analysis made comparison of the texts of different genres and styles in Russian and Kazakh. Using the synergetic theory of information there was carried out a structural analysis of arbitrary texts from the side of their chaos and order by the number of separate letters frequency.

There was developed a linguistic-mathematical model for analyzing the text structure based on the fundamental law of preserving the sum of information and entropy using Shannon's formula. In general characteristic of the entropic-information (entropy is a measure of disorder, information is a measure of disorder elimination) analysis of the texts we used Shannon's statistical formula for determining the text perfection, harmony:

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i \text{, [10]}$$

where p_i is probability of detecting a system unit in their lot N ; $\sum_{i=1}^{N} p_i = 1$, $p_i \geq 0$, $i = 1,2,...,N$.

In the course of the experiment there was carried out the information-entropic analysis of the texts containing 500 characters of the scientific, business-official, journalistic, informal and belles-lettres styles of speech in Kazakh and Russian. To calculate the texts information there were computed the probabilities of the occurring of one letter, two-letter, three-letter, four-letter, five-letter and six-letter combinations.

Note that in this article as an example there are shown the calculations of the texts of the business-official style of speech in Kazakh and Russian. The material for the experiment served a text in Russian from the RK Constitution, the< main law of Kazakhstan acting from the day of adopting the Republic of Kazakhstan Constitution.

To calculate the business-official text information containing 500 characters there were calculated the probabilities of the occurring of one letter, two-letter, three-letter, four-letter, five-letter and six-letter combinations. The selected abstract contains 500 characters with blanks, 441 without blanks [11].

Since the Russian alphabet contains 32 letters (31 letters and a blank), according to this result

$$H_0 = \log 32 = 5 \text{ bit,}$$

where H_0 is the maximum value of the text entropy consisting in acceptance of one letter of the Russian text (information contained in one letter) under the condition that all the letters are considered equally probable;

bit is a unit of information measuring.

To calculate the text information we computed the probabilities of each letter occurrence in this abstract. In calculation there were considered 32 letters of the Russian alphabet and a blank, all the other characters (brackets, quotations, commas, etc.) were not considered. The numerical data contained in the text were written in words. The calculation of probability (p) of various letters occurrence in the text is performed by calculating the relative frequency of individual letters. To define one letter occurrence probability in the text there was used the classical formula for determining probability:

$$P(oneletter) = \frac{m}{n},$$

where P is the relative frequency;
m is the number of one letter occurrence in the text;
n is the number of all letters occurrence in the text.

In the course of calculations there were obtained the following results:

The text entropy with accounting one letter is equal:

$$H_1 = H(\alpha_1) = -0{,}118 \cdot \log_2(0{,}118) - 0{,}11 \cdot \log_2(0{,}11) - ...$$
$$- 0{,}002 \cdot \log_2(0{,}002) \approx 4{,}2746$$

The text entropy with accounting two-letter combinations H_2:

$$H_2 = H\alpha_1(\alpha_2) = H(\alpha_1\alpha_2) - H(\alpha_1) = -0{,}04 \cdot \log_2(0{,}04) - 0{,}038 \cdot \log_2(0{,}038) - ...$$
$$- (0{,}002) \cdot \log_2(0{,}002) + 0{,}118 \cdot \log_2(0{,}118) + 0{,}11 \cdot \log_2(0{,}011) + ...$$
$$+ 0{,}002 \cdot \log_2(0{,}002) \approx 2{,}6721$$

The text entropy with accounting three-letter combinations H_3:

$$H_3 = H\alpha_1\alpha_2(\alpha_3) = H(\alpha_1\alpha_2\alpha_3) - H(\alpha_1\alpha_2) = -0{,}02 \cdot \log_2(0{,}02) - ...$$
$$- 0{,}02 \cdot \log_2(0{,}02) - 0{,}002 \cdot \log_2(0{,}002) + 0{,}04 \cdot \log_2(0{,}04) + ...$$
$$+ 0{,}038 \cdot \log_2(0{,}038) + 0{,}002 \cdot \log_2(0{,}002) \approx 0{,}9196$$

The text entropy with accounting four-letter combinations H_4:

$$H_4 = H\alpha_1\alpha_2\alpha_3(\alpha_4) = H(\alpha_1\alpha_2\alpha_3\alpha_4) - H(\alpha_1\alpha_2\alpha_3) =$$
$$= -0{,}01 \cdot \log_2(0{,}01) - 0{,}01 \cdot \log_2(0{,}01) - 0{,}002 \cdot \log_2(0{,}002) + 0{,}02 \cdot \log_2(0{,}02) + ...$$
$$+ 0{,}02 \cdot \log_2(0{,}02) + ... + 0{,}002 \cdot \log_2(0{,}002) \approx 0{,}3290$$

The text entropy with accounting five-letter combinations H_5:

$$H_5 = H\alpha_1\alpha_2\alpha_3\alpha_4(\alpha_5) = H(\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5) - H(\alpha_1\alpha_2\alpha_3\alpha_4) =$$
$$= -0{,}01 \cdot \log_2(0{,}01) - 0{,}01 \cdot \log_2(0{,}01) - ...$$
$$- 0{,}002 \cdot \log_2(0{,}002) + 0{,}01 \cdot \log_2(0{,}01) + ... + 0{,}002 \cdot \log_2(0{,}002) \approx 0{,}1517$$

At last, the text entropy with accounting six-letter combinations H_6:

$$H_6 = H\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5(\alpha_6) = H(\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5\alpha_6) - H(\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5) =$$
$$= -0{,}01 \cdot \log_2(0{,}01) - 0{,}01 \cdot \log_2(0{,}01) - ... - 0{,}002 \cdot \log_2(0{,}002) +$$
$$+ 0{,}01 \cdot \log_2(0{,}01) + ... + 0{,}002 \cdot \log_2(0{,}002) \approx 0{,}01046$$

As a result of all performed calculations of the number of different letter combinations in the business-official text of Russian we came to the following indicators:

H_1	H_2	H_3	H_4	H_5	H_6
4,2746	**2,6721**	**0,9196**	**0,3290**	**0,1517**	**0,1046**

The calculations carried out confirm the fact that with information increasing there takes place the text uncertainty degree (entropy) decrease. Thus, we can conclude that the amount of information and entropy in the business-official style differs principally from the texts of other styles. It is probably due to the exact form of exposition, rule-boundedness, stability and high information capacity of the speech means.

Let's analyze a text of the business-official style in Kazakh that contains 500 characters with blanks and 434 without blanks [12].

For more accurate calculation of information contained in one letter of the Kazakh text we need to know probabilities of different letters occurrence. These probabilities can be approximately determined if we take an abstract and compute for it relative frequencies of individual letters.

Since the Kazakh alphabet contains 43 letters (42 letters and a blank), according to this result:

$$Ho = \log 43 = 5{,}4 \text{ bit}$$

- the entropy of the experiment consisting in the acceptance of one letter of the Kazakh text (information contained in one letter) under the condition that all letters are considered equally probable**.**

Using the classical formula of determining probability, let's calculate for it the relative frequencies of individual letters and relative frequencies in the diminishing order.

Equating these frequencies to the probabilities of corresponding letters occurrence we will obtain based on Shannon's information entropy formula for calculating the entropy maximum value with accounting one letter of the Kazakh text:

$$H_1 = H(\alpha_1) = -0{,}132 \cdot \log_2(0{,}132) - 0{,}130 \cdot \log_2(0{,}130) - ...$$
$$- 0{,}002 \cdot \log_2(0{,}002) \approx 4{,}3443$$

Now let's calculate the text conditional entropy with accounting two-, three-, four-, five-, six-letter combinations.

The results show the following figures:

$$H_2 = H\alpha_1(\alpha_2) = H(\alpha_1\alpha_2) - H(\alpha_1) = -0{,}028 \cdot \log_2(0{,}028) - 0{,}028 \cdot \log_2(0{,}028) - ...$$
$$- (0{,}002) \cdot \log_2(0{,}002) + 0{,}132 \cdot \log_2(0{,}132) + 0{,}130 \cdot \log_2(0{,}130) + ...$$
$$+ 0{,}002 \cdot \log_2(0{,}002) \approx 2{,}6006$$

$$H_3 = H\alpha_1\alpha_2(\alpha_3) = H(\alpha_1\alpha_2\alpha_3) - H(\alpha_1\alpha_2) =$$
$$= -0{,}016 \cdot \log_2(0{,}016) - 0{,}012 \cdot \log_2(0{,}012) - ... - 0{,}002 \cdot \log_2(0{,}002) +$$
$$+ 0{,}028 \cdot \log_2(0{,}028) + 0{,}028 \cdot \log_2(0{,}028) + ... + 0{,}002 \cdot \log_2(0{,}002) \approx 1{,}0225$$

$$H_4 = H\alpha_1\alpha_2\alpha_3(\alpha_4) = H(\alpha_1\alpha_2\alpha_3\alpha_4) - H(\alpha_1\alpha_2\alpha_3) =$$
$$= -0{,}012 \cdot \log_2(0{,}012) - 0{,}008 \cdot \log_2(0{,}008) - ... - 0{,}002 \cdot \log_2(0{,}002) +$$
$$+ 0{,}016 \cdot \log_2(0{,}016) + 0{,}012 \cdot \log_2(0{,}012) + ... + 0{,}002 \cdot \log_2(0{,}002) \approx 0{,}2665$$

$$H_5 = H\alpha_1\alpha_2\alpha_3\alpha_4(\alpha_5) = H(\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5) - H(\alpha_1\alpha_2\alpha_3\alpha_4) =$$
$$= -0{,}008 \cdot \log_2(0{,}008) - 0{,}006 \cdot \log_2(0{,}006) - ... - 0{,}002 \cdot \log_2(0{,}002) +$$
$$+ 0{,}012 \cdot \log_2(0{,}012) + ... + 0{,}002 \cdot \log_2(0{,}002) \approx 0{,}2012$$

$$H_6 = H\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5(\alpha_6) = H(\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5\alpha_6) - H(\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5) =$$
$$= -0{,}006 \cdot \log_2(0{,}006) - 0{,}006 \cdot \log_2(0{,}006) - 0{,}006 \cdot \log_2(0{,}006) - ...$$
$$- 0{,}002 \cdot \log_2(0{,}002) + 0{,}008 \cdot \log_2(0{,}008) + ... + 0{,}002 \cdot \log_2(0{,}002 \approx 0{,}095$$

As a result there were obtained the following values (in bits) in Kazakh:

H_1	H_2	H_3	H_4	H_5	H_6
4,3443	**2,6006**	**1,0225**	**0,2665**	**0,2012**	**0,095**

Thus, the further calculations of the texts from one to six-letter combinations for Kazakh and Russian are not similar. Based on the carried out calculations it can be supposed that in the business-official style of both languages with information increasing there decreases the degree of uncertainty (entropy). In Kazakh and Russian entropy is equal to (in bits):

In Russian

H_1	H_2	H_3	H_4	H_5	H_6
4,2746	**2,6721**	**0,9196**	**0,3290**	**0,1517**	**0,1046**

In Kazakh

H_1	H_2	H_3	H_4	H_5	H_6
4,3443	**2,6006**	**1,0225**	**0,2665**	**0,2012**	**0,095**

From here we can conclude that the dynamics of the text entropy decreases with the transition to a higher level of organization, besides, there increases the text information capacity that confirms the language development according to the law of preserving the sum of information and entropy.

According to the obtained data, the text possessing its inherent signs, namely: information capacity, coherence, completeness and characterized by the style and genre tinge, is built according to certain laws that impact the information and entropy presence in the text.

Within the frames of the study we presented an attempt of theoretical substantiation of the possibility and necessity of using a synergetic paradigm for analyzing a linguistic text, developed a linguistic-mathematical model of the text based on C. Shannon's formula. We tried to demonstrate the possibility of using the synergetic approach to the text analysis in Russian and Kazakh. The study results confirmed once more that synergy explains the evolution of any complex dissipative system as self-development.

2. Conclusion

Thus, using entropy in linguistics is one of the most universal characteristics of the text, the indicator of its complexity in the information-entropic sense. That's why a language is a complex multi-level object of synergy possessing energy and inner life.

Drawing up the balance of this study we would like to note that this fact is explained by the different numbers of the hierarchic system elements, different number of letters in the considered alphabets of Russian and Kazakh. The text entropy decrease at higher levels prove the fact that for a multi-level hierarchic system it is very important to describe a lower level as an interaction of interconnected systems, each of which possesses its information properties. We established that with the transition to a higher level of the hierarchic system that is based on accounting the increasing letter combinations, the text information capacity increases. The approach considered, in our opinion, corresponds completely to the basic requirements of the system entropic-information analysis as it provides, in the hierarchic system modeling, the integrity of its consideration due to general-theoretical and methodological conceptions.

In the conclusion we will note that there was developed algorithmic software for the system that can be used for carrying out complex text studies in any language. The methodology suggested and its program realization will permit to reduce labor hours for calculating any text entropy and to increase the calculation accuracy.

Corresponding Author:
Dr. Ospanova,
Karaganda State Technical University, Mira blvd, 56,
100027, Karaganda, Kazakhstan

References

1. Ponomareva A.I., 1989. The issues of history of the national economy and economic thought. Istoki. V.1., pp: 227.
2. Entropy (information theory). Date Views 03.11.2013 www.en.wikipedia.org/wiki/Entropy_%28information_theory%29
3. Angrist S.W., L.G. Hepler, 1967. Order and Chaos: Laws of Energy and Entropy. N.Y.: Basic Books, pp: 146.
4. Shannon C.E., 1948. A Mathematical Theory of Communication. Bell System Technical Journal, 27: 623–656.
5. Arnheim R., 1971. Entropy and Art. An Essay on Disorder and Order. Berkley and Los Angeles: University of California Press, pp: 61.
6. Whyte L., 1965. Law Atomism, Structure, and Form. Structure in Art and Science. Kepes Gyorgy (ed.). New York: Braziller: 20-28.
7. Narashimha R., 1994. Linguistic Entropy in Othello of Shakespeare. New Delhi: M D Publications Ltd, pp: 95.
8. Carnap R., 1977. Two Essays on Entropy. Berkley and Los Angeles: University of California Press, pp: 115.
9. Oleshkov M.Yu., 2006. Bases of functional linguistics: discourse aspect: Tutorial for students of Rus. Lang. and liter. departments. Nizhni Tagil, pp: 146.
10. Shannon C.E., 1963. Mathematical theory of communication. The works on information theory and cybernetics. M.: IL, pp: 243-332.
11. Constitution of the Republic of Kazakhstan, 2012. Almaty: LLP «88i8», pp: 56.
12. Duisembekova L., 2010. Kazakh language: introduction in office work. Almaty: Institute of the state language development, pp: 400.

11/29/2013